UNIVERSIDADE DO ESTADO DA BAHIA

DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

DIEGO DOS SANTOS

ANÁLISE DO ALGORITMO *RANDOM FOREST* NA CLASSIFICAÇÃO DE SINTOMAS DAS DOENÇAS ARBOVIRAIS

SALVADOR

2022

DIEGO DOS SANTOS

ANÁLISE DO ALGORITMO *RANDOM FOREST* NA CLASSIFICAÇÃO DE SINTOMAS DAS DOENÇAS ARBOVIRAIS

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito à obtenção do grau de Bacharel em Sistemas de Informação. Área de Concentração: Ciência da Computação

Orientadora: Maria Inés Valderrama Restovic

SALVADOR

2022

FICHA CATALOGRÁFICA
Sistema de Bibliotecas da UNEB

S237a Santos, Diego dos

Análise do algoritmo Random Forest na classificação de sintomas das doenças arbovirais / Diego dos Santos. - Salvador, 2022.
60 fls : il.

Orientador(a): Maria Inés Valderrama Restovic.
Inclui Referências
TCC (Graduação - Sistemas de Informação) - Universidade do Estado da Bahia. Departamento de Ciências Exatas e da Terra. Campus I. 2022.

1.Aprendizado de Máquina. 2.Random Forest. 3.Arbovírus. 4.Vírus da dengue. 5.Vírus da chikungunya.

CDD: 604

DIEGO DOS SANTOS

ANÁLISE DO ALGORITMO *RANDOM FOREST* NA CLASSIFICAÇÃO DE SINTOMAS DAS DOENÇAS ARBOVIRAIS

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito à obtenção do grau de Bacharel em Sistemas de Informação. Área de Concentração: Ciência da Computação

Aprovada em:

BANCA EXAMINADORA

Maria Inés Valderrama Restovic (Orientadora)
Universidade do Estado da Bahia – UNEB

Antonio Carlos Fontes Atta
Universidade do Estado da Bahia – UNEB

Diego Gervasio Frías Suárez
Universidade do Estado da Bahia – UNEB

AGRADECIMENTOS

Em primeiro lugar gostaria de agradecer a Deus, que fez com que meus objetivos fossem alcançados durante todos os meus anos de estudos. Aos membros da minha família, que me incentivaram nos momentos difíceis e compreenderam a minha ausência enquanto eu me dedicava aos estudos. A professora Inês Restovic, por ter sido minha orientadora e ter desempenhado uma função fundamental na minha formação com tanta dedicação e amizade. Aos meus colegas de curso, com quem convivi intensamente durante os últimos anos, compartilhando momentos dentro e fora do âmbito acadêmico, que me permitiram crescer não só como pessoa, mas também como formando. Por último e não menos importante, a minha esposa Flávia Souza que sempre esteve ao meu lado, me incentivou e ajudou em todos esses anos de graduação.

A vitalidade é demonstrada não apenas pela persistência, mas pela capacidade de começar de novo. (F. Scott Fitzgerald)

RESUMO

As mudanças climáticas provocadas pelo aquecimento global aumentam a temperatura do planeta, beneficiando a proliferação dos vírus. Os mosquitos *Aedes Aegypti* e *Aedes Albopictus* são os principais transmissores de arbovírus, especificamente dos vírus da dengue (DENV) e vírus da chikungunya (CHIKV). Os pacientes infectados por essas arboviroses apresentam sintomas semelhantes que dificultam o trabalho inicial do diagnóstico médico. A integração da tecnologia na área médica traz uma série de benefícios, desde o atendimento médico até em momentos cirúrgicos. A introdução do aprendizado de máquina vem crescendo em termos de relevância nos últimos anos, graças à quantidade massiva de dados gerados. Vários algoritmos são analisados e comparados para identificar padrões e correlações com dados das arboviroses. O uso do *Random Forest* (RF) para o entendimento das arboviroses está em fase inicial e não foi utilizado em uma análise conjunta com DENV e CHIKV. De todo modo, os estudos na sua maior parte são executados de forma isolada com DENV. As características do algoritmo RF chamam bastante atenção por resolver problemas comuns dos algoritmos de aprendizado de máquina, com a criação de árvores de decisão que trabalham de forma isolada, mas têm fator decisivo no resultado final do modelo, além do seu processo de aleatoriedade das amostras para gerar as árvores de decisão. Neste estudo foi desenvolvido um modelo classificador com a RF que apresentou comportamento muito sensível em relação ao conjunto de dados, onde os rótulos imprecisos reduziram as métricas de desempenho. Os ajustes realizados inicialmente com o conjunto de dados, demonstraram evolução nas métricas de desempenho. Outras características marcantes foram: o alto consumo de recursos computacionais e o curto tempo de treinamento para obter um modelo. No primeiro momento, o modelo teve uma acurácia de 59%, mas com todos os ajustes realizados durante o desenvolvimento, obteve-se 76% de acurácia no classificador final. Apesar do resultado geral, as métricas de desempenho foram melhores para CHIKV, pois os sintomas característicos foram presentes em muitas amostras de pacientes rotulados por esse arbovírus.

Palavras-chave: Aprendizado de máquina. Random Forest. Arbovírus. Vírus da dengue. Vírus da chikungunya.

ABSTRACT

Climate change caused by global warming increases the temperature of the planet, benefiting the prospect of viruses. The *Aedes Aegypti* and *Aedes Albopictus* mosquitoes are the primary transmitters of arboviruses, specifically DENV and CHIKV. Patients infected with these arboviruses have similar symptoms that make the initial work of medical diagnosis difficult. The integration of technology in the medical field brings a series of benefits, from medical care to surgical procedures. The introduction of machine learning has been growing in terms of relevance in recent years, thanks to the massive amount of data generated. Several algorithms are analyzed and compared to identify patterns and correlations with arboviral data. The use of RF for understanding arboviruses is in its infancy and was not used in a joint analysis with DENV and CHIKV. Anyway, most studies are executed in isolation with DENV. For solving common machine learning algorithms problems, the RF algorithm has gained attention because of its characteristics, which involve the creation of decision trees that work in isolation but have a decisive influence on the final model outcome, along with the randomness of the samples used to generate the trees. In this study, a classifier model was developed with the RF that presented a very sensitive behavior about the data set, where imprecise labels reduced the performance metrics. The adjustments made initially with the data set showed evolution in the performance metrics. Other striking features were: the high consumption of computational resources and the short training time to obtain a model. At first, the model had an accuracy of 59%, but with all the adjustments made during development, we obtained 76% accuracy in the final classifier. Despite the general result, the performance metrics were better for CHIKV, as the characteristic symptoms were present in many samples of patients labeled by this arbovirus.

Keywords: Machine Learning. Random Forest. Arbovirus. Dengue virus. Chikungunya virus.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE CÓDIGOS-FONTE

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Networks*
AUC	*area under the ROC curve*
BPNN	*Back-Propagation Neural Networks*
CHIKV	vírus da chikungunya
CV	*Cross Validation*
DATASUS	Departamento de Informática do Sistema Único de Saúde
DENV	vírus da dengue
DENV-4	vírus da dengue - sorotipo 4
DENV-3	vírus da dengue - sorotipo 3
DENV-2	vírus da dengue - sorotipo 2
DENV-1	vírus da dengue - sorotipo 1
DF	febre da dengue
DHF	febre hemorrágica da dengue
DSR	*Design Science Research*
DTR	*Decision Trees Regression*
GBM	*Gradient Boosting Machine*
GCP	Google Cloud Platform
GLM	*Generalized Linear Model*
Kappa	*Cohen's Kappa Coefficient*
KNN	*K-nearest neighbors*
LM	*Linear Regression Model*
MCC	*Matthews Cohen Correlation*
ML	*Machine Learning*
NB	*Naive Bayes*
NN	*Neural Networks*
OMS	Organização Mundial da Saúde
OOB	*out-of-bag*
PCR	*Polymerase Chain Reaction*
RF	*Random Forest*

ROC	*Receiver Operating Characteristic*
RT-PCR	*Reverse Transcription PCR*
SINAN	Sistema de Informação de Agravo de Notificação
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
TABWIN	*Tab para Windows*
ZIKV	vírus da zika

SUMÁRIO

1 INTRODUÇÃO

As recentes mudanças climáticas provocadas pelo aquecimento global aumentam a temperatura média dos oceanos e da atmosfera devido à emissão massiva de gases de efeito estufa como vapor de água, metano, ozônio, dióxido de carbono, clorofluorcarbonos e óxido nitroso (1). O efeito estufa, por outro lado, é essencial para a manutenção da temperatura do planeta porque evita o resfriamento extremo. Atividades industriais, agrícolas e de desmatamentos são fatores que colaboram na intensificação das emissões de gases que em demasia causam sérias consequências.

Em decorrência destas atividades, a temperatura em diversas regiões do planeta tem se elevado ao longo dos anos, causando problemas como derretimento das calotas polares e escassez de água, onde mais regiões se tornam propícias à disseminação de doenças transmitidas por arbovírus (*Arthropod borne virus*), um conjunto de vírus que se replicam em artrópodes como insetos e aracnídeos e podem ser transmitidos a hospedeiros vertebrados (2).

A transmissão é fortalecida em regiões tropicais e subtropicais que beneficiam a reprodução dos vetores virais devido à alta temperatura, como os mosquitos *Aedes Aegypti* e *Aedes Albopictus*. São eles os principais transmissores de arbovírus mais conhecidos por casos de infecção do DENV. Estes mosquitos não se limitam a uma única doença viral, causando preocupação na sociedade pela variedade de vírus que podem ser transmitidos. As outras doenças virais cujo o número de pesquisas na área de medicina e tecnologia da informação são a febre-amarela, chikungunya e o zica, que também são transmitidas por estes mosquitos.

Devido à preocupação de casos epidêmicos dessas doenças ocorrerem com mais frequência a cada ano, as pesquisas têm contribuído para investigar fatores ocasionados por estas doenças e as dificuldades que elas impõem ao prognóstico médico. É recorrente a dificuldade de detecção dos vírus em certos estágios de infecção em razão da similaridade de alguns sintomas, por exemplo, febre e dor no corpo (3). Entretanto, em determinados estágios da doença surgem características únicas que possibilitam a análise mais precisa para o diagnóstico médico, como as dores nas articulações causadas pela infecção CHIKV. Essa similaridade dificulta o tratamento ágil baseado somente nos sintomas apresentados pelos pacientes. Exames como *Polymerase*

Chain Reaction (PCR) são fundamentais em regiões epidêmicas para efetuar o diagnostico, mas seu custo para distribuição em massa é elevado, em países com menores recursos financeiros se torna um desafio.

Complicações pós-infecção atraíram a atenção de pesquisadores e, novos estudos baseadas na obtenção de dados sintomatológicos dos pacientes infectados pelas doenças arbovirais visam entender e auxiliar na rápida detecção e prevenção. Infelizmente, existe uma escassez de ferramentas e não existe vacinas para o controle do DENV (4). A detecção precoce é relevante para iniciar o tratamento adequado, além de ser uma forma de alerta sobre as ações de vigilâncias ambientais e epidemiológicas. A importância se dá pela carência de medicamentos para evitar epidemias, e como consequência os principais controles dos vetores virais são baseados em campanhas de conscientização da sociedade, principalmente por água parada nas comunidades (5). Mesmo que os meios de comunicação tenham evoluído para divulgar informações sobre epidemias de forma rápida, existem áreas que estão em desenvolvimento onde a tecnologia pode ajudar a gerar novos conhecimentos e auxiliar na prevenção de epidemias.

A informática clínica e a informática em saúde pública podem ser essenciais para melhorar a capacidade de produzir resultados de pesquisa básica e avaliar a eficiência das intervenções nas comunidades (3). A ciência de dados tem evoluído, assumindo papel significante para auxiliar em vár aias áreas do conhecimento, inclusive na área da saúde. Segundo Fahmi et al. (6) as técnicas como mineração de dados têm um papel essencial na previsão de doenças, sendo utilizada para localizar padrões, conexões, correlações e anomalias em conjuntos de dados, produzindo conhecimento para ajudar nas tomadas de decisões. No subcampo da mineração de dados, algoritmos de aprendizado de máquina foram desenvolvidos com habilidade de aprendizado para resolução de problemas, como classificação.

O algoritmo RF que é em português significa floresta aleatória, como o seu próprio nome indica, opera com muitas árvores de decisão de forma aleatória. Sendo sua principal característica, cada árvore é utilizada na composição do resultado final, tornando o método mais robusto. Como consequência, estudos aplicam com sucesso o algoritmo para resolver problemas de classificação (3). Por conta da sua aleatoriedade, o RF é visto como um bom algoritmo para diminuir variância em problemas de classificação, aumentando a precisão dos resultados (7).

Estudos correlatos de comparação entre algoritmos envolvendo o RF tiveram como objetivo testar e avaliar o desempenho de algoritmos de classificação para casos de infecção na

previsão do vírus da DENV e vírus da zika (ZIKV). Com base em suas métricas de desempenho obtidas através dos resultados de treinamento e teste, foi possível notar que o RF ficou bem próximo dos melhores algoritmos utilizados nas pesquisas (6, 8). O algoritmo apresenta aspectos importantes que promovem a sua utilização: 1. provê fácil configuração inicial e baixo tempo de processamento para gerar o classificador, quando comparado a outros algoritmos com desempenho similar; 2. aleatoriedade em todo o processo de separação das amostras de dados, a fim de diminuir a variância, 3. multifuncional, pode ser utilizado para problemas de classificação e regressão; 4. diversidade, vários critérios para avaliação das variáveis importantes, com propósito de melhorar a precisão dos modelos.

Com base nas lacunas e nos questionamentos apresentados anteriormente sobre a utilização da RF em epidemias para problemas de classificação das arboviroses, este trabalho analisou o potencial do RF para identificar doenças arbovirais baseado nos dados sintomatológicos de pacientes infectados por DENV e CHIKV. O modelo classificador deste trabalho obteve 76% de acurácia, demonstrando potencial para fornecer auxílio aos profissionais da área de saúde. As características de versatilidade para tratar de problemas diversos, facilidade para configurar os hiperparâmetros e curto período de treinamento do RF obtendo bons resultados, colocam o algoritmo como uma boa opção para estudos com arboviroses. Os estudos podem ser aprofundados a partir deste trabalho, pois o artefato concebido, permite a continuação do desenvolvimento, graças a metodologia da pesquisa *Design Science Research* (DSR) aplicada.

O caminho percorrido para analisar e criar o classificador com RF será descrito sua estrutura nesta monografia, a seguir. O capítulo 2 de fundamentação teórica contextualiza sobre as arboviroses, seus meios de transmissões, família e os principais vírus DENV, ZIKV e CHIKV. De forma separada será abordado sobre os arbovírus com foco na similaridade dos sintomas. Neste capítulo também é abordado sobre aprendizado de máquina e como esta técnica está auxiliando a área médica para identificar padrões e gerar novos conhecimentos.

A metodologia de pesquisa é explicada no capítulo 3, onde discutimos como o objetivo de resolver um problema prático com a construção de um artefato se encaixou perfeitamente com a metodologia proposta. A partir dos ciclos de pesquisa, o processo garantiu a geração de conhecimento enquanto o artefato foi desenvolvido com base nas conjecturas teóricas.

O capítulo 4 discorre sobre os processos de desenvolvimento de pesquisa e do artefato ao relacionar os princípios do DSR aplicado no problema da pesquisa de analisar a RF

na classificação dos sintomas relacionados a arbovírus. As ferramentas, métricas e os ciclos de desenvolvimento são detalhados com o propósito de elucidar e garantir que as etapas possam ser replicadas. Em cada ciclo existe uma análise sobre as métricas de desempenho, com o propósito de que, a cada ciclo sejam melhorados os resultados obtidos.

No capítulo 5 é apresentada a conclusão do trabalho de forma detalhada com a percepção sobre todo o desenvolvimento do trabalho, assim como, a possibilidade de continuação deste trabalho com os possíveis caminhos a serem trilhados.

2 APRENDIZADO DE MÁQUINA NO ESTUDO DO ARBOVÍRUS

O desenvolvimento computacional abriu portas para inúmeros benefícios aos seres vivos. Máquinas com mais capacidade de processamento e armazenamento, puderam colaborar para o avanço da tecnologia em diferentes partes, incluindo da biologia. O volume de dados gerados com informações de pacientes infectados por vírus em banco de dados médicos podem ser aproveitados para aplicação de estudos e geração de novos conhecimentos. Estudos estão sendo feitos mundialmente utilizando algoritmos de *Machine Learning* (ML) em arbovírus para amparar as previsões de epidemias e classificar as doenças ocasionadas por eles.

2.1 ARBOVÍRUS

Conhecidos amplamente como arbovírus, *Arthropod Borne Virus* são vírus que podem ser transmitidos aos humanos através de artrópodes (2), em razão de seu ciclo de reprodução principal ocorrer com insetos. A transmissão acontece por meio da picada dos artrópodes que se alimentam de sangue, provocando a infecção em humanos e animais. Nesse processo, dependendo da região, pode ser bem complexo identificar a qual das cinco famílias pertencem o arbovírus envolvido: *Bunyaviridae*, *Togaviridae*, *Flaviviridae*, *Reoviridae* e *Rhabdoviridae*. Lopes, Nozawa e Linhares (9), em 2014, informaram sobre a quantidade estimada de espécies de arbovírus no mundo com base nas famílias apresentadas.

> *Estima-se que haja mais de 545 espécies de arbovírus, dentre as quais, mais de 150 relacionadas com doenças em seres humanos, sendo a maioria zoonótica. São mantidos em ciclo de transmissão entre artrópodes (vetores) e reservatórios vertebrados como principais hospedeiros amplificadores (9).*

Dentro destas espécies temos o vírus da dengue (DENV), vírus da Zika (ZIKV) e o vírus da chikungunya (CHIKV). Com o passar do tempo os arbovírus vem se espalhando pelo planeta, a ponto de causar preocupação mundial, tendo como consequência a necessidade medidas preventivas no combate a proliferação das arboviroses. Vários fatores dificultam o controle das doenças e fortalecem a transmissão viral, principalmente em regiões tropicais e subtropicais, em razão de desmatamentos, emissão de gases e sistema sanitário deficiente (9, 2). As regiões endêmicas têm dificuldades até na doação de sangue pela possibilidade de transmitir

o vírus. Segundo Lopes, Nozawa e Linhares (9), o continente Antártico é o único que não é afetado, mas com o processo de aquecimento global e evolução das arboviroses o futuro se torna incerto.

2.1.1 Vírus Dengue

Existem quatro sorotipos de DENV, categorizados como: vírus da dengue - sorotipo 1 (DENV-1), vírus da dengue - sorotipo 2 (DENV-2), vírus da dengue - sorotipo 3 (DENV-3) e vírus da dengue - sorotipo 4 (DENV-4). Há uma grande variedade de doenças causadas pelos DENVs, que vão desde infecção assintomática, febre da dengue (DF) à febre hemorrágica da dengue (DHF) (10), sendo a última um caso de alto risco. A origem dos quatro sorotipos é incerta, mas as histórias contadas são semelhantes entre todos, com origem em primatas na região da Ásia sendo fortemente apoiada por evidências ecológicas e filogenéticas (11). Contudo, estudos recentes sugerem origem africana do progenitor *Flavivirus*.

O gênero *Flavivirus* é oriundo da família *Flaviviridae* que ramificou em quatro subgrupos: 1. Os vírus específicos de insetos são isolados apenas de várias espécies de mosquitos; 2. Os vírus de vertebrados que não têm vetor artrópode conhecido, isolados apenas de roedores e morcegos; 3. Os vírus são transmitidos por mosquitos; 4. Os vírus transmitidos por carrapatos (10). É possível acreditar que o *Flavivirus* tenha origem em carrapatos e mosquitos. Porém, divergiu ao entrar em contato com diversos tipos de hospedeiros.

O primeiro caso de isolamento do vírus ocorreu em 1943 por Hotta e Kimura (12). Outros casos semelhantes foram identificados em soldados americanos nas regiões da Índia, Nova Guiné e Havaí em 1944 (13). Visto como antigenicamente semelhante, o vírus foi conhecido como DENV-1 e outros casos antigenicamente distintos de DENV-2 (10). A origem dos outros dois sorotipos só foi reconhecida anos depois em uma epidemia em Manila, nas Filipinas, em 1956 (14).

Os autores Amudhan Murugesan e Mythreyee Manoharan (10) defendem que a origem do vírus seja do continente africano pelos seguintes fatos: 1. Não existia mosquitos *Aedes* na América, mas existiam numerosas espécies do mesmo subgênero na Etiópia e regiões orientais; 2. O mosquito transmissor do *Aedes Aegypti* tem origem nas florestas africanas. O processo de proliferação para o meio doméstico foi em aldeias africanas ocidentais por meio dos armazenamentos de águas, suficiente para a evolução antes do processo de tráfico dos escravos

africanos para o novo mundo. O transmissor secundário é o mosquito *Aedes Albopictus* apontado como a principal causa epidêmica na Ásia durante a Segunda Guerra Mundial.

Em 1952, houve um grande sucesso na contenção do vírus no Brasil com inseticidas extinguindo a existência do vírus por um breve período (10). Com o relaxamento das medidas adotadas para o gênero dos *Flavivirus*, a DENV voltou com força em diversas regiões nacionais. Entre 1990 e 2000, os sorotipos DENV-1 e DENV-2 causaram sérios problemas nos centros urbanos do Sudeste e Nordeste, levando a uma série de internações, com os mais variados sintomas causados pelas doenças da DENV nos infectados (15). Desde então, em dados mais atuais, os quatro sorotipos circulam no Brasil, necessitando de um o trabalho contínuo para conter a epidemia.

2.1.2 Vírus Zika

O ZIKV pertence à família *Flaviridae* do gênero *Flavivirus* filogeneticamente ligado com DENV, teve sua origem descoberta na floresta Zika no Uganda em 1947 (16). Na mesma região, em 1952, houve os primeiros casos de infecção em humanos, no ano seguinte foram detectados novos casos na Nigéria. Com o passar dos anos o arbovírus foi se espalhando pelo continente africano, reconhecido entre os anos de 1975 a 1977 nas localidades de Serra Leoa, Nigéria, Senegal, Gabão, Costa do Marfim e em países da África Central (17).

Na Indonésia, entre 1977 e 1978, surgiram os primeiros casos do ZIKV fora do continente africano, internando vários moradores com sintomas de febre aguda. O pior caso ocorreu na Micronésia, afetando 75% da população nas primeiras crises epidêmicas conhecidas do arbovírus (18). A transmissão fugiu do controle ao chegar nas ilhas do Oceano Pacífico e rapidamente, em 2013, ocorreu a epidemia na Polinésia ao gerar 19.000 indivíduos suspeitos e 284 casos confirmados (17). Logo, casos começaram a surgir na América, o Chile foi o primeiro país afetado pelo vírus em 2014, em virtude da distância das ilhas afetadas no Oceano Pacífico.

Em 2015, o Brasil teve seus primeiros casos confirmados nos estados do Rio Grande do Norte e na Bahia (16). Não demorou muito para que o número alarmante de casos ocorresse em um curto período, levantando a possibilidade de uma disseminação silenciosa pelos infectados assintomáticos, logo, veio a ser considerado como o principal meio de transmissão do vírus os mosquitos *Aedes Aegypti* e *Aedes Albopictus*. Zhao e Musa (16) apontam também outras formas de contaminação por meio de materno-fetal, relação sexual e via transfusão de sangue.

A falta de medicamentos e vacinas tornaram a ZIKV uma preocupação mundial declarada pela Organização Mundial da Saúde (OMS) em 2016. Ocorrências de microcefalia durante o mesmo período foram descobertas em gestantes contaminadas pelo ZIKV no Brasil. Várias pesquisas foram iniciadas para entender a gravidade do vírus no corpo das pessoas gestantes, descobrindo novas relações com uma síndrome congênita e da síndrome de Guillain-Barré (19, 20).

2.1.3 Vírus Chikungunya

O vírus Chikungunya é um alfavírus da família Togaviridae, transmitido principalmente aos humanos por mosquitos *Aedes Aegypti* e *Aedes Albopictus*. O primeiro caso de isolamento do vírus ocorreu em 1953, na Tanzânia, com sintomas de febre no indivíduo. Na mesma época, estava ocorrendo o surto de uma doença caracterizada por causar dores nas articulações e febre alta, localmente conhecida como chikungunya (21). Novos surtos foram registrados entre as décadas de 1960 e 1980 no continente africano e asiático. Após um período de 20 anos em esquecimento, o CHIKV ressurgiu em um novo surto no Congo, em 2000 (22). Esses registros do vírus não foram considerados um problema de saúde pública até 2004, quando provocaram surtos explosivos que afetaram progressivamente o mundo pela África Oriental no Quênia, Ilhas do Oceano Índico e Índia.

Países do continente europeu também foram afetados, como a Itália em 2007. Logo após, no ano de 2010, foi a vez da França, bem como houve relatos de casos da CHIKV no Sul da China, Península Arábica e Nova Caledônia no Oceano Pacífico, mostrando a força epidêmica deste arbovírus (23, 24). Devido à proximidade das regiões e dado o contexto histórico, casos virais nas Ilhas do Oceano Pacífico deixaram o continente Americano em alerta, onde os mosquitos *Aedes*, que são os principais transmissores, já estão estabelecidos. A esse respeito, Silva et al. (25) registra:

> *A importância da infecção nas Américas foi destacada em dezembro de 2013, após a Organização Pan-Americana da Saúde (OPAS) publicar um alerta epidemiológico sobre as evidências dos primeiros casos autóctones da doença. Até a 52ª semana epidemiológica (SE) do ano posterior, 2014, foram notificados 1.071.696 casos suspeitos da doença em mais de 30 países do continente americano, a exemplo de México, El Salvador, Nicarágua, Guiana Francesa, Porto Rico, Colômbia, Venezuela, Brasil e Suriname, entre outros, com 169 óbitos atribuídos à chikungunya (25).*

Esses autores informam que os casos no Brasil ocorreram pioneiramente nos estados do Amapá e Bahia, identificando duas linhagens do arbovírus depois de uma série de exames. O clima tropical do Brasil tornam o país um lugar propício para disseminação de arboviroses, onde mais de 4.000 municípios registram ocorrências do *Aedes aegypti* e outros 3.285 do *Aedes Albopictus* (26, 27).

2.1.4 A similaridade dos sintomas das doenças arbovirais

A sintomatologia das doenças arbovirais tem muitos fatores em comum que confundem nos estágios iniciais aos profissionais da saúde. Essa característica, dificulta a análise mais superficial baseadas apenas em perguntas sobre os sintomas do paciente. Países como o Brasil que têm uma variedade de arboviroses deixam o processo ainda mais delicado, tornando essencial que exames sejam feitos para garantir o diagnóstico médico. Isso porém, envolve altos custos e planejamento para garantir que todo território tenha insumos para combater os arbovírus. Nos próximos parágrafos deste capítulo será abordado a similaridade entre DENV, CHIKV e ZIKV.

Os sintomas causados pela doença do vírus da zika podem variar de adulto para criança, segundo Musso e Gubler (28), os sintomas são: erupção cutânea, febre baixa, artralgia, mialgia e conjuntivite. Os casos de infecção por ZIKV assintomáticos são um problema grave para o controle de epidemia, pois trabalham silenciosamente na sua proliferação e dificultando o combate ao vírus. As preocupações vão além, existem complicações neurológicas encontradas pelo neurotropismo do vírus como meningoencefalite, mielite ou Síndrome de Guillain-Barré (19, 20).

Na fase sintomática, a identificação do vírus é feita por exame de sangue, contudo o material genético do ZIKV pode ser encontrado por mais de 10 dias na urina e durante a gravidez pode permanecer até 3 vezes mais do que o esperado. A situação foi analisada por Karkhah et al. e Meaney-Delman et al. (29, 30), em mulheres grávidas entre 18 e 39 anos, em Porto Rico. *Reverse Transcription PCR* (RT-PCR) é um diagnóstico laboratorial para análise de doenças virais que permite identificar material genético de amostras colhidas pelos profissionais da saúde, no entanto, em casos epidêmicos se torna custoso suprir a necessidade de várias localidades com esse tipo de exame e o resultado é liberado após dias.

A doença da dengue tem muitos casos com impacto leve no corpo humano, que pode

ser considerado pelos sintomas de febre, náuseas, vômitos, exantema, cefaleia, dor retro orbitária, artralgia, mialgia e leucopenia. Os sintomas mais severos podem causar danos ao corpo humano, tais como dor ou sensibilidade abdominal, vômitos persistentes, acúmulo de líquido clínico (derrame pleural, ascite, espessamento da parede da vesícula biliar), sangramento de mucosa (epistaxe, sangramento gengival, sangramento gastrointestinal, hematúria, sangramento vaginal, sangramento da pele), letargia ou inquietação, hepatomegalia maior que 2 cm e hematócrito aumentado com diminuição das plaquetas (31, 17). Vários órgãos podem ser afetados nestas situações mais graves, como coração, olhos, rins, cérebro, fígado e pâncreas, causando até a morte do indivíduo, no pior dos casos (32).

A maioria dos pacientes em áreas endêmicas com doença da dengue são assintomáticos facilitando a proliferação do vírus. Os sintomáticos têm um período entre 3 a 14 dias para sentir os primeiros sinais dos sintomas provocados pela infecção (32). Tornando a situação preocupante, pois pessoas que transitam entre regiões para trabalho ou viagem podem levar o vírus para outras áreas sem terem conhecimento da situação.

A doença do chikungunya tem sintomas bem similares aos da dengue e zika, que se desenvolvem nos infectados depois de 2 a 6 dias da picada do mosquito. Sintomas como febre alta, artralgias, dor nas costas, cefaleia, fadiga intensa, anorexia, mialgias, náuseas e vômitos (33). O reumatismo agudo bilateral e simétrico é tipicamente extenso e progressivo em poucos dias (23, 34). As articulações são frequentemente afetadas, provocando muita dor e inchaços, especialmente nas articulações interfalângicas, punhos e tornozelos.

Sintomas atípicos podem se desenvolver no corpo humano, como no sistema nervoso, cardiovascular, pele e rins. Os sintomas não relatados nas pesquisas do parágrafo acima são meningoencefalites, encefalopatia, convulsões, síndrome de Guillain-Barré, síndrome cerebelar, paresias, paralisias, neuropatias, neurite óptica, iridociclite, episclerite, retinite, uveíte, miocardite, pericardite, insuficiência cardíaca, arritmia, instabilidade hemodinâmica, hiperpigmentação por fotossensibilidade, dermatoses vesiculobolhosas, nefrite, insuficiência renal aguda, discrasia sanguínea, pneumonia, insuficiência respiratória, hepatite, pancreatite, síndrome da secreção inapropriada do hormônio antidiurético e insuficiência adrenal

A incidência de mais de uma doença causada por arboviroses pode ser um caso complicado para análise de prognósticos médicos e controle de epidemias, o custo para manter exames laboratoriais com RT-PCR são altos e levam dias para emitir o resultado. Existe

similaridade dos sintomas dessas doenças que dificultam o prognóstico médico, em razão do comportamento dos vírus no corpo ao longo dos dias. É complexo se basear somente nas informações dadas e observadas nos pacientes para conduzi-lo ao melhor tratamento dos sintomas. A falta de medicamentos adequados para tratamento é um fator impactante para mitigar os danos causados.

2.2 APRENDIZADO DE MÁQUINA

Nesta época em que cada vez mais pessoas são adeptas da tecnologia, grandes quantidades de dados são gerados com objetivos distintos, dentre deles centralizar e facilitar o acesso aos dados de forma global. O aprendizado de máquina foi criado para emular o trabalho árduo de aprendizado humano de extrair conhecimento através da análise de dados (35). As técnicas de aprendizado de máquina são aplicadas em variadas áreas, entre reconhecimento de padrões, visão computacional, engenharia de naves espaciais, finanças, entretenimento e biologia computacional até aplicações biomédicas e médicas (35, 36).

Para Dizo (36), os algoritmos de aprendizado de máquina têm direcionado atenção para as questões médicas nos últimos anos em razão de compreender padrões de dados volumosos com desempenho de classificação precisa. A aplicação desses algoritmos em problemas relacionados às arboviroses, busca compreender fatores significativos para a proliferação, tais como os aspectos climáticos, socioeconômicos e fisiológicos, e vem crescendo nos últimos anos. Os estudos com objetivo de prever epidemias em áreas endêmicas é conduzido em menor escopo por diferentes regiões; é muito comum ver dados diferentes serem associados para validar e gerar novos conhecimentos nas pesquisas. Os algoritmos de aprendizado de máquinas são utilizados isoladamente, ou com propósito de comparação entre eles, para avaliar a melhor precisão no contexto definido.

Na construção de um modelo de aprendizado de máquina existem diferentes abordagens para o aprendizado. As quatro abordagens que compõem o ecossistema do aprendizado de máquina são concebida para resolver diferentes problemas, são eles: aprendizado supervisionado, aprendizado semi-supervisionado, aprendizado não supervisionado e aprendizado por reforço. A escolha de cada tipo varia de acordo com o problema no qual está tentando resolver.

O processo de aprendizado supervisionado é baseado no conhecimento de padrões, com três aspectos importantes: 1. Os dados de entrada para realizar a tarefa desejada; 2. Processo

repetitivo chamado treinamento, em que é adquirida experiência por parte do algoritmo; e 3. Generalizar para produzir o resultado desejado a partir de dados novos e não vistos anteriormente (35). Os dados de entrada são decisivos para o aprendizado do algoritmo, sendo capaz de ser examinado e selecionado. Mesmo que os dados iniciais gerem baixa precisão, novos dados selecionados podem ser inseridos para que o algoritmo continue evoluindo no aprendizado durante processo de treinamento de forma supervisionada.

Aprendizado de máquina não supervisionado se baseia nos resultados de cada etapa do treinamento como calibração dos resultados. O algoritmo não recebe uma configuração para cada entrada associando a um resultado específico, onde encontra seu próprio caminho. Por último, temos o aprendizado de máquina semi supervisionados, em que parte dos dados são rotulados e outra não. Os dados rotulados podem ser utilizados para auxiliar no aprendizado dos dados não rotulados (35).

2.2.1 Classificação

Na aplicação do aprendizado de máquina supervisionado, na qual a saída é rotulada, são utilizados algoritmos de classificação e regressão. O algoritmo de classificação permite lidar com semelhanças e diferenças quando os objetos a serem classificados possuem muitas características dentro de sua própria classe, mas ainda possuem fatores fundamentais que os identificam. Diferente do algoritmos de classificação, os algoritmos de regressão não buscam classificar os itens, o objetivo é fazer uma regressão para tentar prever um número, como por exemplo o número de casos de DENV.

Os classificadores são divididos em probabilísticos e em lineares. O classificador *Naive Bayes* (NB), um exemplo amplamente utilizado em aprendizado de máquina, faz cálculo da possibilidade posterior de uma classe. Os classificadores lineares agrupam os itens que possuem os mesmos valores, um dos algoritmos populares se chama *Random Forest* (37).

2.3 *RANDOM FOREST*

RF é um algoritmo de aprendizado de máquina supervisionado concebido por Breiman (7), que combina a simplicidade de um conjunto de árvores de decisão com a flexibilidade e aleatoriedade para alcançar um único resultado com a melhor precisão para problemas de classificação e regressão. O fator aleatoriedade é importante para resolver um dos maiores

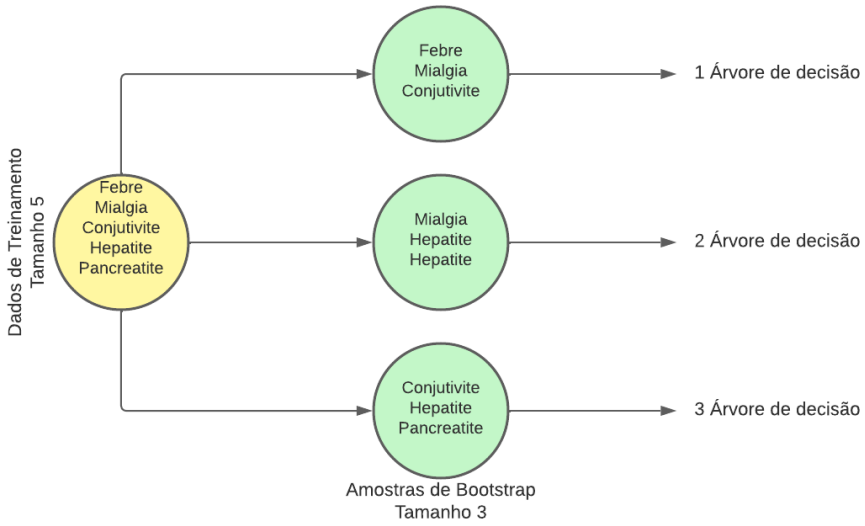
problemas das árvores com o *overfftting*. Como a RF usa alto número de árvores de decisão, não sofrerá *overfftting* pelo cálculo médio das árvores não correlacionadas, reduzindo a variação geral.

No estudo de Breiman sobre o artigo de seleção aleatória elaborado por Amit e Geman em 1997 (38), o autor apresenta a ideia para o método de *bagging* que constrói subconjuntos aleatórios de dados com propósito de criar uma floresta não correlacionada de árvores de decisão. RF sendo desenvolvida para ser concorrente *boosting* tornou-se uma extensão do método de *bagging* (39, 7). A diferença principal entre árvores de decisão e o RF é que a primeira considera todas as divisões de dados possíveis e a segunda utiliza somente um subconjunto aleatório desses dados.

Para utilizar o algoritmo de forma correta, existem três hiperparâmetros que devem ser configurados antes do processo treinamento, são eles: tamanho do nó, número de árvores e o número de recursos amostrados. Após definidos, o *Bootstrap Dataset* é a fase inicial do algoritmo para criar um subconjunto de dados com base nos originais de forma aleatória, somente em algumas amostras e não em sua totalidade. O processo consiste na seleção dos dados aleatoriamente para prover a diversificação reduzindo a correlação. Por esse fator, também pode ser selecionado os mesmos dados mais de uma vez durante o processo (7), como ilustrado na figura 1. O processo é repetido até se encaixar nas três condições de paradas: 1. Selecione m preditores aleatoriamente dos p preditores disponíveis; 2. Encontre a melhor divisão binária entre todas as divisões binárias nos m preditores da etapa 1; e 3. Divida o nó em dois nós descendentes usando a divisão da etapa 2 (39).

Depois da etapa de *Bootstrap Dataset*, inicia-se o processo de seleção de características (*features*) aleatoriamente para criação das árvores. Diferente das árvores de decisão, a RF irá escolher de maneira aleatória, duas ou mais *features* para definir qual delas será utilizada no primeiro nó. O processo se repete para a seleção do próximo nó, excluindo as *features* escolhidas anteriormente até finalizar as possibilidades de escolha. Mesmo não sendo considerada a melhor estratégia de construção de árvores, pois podem ser escolhidas às duas piores *features* nos primeiros nós. Entretanto, como serão criadas várias árvores, essa estratégia tende a evitar o *overfitting* que pode ocorrer. Todo esse processo será repetido para a criação das próximas árvores de decisão, baseado na quantidade estabelecida dos seus hiperparâmetros, citado no parágrafo anterior.

Figura 1 – Geração de amostras *bootstrap*



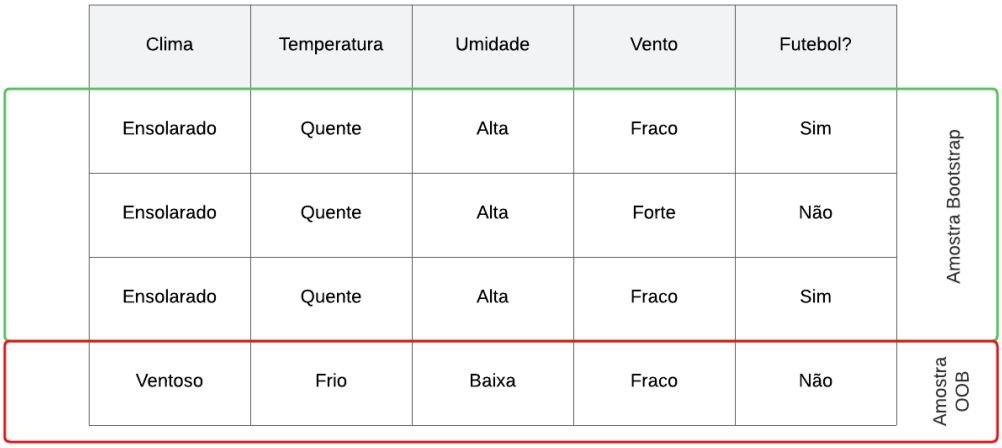
Fonte: Elaborado pelo autor

Durante a divisão das amostras ocorre o processo conhecido como *out-of-bag* (OOB), que reserva um terço dos registros. Para melhor entendimento, será descrito um exemplo, onde existe um conjunto de dados de treinamento com quatro registros. A primeira amostra *bootstrap* é selecionada nos três primeiros registros para gerar a primeira árvore de decisão. O quarto registro não será selecionado para criar uma amostra OOB, no exemplo somente um registro ficou de fora, mas é possível ter outros com um conjunto de dados maior, para o melhor entendimento pode ser observada a Figura 2. Após o treinamento da árvore de decisão, a amostra de dados não observados no OOB será utilizada para validação cruzada do modelo gerado. Todos os resultados dos modelos são considerados na floresta de árvores, criada baseando-se em duas estratégias: para casos de regressão será calculada a média das árvores de decisão individuais, para uma tarefa de classificação será realizada uma votação majoritária.

2.4 TRABALHOS CORRELATOS

Dey e Mukhopadhyay (40) se propuseram a analisar diferentes algoritmos para prever os casos de dengue com interação em proteínas humanas, manuseando algoritmos de aprendizagem supervisionados como *Support Vector Machine* (SVM), RF, NB e *K-nearest neighbors* (KNN). Ao avaliar os quatro algoritmos, a RF teve resultados melhores em acurácia, *recall*, F1-score enquanto não desempenhou tão bem em especificidade e precisão para a SVM.

Figura 2 – Processo de seleção das amostras



Fonte: Elaborado pelo autor

Entretanto, Jiang et al. (41) utilizam dados de fatores externos para ter a possibilidade de gerar um modelo mais eficiente, os dados climáticos que em estudos anteriores mostraram uma relação muito forte na proliferação dos mosquitos. Foram acrescentados dados de características ambientais como umidade, geolocalização e socioeconômico que afetam as condições de moradia da população. Jiang et al. (41) usam também *Back-Propagation Neural Networks* (BPNN), *Gradient Boosting Machine* (GBM) e RF na análise espacial de 5x5km para prever risco de transmissão do ZIKV. O desempenho obtido pela RF esteve bem próximo dos resultados dos outros algoritmos avaliados com a métrica *area under the ROC curve* (AUC) de 0,963 sendo que a BPNN obteve 0,966.

Uma análise global por Mudele et al. (42) para relacionar fatores ambientais, incluindo temperatura, condição da vegetação, umidade e precipitação na transmissão de arbovírus, uma melhora nos resultados devido à quantidade diferentes de dados disponíveis na análise. Nesse caso, RF foi a base da pesquisa utilizando outros algoritmos, tais como *Linear Regression Model* (LM), *Generalized Linear Model* (GLM), *Artificial Neural Networks* (ANN), *Support Vector Regression* (SVR), KNN e *Decision Trees Regression* (DTR), visando realizar uma análise comparativa para provar a robustez em analisar relações complexas e demonstrar sua capacidade preditiva como uma boa abordagem para o estudo.

Em comparação com outros algoritmos como SVM, *Neural Networks* (NN) e também com métodos estatísticos, a RF apresenta resultados relevantes, apontando-a como uma boa

opção para criar um modelo preditivo. Yin et al. (43) conseguiram obter resultados melhores que a NN. Nesta pesquisa a RF mostra resultados consistentes para criar modelos preditivos voltados aos arbovírus e de forma secundária podemos ver indícios que ela pode ser uma boa opção para algoritmos como a NN, mesmo tendo pequenas variações nos estudos analisados. Entretanto, em determinados casos não se mostra tão precisa quanto como pode ser observado por Jiang et al. (41), que a BPNN teve um melhor resultado e Huang et al. (44), que obteve melhores resultados com a ANN.

Observou-se que a ANN geralmente tem vantagem sobre os demais algoritmos, pois depende de fatores como o conjunto de dados e os ajustes feitos para extrair o máximo do potencial do algoritmo. Fahmi et al. (45) analisou que, enquanto a RF utiliza ajustes em dois hiperparâmetros, a ANN usa 100 neurônios, ativação com Relu, solucionador com Adam, alfa em 0,0001 e um número máximo de 200 interações, sendo o custo de implantação e execução para RF mais baixo e com resultados próximos. Todavia, o autor conclui que a ANN é a melhor opção, os estudos coletados não têm indicações explícitas da substituição da ANN pela RF, tendo foco mais na comparação para utilizar a que obtém o melhor resultado

Os resultados para comparações são efetuados por meio das métricas de desempenho coletadas, todos os estudos conduzem ao uso destas métricas de modo a comparar ou somente validar o modelo criado por estes algoritmos, tornando uma análise quantitativa. As métricas mais vistas são de acurácia para previsão geral, precisão que é métrica de um único teste, *recall* é o número de previsões corretas recuperadas (46), bases de quase todos os artigos, pois validam a precisão do modelo preditivo concisamente e simples. Entretanto, outras métricas mais complexas são utilizadas como complemento da avaliação de desempenho.

A *F–measure* ou *F–score* é a média harmônica de precisão, considerado valores perfeitos ao atingir o valor 1, baseado no cálculo de precisão e *recall* (47) utilizada para avaliar os resultados obtidos. Previamente, Pushphavathi, Suma e Ramaswamy (48) utilizam AUC e *Receiver Operating Characteristic* (ROC) para avaliar o desempenho dos classificadores. Existem outras métricas utilizadas, mas aparecem com menor frequência durante a extração dos dados desta revisão literária, tais como *Cohen's Kappa Coefficient* (Kappa) (40) que mede a concordância entre a classificação e os valores de verdade, *Matthews Cohen Correlation* (MCC) utilizada como uma medida de qualidade (46), utilizadas para objetivos específicos de cada estudo para complementar a análise.

3 METODOLOGIA DE PESQUISA

A metodologia selecionada para este estudo foi o DSR (49) porque fundamenta o desenvolvimento de artefatos como meio para a produção de conhecimentos científicos do ponto de vista epistemológico (50), em que o artefato é criado para resolver algum problema real no contexto natural e social.

O pesquisador tem dois objetivos: primeiro resolver um problema prático no contexto específico do artefato e segundo gerar um novo conhecimento, criando dois ciclos de pesquisa inter-relacionados chamados Ciclo de Design e Ciclo de Engenharia, que indicam que as conjecturas teóricas servem como alicerce para a construção do artefato, servindo para validar as conjecturas (50). Exposto um problema, o artefato é desenvolvido para afirmar ou colocar em dúvida as conjecturas teóricas que geram um ciclo de melhoria/direcionamento, conforme apresentado na Figura 3.

Figura 3 – Ciclos em Design Science Research

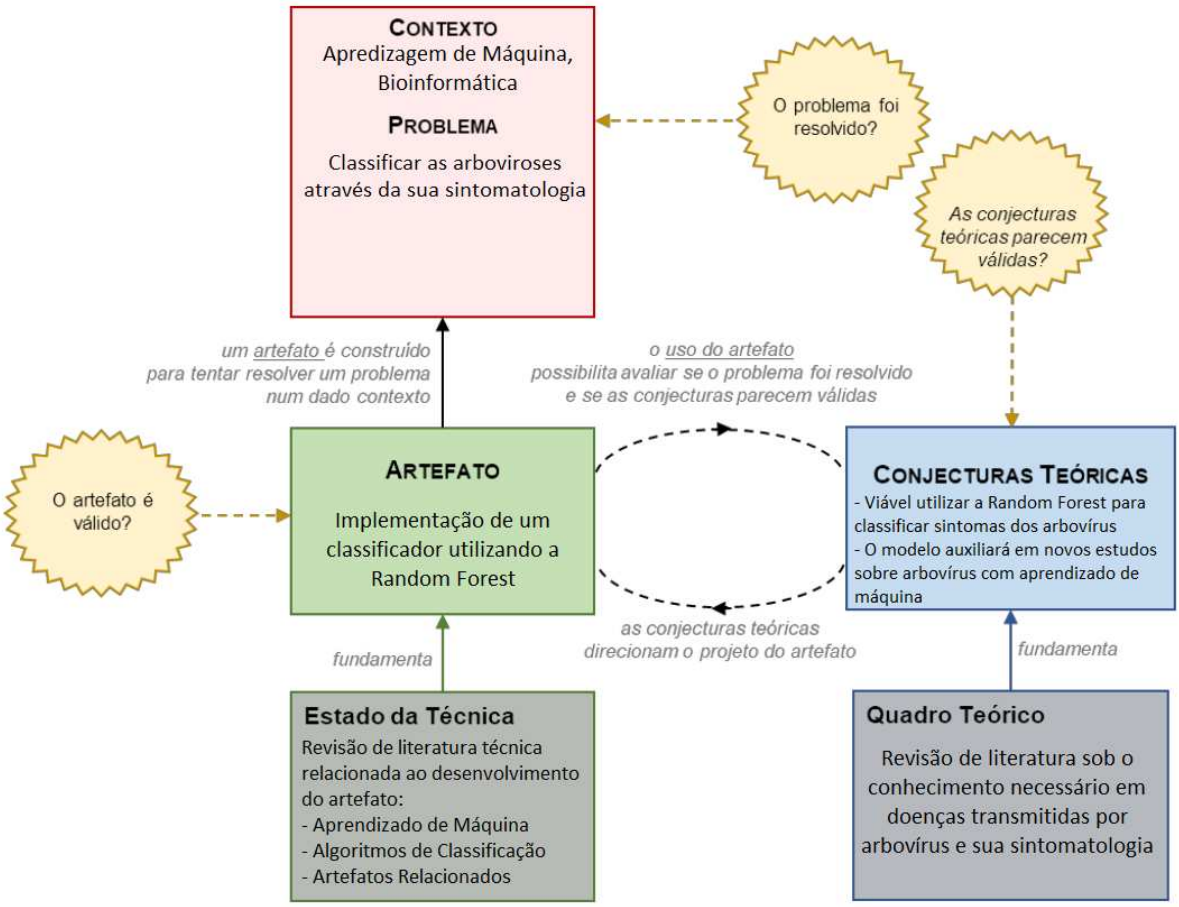
Fonte: Retirada de Pimentel (49)

O Ciclo de Engenharia tem como principais características as avaliações do artefato, buscando melhorias e refinamento do projeto (50), ao seguir às cinco etapas: investigação do problema, design da solução, validação da solução, implementação da solução e avaliação da implantação.

O Ciclo Empírico baseia-se em teorias e métodos científicos para garantir que a condução da pesquisa seja realizada a rigor teórico e metodológico (50) em sete etapas: contexto da pesquisa, análise de problema de pesquisa, pesquisa e inferência do design, validação, execução da pesquisa, análise de dados e contribuição da pesquisa.

Por último, o Ciclo de Relevância relaciona o contexto ao artefato projetado para atingir o objetivo da pesquisa levando em consideração o ambiente, pessoas, problemas e oportunidades. Desta maneira, identificamos ser necessário realizar três avaliações em pesquisas concebidas no paradigma da DSR na figura 4: se o artefato satisfaz aos requisitos; se o problema foi resolvido satisfatoriamente; e se as conjecturas teóricas parecem válidas (50). A Figura 4 consiste no modelo DSR adaptado para o projeto de pesquisa.

Figura 4 – Diagrama da pesquisa



Fonte: Adaptado de Pimentel (50)

Com base no diagrama, definiu-se o objetivo de criar um classificador através dos dados clínicos baseado nas conjecturas teóricas da área de epidemiologia molecular e nas técnicas

para a criação do artefato. Será desenvolvido de forma cíclica adaptativa, de modo a testar, validar e corrigir quando for necessário, até atingir o objetivo em produzir conhecimento suficiente para suprir a lacuna existente.

4 DESENVOLVIMENTO DO CLASSIFICADOR

Neste capítulo será abordado o problema de pesquisa, assim como a arquitetura proposta, desenvolvimento, treinamento do algoritmo, testes e avaliação dos resultados detalhadamente. Todo desenvolvimento do artefato é abordado na seção 4.1, baseado em três ciclos, onde são apresentadas as mudanças a nível de dados, código e avaliação das métricas de desempenho. O intuito é, analisar o algoritmo RF a cada ciclo, no qual diferentes pontos no desenvolvimento do artefato podem impactar nas métricas de desempenho.

4.1 DESENVOLVIMENTO

O processo será baseado na metodologia DSR citada anteriormente, aplicada com ciclos de melhorias para avaliar o comportamento do RF durante o período de desenvolvimento do classificador. O foco do trabalho está no algoritmo do RF por sua crescente popularidade entre os algoritmos de aprendizado de máquina e pela característica fortemente inserida com aleatoriedade dos seus processos de utilização dos dados, bem como o conjunto de árvores para resolver os problemas de classificação e regressão.

Nos últimos cinco anos, houve um aumento significativo nas pesquisas envolvendo as arboviroses e algoritmos de aprendizado de máquina. A diversidade de dados utilizados e algoritmos é muito grande nessas pesquisas. A utilização do RF tem se tornado mais frequente quando o assunto é relacionado a DENV. Neste estudo, o objetivo principal será a avaliação do algoritmo fundamentado em suas métricas de desempenho. O treinamento do modelo foi aplicado no conjunto de dados sintomatológicos de pacientes infectados por arbovírus, oriundos do Departamento de Informática do Sistema Único de Saúde (DATASUS), como visto na subseção 2.1.4.

4.1.1 Ambiente de desenvolvimento

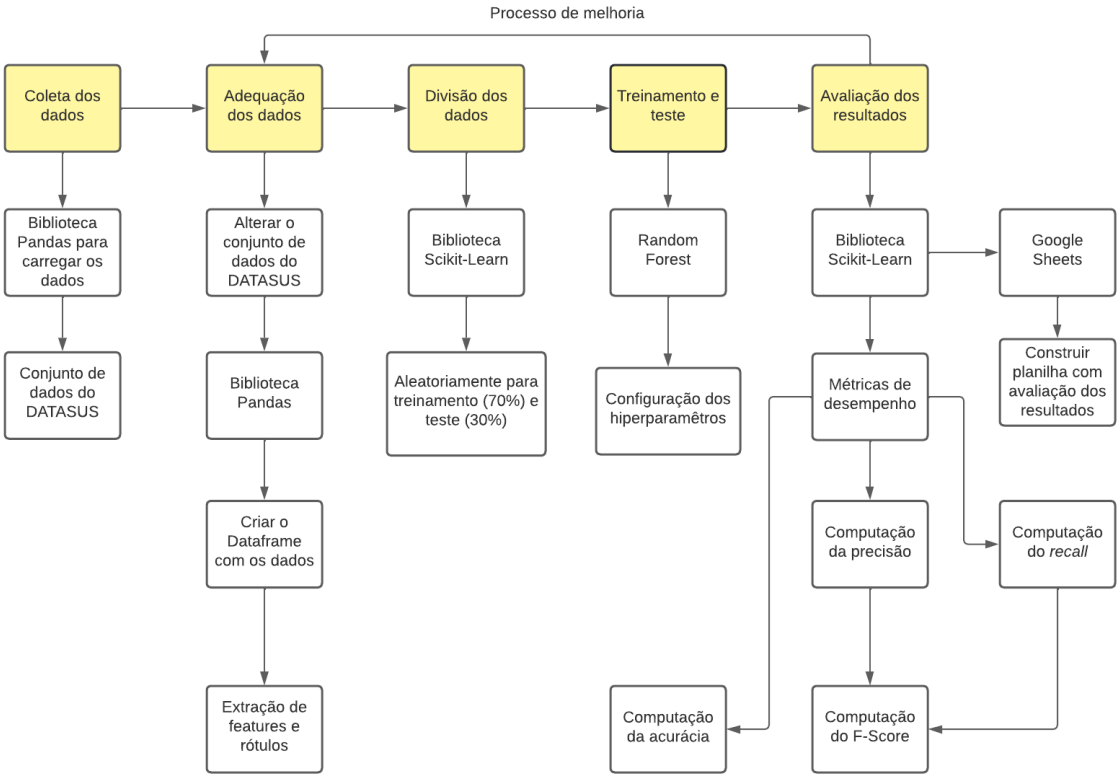
Todo o processo de desenvolvimento foi elaborado no sistema operacional *Windows*. A linguagem base utilizada foi *Python*, pelo seu conjunto extensivo de bibliotecas, como *Scikit-Learn* que contém vários algoritmos de aprendizado de máquina, incluindo o RF. A biblioteca

também permite coletar métricas de desempenho utilizadas na avaliação dos resultados. Uma outra biblioteca importante para este trabalho foi *Pandas*, que é uma ferramenta de análise e manipulação de dados importante no processo de transformação dos dados obtidos para o modelo de *DataFrame*, para ser manipulado pelo algoritmo do RF. Por último para construir a representação gráfica foi utilizada a biblioteca *Matplotlib*, com intuito de facilitar o entendimento sobre os dados apresentados.

4.1.2 Implementação da arquitetura proposta

A proposta de arquitetura foi planejada com cinco etapas: coleta, adequação, divisão, treinamento/teste e avaliação dos resultados. São etapas fundamentais para alcançar o resultado desejado, de analisar o algoritmo com base nas métricas de desempenho. Para elucidar, a Figura 5 apresenta todas as etapas e subetapas. As próximas subseções abordam cada uma detalhadamente.

Figura 5 – Arquitetura do problema



Fonte: Elaborado pelo autor

4.1.2.1 Primeiro ciclo

A estrutura base para o desenvolvimento dos próximos experimentos estão nesta subseção, será abordado desde os desafios iniciais como a escolha do conjunto de dados é importante até as avaliações das métricas de desempenho para direcionar o próximo ciclo.

4.1.2.1.1 Coleta dos dados

Os dados constituem a parte fundamental para o desenvolvimento do classificador, a quantidade e a qualidade deles são fatores críticos. Para Thomas et al. (51) um *dataset* construído e normalizado com especialista da área médica, tem o proposito de entregar dados com qualidade para serem utilizados em estudos. Baseado nisto os dados deste trabalho foram obtidos de duas bases de dados entre os anos de 2015 a 2020 pelo Sistema de Informação de Agravo de Notificação (SINAN) do estado do Amazonas e do portal de dados abertos de Recife que utilizam como fonte de dados o DATASUS, estes estados disponibilizaram os dados que foram devidamente depurados por especialistas médicos.

Durante a normalização dos *datasets* foram descartados os registros com as características de não relacionar sinais ou sintomas e *features* com as variáveis com mais de 50% de dados ausentes também foram removidas (51). O *dataset* consolidado apresenta 40.376 registros com 27 variáveis de informações clínicas e sociodemográficas de pacientes confirmados de DENV e CHIKV, assim como pacientes descartados dessas arboviroses. O número maior de casos "não confirmados" poderia causar desequilíbrio e enviesar a geração do modelo de aprendizado de máquina. Thomas (51) aplicou uma técnica de subamostragem aleatória para balancear a quantidade de registros entre as classes, resultando em 17.172 registros, sendo 5.724 para cada uma das três classes.

Existe escassez de dados clínicos de pacientes reais com sintomas do ZIKV, as fontes de dados já coletadas não tem informações que confirmem casos de infecção por ZIKV. Iniciaram-se novas buscas para obter os dados necessários para ser integrado ao *dataset* existente, as informações no Departamento de Informática do DATASUS, SINAN e no portal de dados abertos de Recife não contém as informações sobre os sintomas, somente informações com características demográficas e biológicas. Os dados relacionados ao ZIKV são fornecidos pelo DATASUS, tentando simular o mesmo processo na extração dos dados feito por Thomas (51)

com a ferramenta *Tab para Windows* (TABWIN). Foi necessário inserir o arquivo na ferramenta para visualizar as informações referente a ZIKV, mas infelizmente não foi possível encontrar nada relacionado aos sintomas.

4.1.2.1.2 Adequação dos dados

O propósito principal é treinar o modelo com as *features* relacionadas aos sintomas, o *dataset* inicial contém 27 *features* nas quais envolvem dados demográficos, biológicos e doenças pré-existentes. As *features* são: idade do paciente, sexo, tempo de gestação, raça, área onde mora, febre, mialgia, cefaleia, exantema, vômito, náusea, dor nas costas, conjuntivite, artrite, artralgia, petéquia, dor retro orbitária, diabetes, doenças hematológicas, doenças do fígado, doenças renais, hipertensão, doença do ácido péptico, doença autoimune, número de dias em que o paciente está apresentando sintomas e classificação final do paciente. Um novo *dataset* foi gerado a partir das *features* relacionadas aos sintomas febre, mialgia, cefaleia, exantema, vômito, náusea, dor nas costas, conjuntivite, artrite, artralgia, petéquia e dor retro orbitária conforme apresentado na tabela 1. Os valores das colunas relacionadas aos sintomas representam somente duas possibilidades: o número 1 (um) é representado como "SIM" para sinalizar a presença do sintoma e o número 2 (dois) como "NÃO" para indicar que o paciente não apresentou o sintoma. Os valores da coluna de classificação final (CLASSI FIN) foram alterados para números inteiros para padronizar todo o *dataset*, o valores como chikungunya, dengue, zika e outras doenças são representados por 0, 1, 2 e 3 respectivamente. Para facilitar o entendimento, os valores exibidos na tabela 1 foram traduzidos para sua forma literal.

Tabela 1 – Primeira versão do *dataset*, somente os 20 primeiros registros

FEBRE	MIALGIA	CEFALEIA	EXANTEMA	VOMITO	NAUSEA	DOR_COSTAS	CONJUNTVIT	ARTRITE	ARTRALGIA	PETEQUIA_N	DOR_RETRO	CLASSI_FIN
SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	SIM	SIM	CHIKV
SIM	SIM	SIM	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	SIM	NÃO	NÃO	CHIKV
SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	CHIKV
SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	CHIKV
SIM	SIM	SIM	NÃO	SIM	NÃO	SIM	NÃO	SIM	NÃO	NÃO	SIM	CHIKV
SIM	SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	CHIKV
SIM	SIM	SIM	NÃO	NÃO	SIM	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	CHIKV
SIM	SIM	SIM	SIM	NÃO	NÃO	SIM	NÃO	NÃO	SIM	NÃO	SIM	CHIKV
SIM	SIM	SIM	NÃO	NÃO	SIM	SIM	SIM	SIM	SIM	NÃO	NÃO	CHIKV
SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	CHIKV
SIM	SIM	SIM	NÃO	NÃO	NÃO	SIM	SIM	SIM	SIM	NÃO	SIM	CHIKV
SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	CHIKV
SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	CHIKV
SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	CHIKV
SIM	SIM	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	SIM	NÃO	NÃO	CHIKV
SIM	NÃO	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	CHIKV
SIM	SIM	SIM	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	SIM	NÃO	SIM	CHIKV
SIM	NÃO	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	CHIKV
SIM	SIM	NÃO	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	CHIKV
SIM	SIM	NÃO	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	CHIKV

Fonte: Elaborado pelo autor

O *dataset* está no arquivo em formato *comma-separated values* (CSV), para carregar

os dados foi necessário utilizar a biblioteca *pandas* com sua função *read_csv* para ler os dados do arquivo. Na função *read_csv* se fez necessário passar dois argumentos: caminho para o arquivo e delimitador, depois ter os dados carregados eles são separados em dois subconjuntos: *features* e *labels*. Para esse processo foi elaborada uma função na linguagem Python declarada como *get_dataset* 1. As *features* selecionadas são passadas pelo argumento *features* na função *get_dataset* Código-fonte 1, se nenhum valor for declarado, será utilizado o valor padrão com as 12 *features* dos sintomas ilustrado na figura 1.

Código-fonte 1 – Função para carregar os dados

```
1 def get_dataset(pathname = "dataset_arbovirus.csv",
     features = DEFAULT_FEATURES, labels = ["CLASSI_FIN"]):
2     import pandas as pd
3
4     dataset = pd.read_csv(pathname, delimiter= ';')
5
6     X = dataset[features]  # Features
7     y = dataset[labels]  # Labels
8
9     return X, y
```

4.1.2.2 Divisão dos dados

O conjunto de dados coletados será dividido em 70% para treinamento e 30% para testes do classificador após o treinamento. Para a separação de forma aleatória, o *Sklean* tem uma função chamada *train_tests_plit* para auxiliar nesse procedimento. Definindo os três argumentos: *features*, *labels* (que foram separados na subseção anterior) e *test_size* definido como 0.3, que equivale a 30% da amostra. O Código-fonte 2 demonstra a construção do processo de separação dos dados.

Código-fonte 2 – Etapa de separação dos dados

```
1 import utils
2
```

```
from sklearn.model_selection import train_test_split

X, y = utils.get_dataset()

X_train, X_test, y_train, y_test = train_test_split(X, y,
    test_size=0.30)
```

4.1.2.2.1 Treinamento e teste

Depois da divisão, é realizado o treinamento do modelo com o conjunto de dados de treino e previsão do modelo com o conjunto de dados de teste. Ao importar pela primeira vez o objeto *RandomForestClassifier* da biblioteca *Sklearn*, foi definido inicialmente os seguintes hiperparâmetros: número de árvores da floresta, critério para medir a qualidade de uma divisão e número de trabalhos em paralelo, buscando melhores resultados e um menor tempo de processamento dos dados. Exibido no Código-fonte 3 os primeiros valores para os hiper parâmetros do primeiro ciclo de treinamento e teste, escolhidos de forma aleatória.

Código-fonte 3 – Função para criar as configurações do modelo

```
def get_random_forest_model(
        n_estimators=200,
        n_jobs=5,
        max_features="sqrt",
    ):
    from sklearn.ensemble import RandomForestClassifier
    return RandomForestClassifier(
            n_estimators=n_estimators,
            n_jobs=n_jobs,
            max_features=max_features,
        )
```

A função *get_random_forest_model* retorna o modelo que será treinado com os dados de treinamento da seção 4.1.2.2. O processo de treinamento durou cerca de 0.72 segundos para

ser finalizado em uma máquina com processador AMD Ryzen 5 3600 6-Core (3,95 GigaHetz) e 16 GigaBytes memória RAM (2400 MegaHetz). A etapa de validação do modelo com amostra de teste durou cerca de 0.12 segundos, as informações referentes ao tempo das etapas foram coletados com a biblioteca *time*. É possível visualizar no Código-fonte 4 a simplicidade de configurar essas etapas devido às funções *fit* e *predict* da biblioteca *RandomForestClassifier*.

Código-fonte 4 – Etapa de treinamento e teste

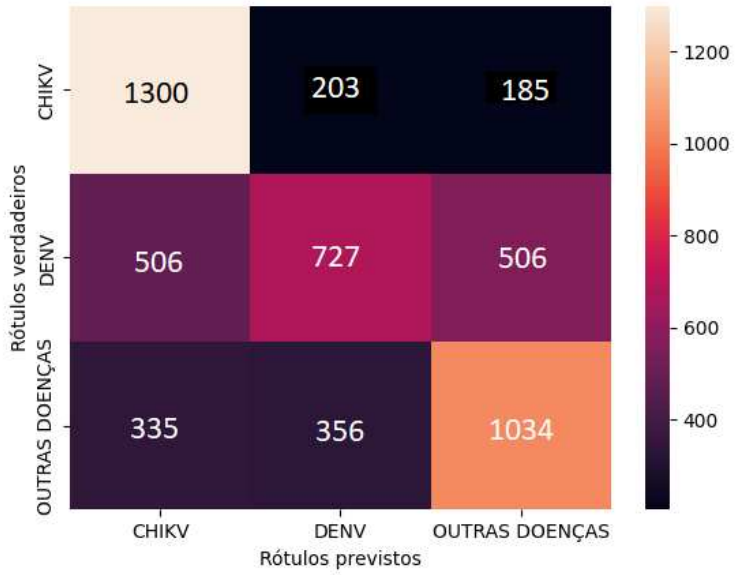
```
1  model = utils.get_random_forest_model()
2
3  start_time = time.time()
4  model.fit(X_train, y_train.values.ravel())
5  print("Training --- %s seconds ---" % (time.time() -
       start_time))
6
7
8  start_time2 = time.time()
9  y_pred = model.predict(X_test)
10 print("Test --- %s seconds ---" % (time.time() -
       start_time2))
```

4.1.2.2.2 Avaliação dos resultados

Os resultados serão avaliados utilizando a matriz de confusão com os valores obtidos. A matriz de confusão classifica seus resultados em quatro tipos, conhecidos como verdadeiro positivo (TP): ocorre quando a classificação previu corretamente, por exemplo: quando o paciente está com DENV e o modelo previu corretamente. Falso positivo (FP): quando no conjunto real foi previsto de forma errônea, por exemplo: o paciente não está com DENV, mas o modelo disse que sim. Falso verdadeiro (TN): no conjunto real, o paciente não está doente, previsto corretamente. Falso negativo (FN): casos que são verdadeiros, mas o modelo previu ao contrário, por exemplo: o paciente está com DENV, mas o modelo previu que não.

Possível visualizar o resultado da matriz de confusão 3x3 na figura 6 devido à quantidade de classes que existem no *dataset* obtém essa dimensão. Para compreender o

Figura 6 – Matriz de confusão 3x3



Fórmula para calcular a matriz de confusão

Chikungunya	Dengue	Outras doenças
TP = $Cell_1$	TP = $Cell_5$	TP = $Cell_9$
FP = $Cell_2 + Cell_3$	FP = $Cell_4 + Cell_6$	FP = $Cell_7 + Cell_8$
TN = $Cell_5 + Cell_6 + Cell_8 + Cell_9$	TN = $Cell_1 + Cell_3 + Cell_7 + Cell_9$	TN = $Cell_1 + Cell_2 + Cell_4 + Cell_5$
FN = $Cell_4 + Cell_7$	FN = $Cell_2 + Cell_8$	FN = $Cell_3 + Cell_6$

Cálculo da matriz de confusão

Chikungunya	Dengue	Outras doenças
TP = 1300	TP = 727	TP = 1034
FP = 203 + 185	FP = 506 + 506	FP = 335 + 356
TN = 727 + 506 + 356 + 1034	TN = 1300 + 185 + 335 + 1034	TN = 1300 + 203 + 506 + 727
FN = 506 + 335	FN = 203 + 356	FN = 185 + 506

Legenda: Verdadeiro positivo (TP), Falso positivo (FP), Verdadeiro negativo (TN) e Falso negativo (FN)

Fonte: Elaborado pelo autor

resultado da matriz de confusão é necessário calcular as células conforme a figura 6 demonstra, calculando as células para chikungunya obtemos os seguintes resultados: 1.300 (TP), 388 (FP), 2.623 (TN) e 841 (FN); os resultados referente a classe da dengue: 727 (TP), 1.012 (FP), 2.854 (TN) e 559 (FN); e os resultados em relação a classe outras doenças: 1.034 (TP), 691 (FP), 2.736 (TN) e 691 (FN). O número de casos de verdadeiro positivo para dengue foi abaixo das outras classificações, um resultado não esperado, devido ao balanceamento das amostras de cada classe

no *dataset*. Outro fator que chamou bastante atenção foi em relação aos resultados do verdadeiro negativo (TN) para todas as classes, que se manteve maior em relação aos valores verdadeiro positivo (TP).

A partir dos dados gerados da matriz de confusão foram geradas as métricas de acurácia, precisão, *recall* e *F1-Score* com a função *classification_report*. Será utilizado a métrica de precisão visando analisar o grau de variação dos resultados a partir da repetição da mesma análise, com a proporção de identificações positivas. Acurácia para validar o quão próximo da realidade estão os valores obtidos com o valor alvo, sendo a razão entre as previsões corretas e o total de previsões. O *recall* para medir a proporção de positivos que foram previstos corretamente. *F1–score* é a média harmônica de precisão, são valores perfeitos ao atingir o valor 1 utilizada para avaliar os resultados obtidos (52).

Tabela 2 – Relatório das métricas de desempenho do primeiro ciclo

Rótulo	**Precisão**	**Recall**	**F1-Score**
CHIKV	0.61	0.77	0.68
DENV	0.57	0.42	0.48
Outras doenças	0.60	0.60	0.60
Acurácia			0.59

Fonte: Elaborado pelo autor

As primeiras métricas da Tabela 2 mostram valores bem abaixo para classificação da dengue em relação aos outros. As amostras de classificação “outras doenças” contém características bem similares aos outros rótulos. Essa similaridade degrada os resultados do modelo, o primeiro ponto é remover as amostras relacionadas ao rótulo “outras doenças”, essa classificação pode ser qualquer tipo de arbovírus por não ter sido validada. Depois avaliar as *features* importantes para o modelo atual, e em relação ao subconjunto, será utilizado uma técnica para gerar 10 subconjuntos de treinamento e teste para entender se o fator aleatoriedade pode influenciar no desempenho do modelo com a diversificação dos dados no segundo ciclo de melhoria.

4.1.2.3 Segundo ciclo

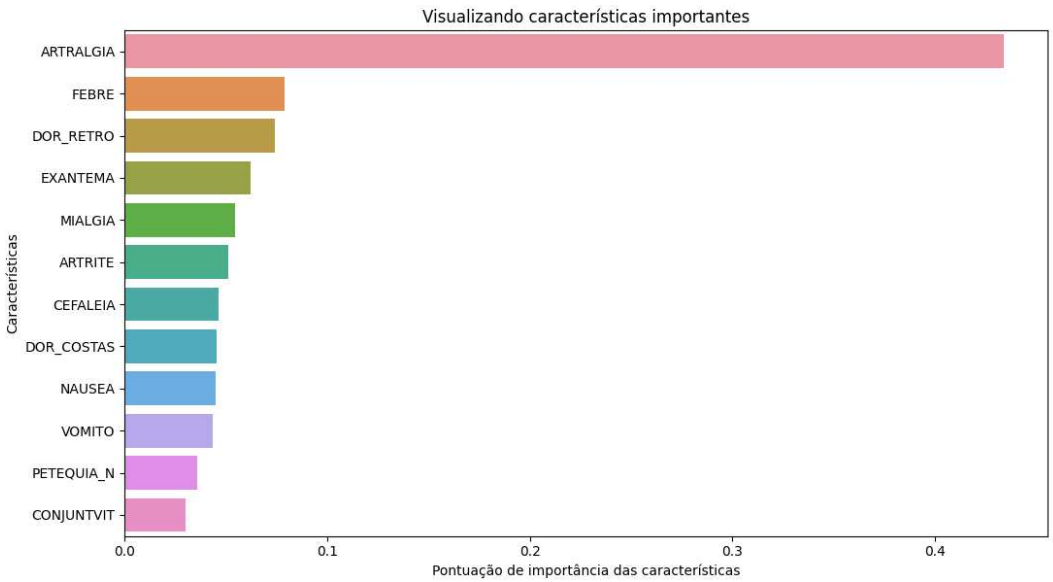
O conhecimento adquirido no segundo ciclo será guia em busca das otimizações para melhorar os resultados do classificador. Essa subseção dará um foco maior em responder se as alterações somente no conjunto de dados podem melhorar a acurácia do classificador.

4.1.2.3.1 Adequação dos dados

Todos os dados da categoria "outras doenças" foram removidos, resultando no total de 11.449 registros. O propósito é diminuir os dados com ruídos que possam interferir na qualidade final do modelo gerado pelo RF. Os dados da categoria que foram removidos tinham muitas similaridades com DENV e CHIKV, eles podem ser classificados como um desses vírus ou até outros.

Com base no modelo do primeiro ciclo foi gerado um gráfico representado na figura 7 com as *features* importantes para o modelo. O objetivo dessa etapa é entender as *features* mais importantes, diminuir a dimensionalidade do modelo com propósito de torná-lo mais simples e melhorar as métricas de desempenho. As *features* conjuntivite e petéquia foram removidas para essa etapa de treinamento devido a sua baixa importância para o modelo.

Figura 7 – Relatório de Classificação



Fonte: Elaborado pelo autor

4.1.2.3.2 Divisão dos dados

No primeiro ciclo foi separado a base de treinamento e teste em uma única amostragem, apesar de ser uma boa abordagem pode não garantir o melhor desempenho. Para maximizar a performance foi utilizado o método *K-fold* que consiste em dividir a base em K subconjuntos previamente definidos com aproximadamente a mesma quantidade de amostras entre eles. Cada interação retornará um valor de acurácia para analisar se o subconjunto tem desempenho melhor que outro, para isso será escolhido um valor de K igual a 10 (dez) para evitar uma interpretação errada do desempenho ou um alto grau de variância (53), esse valor é comumente utilizado para partição do conjunto de dados. Será mantida a proporção da divisão dos dados no sentido que, cada partição separe 70% para treinamento e 30% para teste, assim como foi realizado no primeiro ciclo.

4.1.2.3.3 Treinamento e teste

Nesta etapa aplicado a técnica *Cross Validation* (CV) e as funções *KFold* e *cross val score* da biblioteca *Sklearn* que são fundamentais para executar este experimento. O primeiro argumento *ns_split* passado para a função *Kfold* dividirá base em 10 (dez) subconjuntos com um total de 1.145 amostras cada, o segundo argumento *shuffle* garante o embaralhamento dos dados antes de dividir conforme apresentado no Código-fonte 5.

Código-fonte 5 – Analisando os 10 subconjuntos

```
1  import utils
2  import pandas
3
4  from sklearn.model_selection import KFold
5  from sklearn.model_selection import cross_val_score
6
7  X, y = utils.get_dataset(name = "dataset_arbovirus_v2.csv",
       features = FEATURES_V2)
8  model = utils.get_random_forest_model()
9
10 k_folds = KFold(n_splits = 10, shuffle = True)
```

```
11 scores = cross_val_score(model, X, y.values.ravel(), cv=
       k_folds)
12
13 print("%0.2f (+/- %0.2f)" % (scores.mean(), scores.std() *
       2), "\n")
14
15 list_result = pandas.Series(scores).sort_values(ascending=
       False)
16 list_result = list_result.apply(lambda x: round(x, 2))
17 print(list_result)
```

4.1.2.3.4 Avaliação dos resultados

O resultado do segundo ciclo mostra que as alterações observadas no ciclo anterior tinham influência sobre o desempenho do modelo, conforme ilustrado na figura 8. Acurácia média dos resultados foi de 0,75 ocasionando um aumento de 27% em relação à acurácia do primeiro ciclo 2. Os ajustes em relação aos dados foram satisfatórios, mas não houve nenhum ajuste em relação aos hiperparâmetros da RF. A função *get_random_forest_model* configura três argumentos que contém informações fixadas, tais como: *n_estimators*, *n_jobs* e *max_features*. Existem muitos hiperparâmetros para serem combinado com o propósito de obter o melhor desempenho, no ciclo três avaliado os hiperparâmetros: *n_estimators*, *max_depth*, *class_weight*, *max_features*, *random_state*, *n_jobs* e *criterion*.

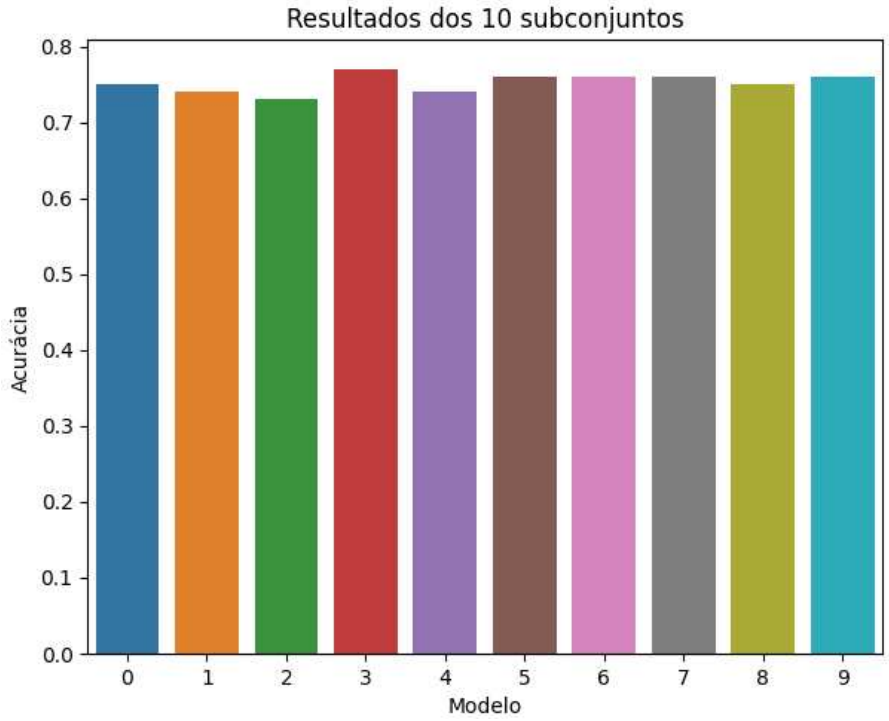
4.1.2.4 Terceiro ciclo

Este ciclo tem como propósito combinar uma série de hiperparâmetros do RF para encontrar a melhor combinação entre eles, nas subseções a seguir será abordado os desafios encontrados e o caminho trilhado para concretizar o experimento.

4.1.2.4.1 Treinamento e teste

Existem muitos hiperparâmetros no RF para serem analisados e ajustados com intuito de otimizar os resultados do modelo criado. Uma análise combinatória de todos os

Figura 8 – Relatório dos resultados da classificação gerados pelos 10 *folds*



Fonte: Elaborado pelo autor

hiperparâmetros se torna um desafio, porque o poder computacional tem que ser grande para analisar todas as combinações possíveis para obter o melhor resultado. Outro fator impactante seria o tempo para montar a estrutura para fazer as combinações necessárias.

Para o primeiro desafio foi necessário limitar a quantidade de hiperparâmetros. Nesse ciclo, 8 (oito) hiperparâmetros foram utilizados. Esses parâmetros foram selecionados por modificações importantes do algoritmo, tais como número de árvores e sua profundidade, avaliação dos resultados de cada resultados das árvores de decisão e pontuação do conjunto de dados de treinamento obtido usando uma estimativa *oob_score*. No Código-fonte 6 é possível ver os valores definidos para a análise combinatória.

Código-fonte 6 – Grade de parâmetros escolhidos para análise combinatória

```
1 param_grid = {
2   'n_estimators': [100, 200, 300, 400, 500],
3   'max_depth': [2, 5, 7, 10, 15, 20],
4   'class_weight': ['balanced', 'balanced_subsample'],
5   'max_features': ['sqrt', 'log2'],
```

```
    'random_state': [24, 42, 123],
    'criterion': ['gini', 'entropy'],
    'n_jobs': [-1],
    'oob_score': [True, False],
}
```

Agora que foi limitado a quantidade dos hiperparâmetros, será necessário realizar a análise, o *Sklearn* tem módulo para automatizar todo o processo de dos parâmetros de um algoritmo. O *GridSearchCV* (54) faz de forma sistemática diversas combinações dos parâmetros e depois de analisar retornar informações importantes como a melhor combinação de parâmetros e melhor resultado de acurácia. Importante ressaltar que todo o conhecimento descoberto foi utilizado na criação da função *grid_search*, possível verificar no Código-fonte 7.

Código-fonte 7 – Função para gerar os resultados necessários do *GridSearchCV*

```
def grid_search(model, X, y, parameters, cv):
    from sklearn.model_selection import GridSearchCV
    from sklearn.metrics import make_scorer, accuracy_score

    grid_obj = GridSearchCV(model, parameters, scoring=
       make_scorer(accuracy_score), cv=cv)
    grid_fit = grid_obj.fit(X, y.values.ravel())

    return grid_fit.best_estimator_, grid_fit.best_score_,
       grid_fit.best_params_, grid_fit.cv_results_
```

O parâmetro *n_jobs* foi definido somente com um valor, para todas as combinações usar todos os processadores, a fim de utilizar o máximo possível do processador da máquina do experimento.

4.1.2.4.2 Avaliação dos resultados

A execução do experimento durou aproximadamente 5 (cinco) horas, foi possível notar um alto consumo dos recursos computacionais de até 60% de CPU e 70% de memoria. No final o resultado foi de 0.76 de acurácia e a melhor combinação de parâmetros pode ser visualizado na tabela 3.

Tabela 3 – Relatório da melhor combinação de parâmetros

Parâmetro	Valor
class_weight	balanced
criterion	gini
max_depth	7
max_features	sqrt
n_estimators	500
n_jobs	-1
oob_score	true
random_state	42

Fonte: Elaborado pelo autor

Apesar de toda complexidade para garantir uma análise completa e mesmo sendo limitado devido a falta de poder computacional, houve uma melhora no resultado da acurácia nesse último ciclo. A tabela 4 mostra a evolução dos resultados, principalmente os valores referente a classe DENV, essa mudança resultou em um aumento aproximado de 29% na acurácia em relação ao relatório do primeiro ciclo. Aplicando o método CV é possível obter resultado até 0.77 de acurácia, mas o intuito com essa avaliação foi de comparar entre o primeiro e último resultado.

A matriz de confusão gerada depois das modificações do conjunto de dados se tornou mais simples para se analisar os resultados por ser uma matriz 2x2, pode ser observado na figura 9 sem a necessidade de calcular as células. Podemos inferir diretamente o modelo classificou CHIKV 1.346 vezes corretamente, classificou DENV 1.264 vezes corretamente, classificou CHIKV 333 vez incorretamente e classificou DENV 492 vezes incorretamente. O

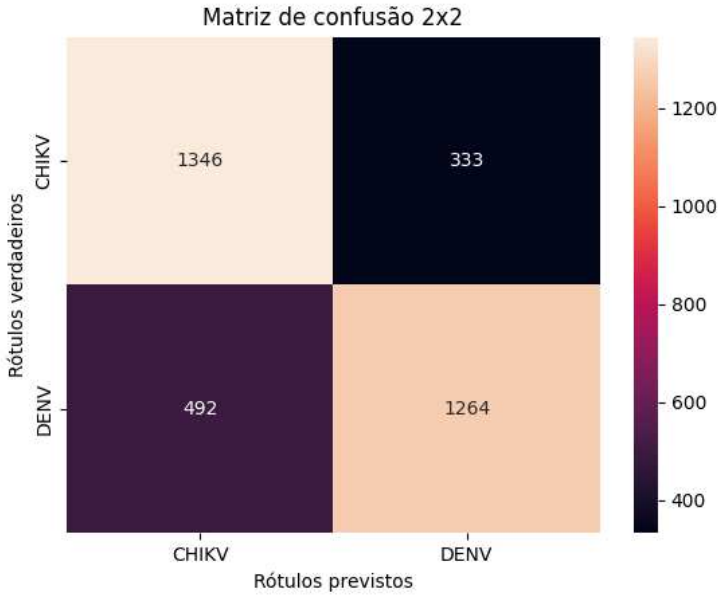
Tabela 4 – Relatório das métricas de desempenho do terceiro ciclo

Rótulo	Precisão	Recall	F1-Score
CHIKV	0.73	0.80	0.77
DENV	0.79	0.72	0.75
Acurácia			0.76

Fonte: Elaborado pelo autor

valor de classificação correta para DENV melhorou drasticamente, esse foi um fator essencial para apresentar resultados significativos durante o processo do terceiro ciclo.

Figura 9 – Matriz de confusão 2x2 do terceiro ciclo



Identificação da matriz de confusão

Chikungunya	Matriz
TP = 1.346	TP = Cell1
FP = 333	FP = Cell2
TN = 1.264	TN = Cell3
FN = 492	FN = Cell4

Dengue	Matriz
TP = 1.264	TP = Cell4
FP = 492	FP = Cell3
TN = 1.346	TN = Cell1
FN = 333	FN = Cell2

Legenda: Verdadeiro positivo (TP), Falso positivo (FP), Verdadeiro negativo (TN) e Falso negativo (FN)

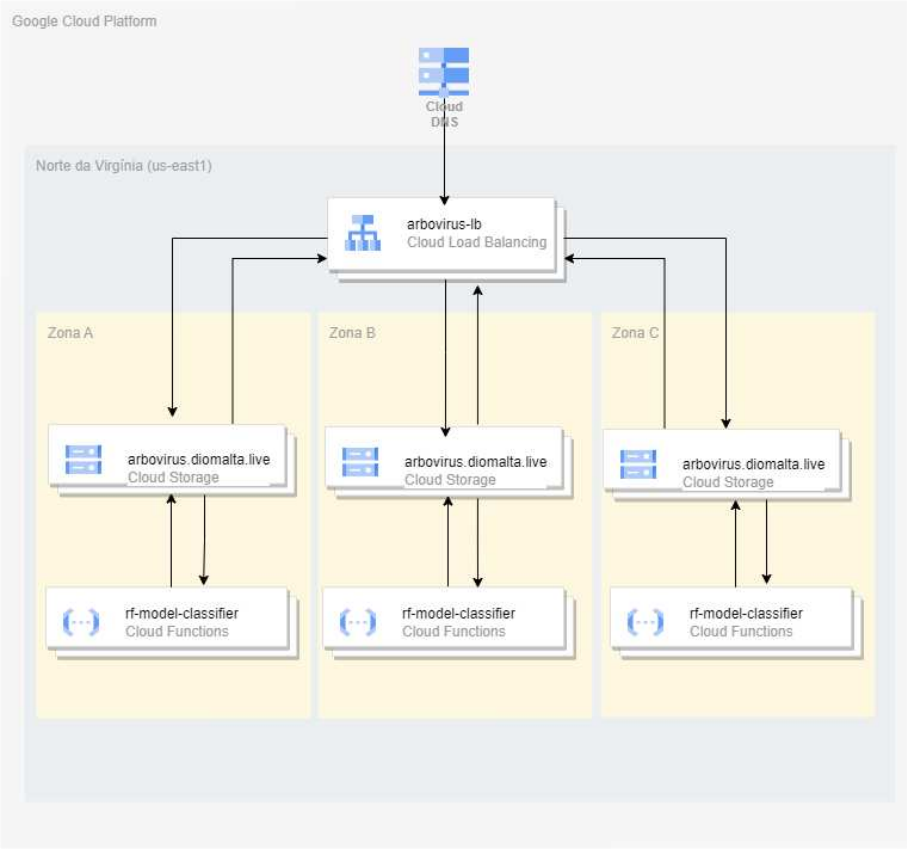
Fonte: Elaborado pelo autor

4.1.2.4.3 Artefato

Com base nessas informações coletadas durante os três ciclos foi construído o modelo classificador final, onde foi persistindo localmente em formato de arquivo. Para ser utilizado é necessário conhecimento em programação e, principalmente com a linguagem *Python* para carregar o modelo e utilizar as funções disponíveis. Reconhecendo essas barreiras para utilização do modelo no dia a dia, foi desenvolvido uma aplicação web de uma única página.

A arquitetura da infraestrutura pode ser visualizada na figura 10, construído no Google Cloud Platform (GCP) com base nos princípios de uma arquitetura de alta disponibilidade, onde a rede configurada na região Norte da Virgínia tem três zonas (A, B, C) de disponibilidade que são isoladas uma das outras, caso uma delas fique indisponível haverá outras para manter a aplicação disponível.

Figura 10 – Arquitetura da infraestrutura



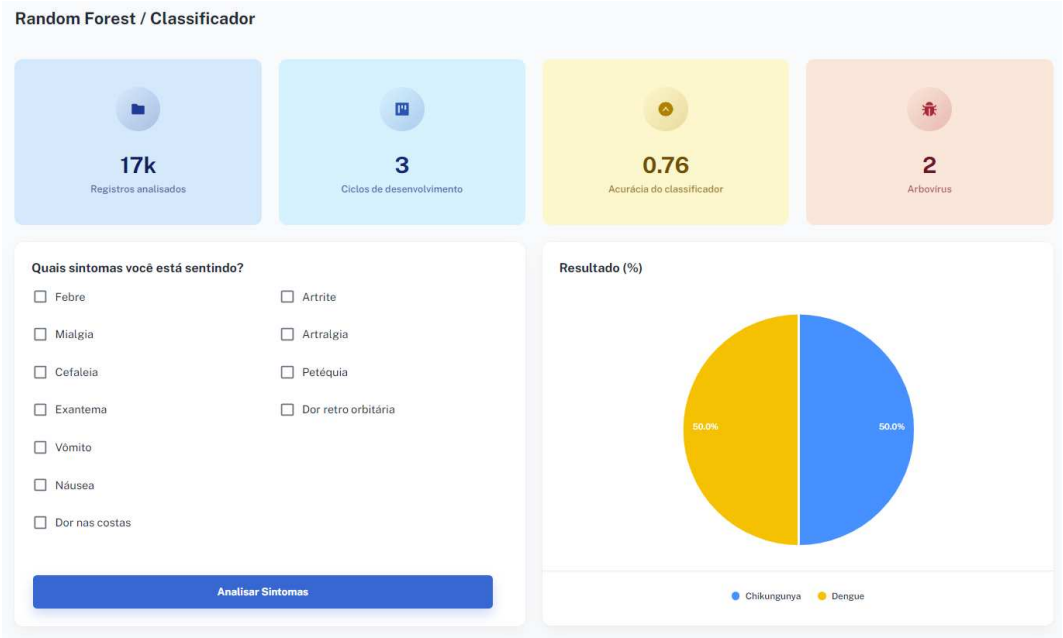
Fonte: Elaborado pelo autor

O principal ponto de entrada é através do serviço *Cloud DNS* com o subdomínio *arbovirus.diomalta.live*, que redirecionado o fluxo para o serviço do *Cloud Load Balancing*

para distribuir a carga de requisições entre as réplicas do *Cloud Storage* que contém os arquivos estáticos para a aplicação web nas zonas de disponibilidade. Por sua vez, a aplicação web faz uma requisição *Hypertext Transfer Protocol* (HTTP) do método *GET* para o serviço do *Cloud Functions* onde está a função *analisys_http*, que ao receber as informações sobre os sintomas, analisa e retorna as informações referente a probabilidade entre os arbovírus.

A camada visual foi construída somente com uma página principal e três seções, a primeira seção traz um resumo referente aos dados deste trabalho, a segunda seção uma lista de sintomas selecionáveis com um botão de analisar e a terceira seção com o resultado de infecção do CHIKV ou DENV, pode ser visualizado na figura 11.

Figura 11 – Aplicação web para demonstrar o funcionamento do classificador desenvolvido com RF



Fonte: Elaborado pelo autor

Apesar de ser uma arquitetura moderna, ela foi construída para ser custo benefício, alinhando alta disponibilidade com baixo custo para garantir a demonstração do artefato com sucesso. No comando para analisar os sintomas o retorno da resposta não se faz de imediato devido a natureza do serviço do *Cloud Functions* de só iniciar sob demanda, quando acionado, começa uma nova instância fazendo a instalação das dependências necessárias para rodar a função e somente depois analisar os sintomas para retornar o resultado para o usuário. Por fim, dando uma resposta visual para o usuário final.

5 CONSIDERAÇÕES FINAIS

O propósito da análise dos conjuntos de dados desde o início foi de utilizar amostras já classificadas e validadas por especialistas da área médica. Inicialmente, foram aceitas todas as amostras contidas, mesmo para aqueles que não tinham uma classificação exata. Durante o primeiro ciclo, a percepção era de que, uma maior variedade de rótulos faria o modelo ter uma capacidade melhor na classificação dos arbovírus. Entretanto, aceitar esses dados ocasionou um modelo impreciso. Durante o primeiro ciclo o resultado obtido se tornou base para a otimização do modelo com o RF.

No segundo ciclo onde passou a ser analisado o conjunto de dados, foi realizado uma série de modificações que culminaram no melhor resultado na classificação do modelo. O resultado obtido mostrou uma importância inesperada naquele momento, as *features* não importantes e amostras com o rótulo que degradava as métricas de desempenho foram removidas, conhecido como “outras doenças”. A partir desse momento, obteve uma quantidade de amostras reduzidas, entretanto essa ação impactou diretamente nas métricas de desempenho apresentadas pelo modelo do segundo ciclo de forma positiva. Também, o algoritmo RF apresentou um tempo de processamento mais curto em relação ao ciclo anterior.

O conjunto de dados utilizado no estudo contém poucas amostras de casos de pacientes infectados por arbovírus, a dificuldade de encontrar dados com essas especificações é grande. O DATASUS até o presente momento não tem dados de pacientes com sintomas do ZIKV, acarretando na diminuição do escopo da pesquisa e impactando na quantidade de amostras. Por outro lado, um conjunto de dados pequeno contribui para um processamento mais rápido, mas nem em todo caso isso se torna verdade. Tem outros fatores que impactam no tempo de processamento do RF, a capacidade computacional da máquina no qual está sendo processado os dados, influencia diretamente nesse tempo, porque o RF exige bastante processamento e memória. Esse desempenho pode ser impactado diretamente pelo conjunto de hiperparâmetros configurados e o tamanho do conjunto de dados.

A importância da *feature* artralgia foi latente na construção do modelo, por ter muitos casos de pacientes com esse sintoma estarem classificados como CHIKV, esse fator afetou muito

a classificação dos sintomas, resultando nas melhores métricas de desempenho em relação a DENV. Essa importância fica muito clara na utilização da aplicação web, onde determinados sintomas selecionados, fazem a probabilidade preditiva inverter o resultado para CHIKV, ou DENV. Mas de forma geral, a diferença final dos resultados foram pequenas, mas em todo o processo os resultados da CHIKV foram os melhores.

O artefato construído durante o processo permite através da metodologia a continuação do desenvolvimento do mesmo, aplicando novos dados, funcionalidades e adicionando novos hiperparâmetros para uma possível melhoria dos resultados obtidos com esse modelo. As informações documentadas entre os ciclos entregam com o máximo de detalhes para o leitor entender a linha de raciocínio, a construção dos protótipos e os resultados que se obteve para ganhar contexto sobre o trabalho.

Para trabalhos futuros sugere-se, com base na experiência desta pesquisa, três possibilidades. Uma de fácil acesso, por exemplo, é a introdução de novas *features* existentes do conjunto de dados utilizado neste estudo, que não diz respeito somente aos sintomas. Algumas *features* existentes adicionais que podem ser utilizadas, tais como idade, gênero, testes médicos, entre outros. O estudo pode ser com objetivo de analisar se a adição de novos dados podem valorizar ou degradar as métricas de desempenho do modelo. Existem estudos que não estão relacionados aos sintomas, mas que tem como intuito analisar fatores socioeconômicos e climatológicos na proliferação do *DENV*. Esse estudo agregaria no entendimento da infecção por arbovírus e uma possível melhoria no artefato gerado neste estudo.

A proposta inicial deste projeto era contar com dados relacionados a ZIKV, infelizmente não foi possível, entretanto se os dados forem liberados pelo DATASUS a continuação poderia ser focada em adicionar esses dados para replicar os passos realizados deste estudo, além de propor novas ações, melhorias e novos ciclos para a realização das etapas. Trazendo uma perspectiva diferente, podem ser construídos modelos separados com os dados rotulados dos arbovírus para entender se o modelo concebido de forma isolada altera o resultado final.

A análise dos hiperparâmetros foi limitada por conta da capacidade computacional utilizada. Existem muitos valores para serem analisados, além de hiperparâmetros não introduzidos no estudo para gerar mais conjuntos de teste. Para esta continuação será necessário uma máquina dedicada para executar o experimento com objetivo de obter o menor tempo possível da análise combinatória, esse requisito pode gerar um custo extra.

REFERÊNCIAS

1 AL-GHUSSAIN, L. Global warming: Review on driving forces and mitigation. **Environmental Progress & Sustainable Energy**, Wiley Online Library, v. 38, n. 1, p. 13–21, 2019.

2 CAMPOS, J. M.; OLIVEIRA, D. M. d.; FREITAS, E. J. d. A.; NETO, A. C. et al. Arboviroses de importância epidemiológica no brasil. Universidade Estadual Paulista, 2018.

3 FATHIMA, A. S.; MANIMEGLAI, D. Analysis of significant factors for dengue infection prognosis using the random forest classifier. **Int J Adv Comput Sci Appl**, v. 6, n. 2, p. 240–245, 2015.

4 SARMA, D.; HOSSAIN, S.; MITTRA, T.; BHUIYA, M. A. M.; SAHA, I.; CHAKMA, R. Dengue prediction using machine learning algorithms. In: IEEE. **2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)**. [S.l.], 2020. p. 1–6.

5 COMBATE à Dengue. [S.l.]: Governo do Paraná. <https://www.dengue.pr.gov.br/Pagina/Como-combater>. Acessado em: 02 nov. 2022.

6 FAHMI, A.; PURWITASARI, D.; SUMPENO, S.; PURNOMO, M. H. Performance evaluation of classifiers for predicting infection cases of dengue virus based on clinical diagnosis criteria. In: IEEE. **2020 International Electronics Symposium (IES)**. [S.l.], 2020. p. 456–462.

7 BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. ISSN 08856125. Disponível em: <http://link.springer.com/10.1023/A:1010933404324>.

8 HA, T. M. P.; TRAN, D. H.; HANH, L. T. M.; BINH, N. T. Experimental study on software fault prediction using machine learning model. In: IEEE. **2019 11th International Conference on Knowledge and Systems Engineering (KSE)**. [S.l.], 2019. p. 1–5.

9 LOPES, N.; NOZAWA, C.; LINHARES, R. E. C. CaracterÃsticas gerais e epidemiologia dos arbovÃrus emergentes no Brasil. **Revista Pan-AmazÃ´nica de SaÃ**, scielo, v. 5, p. 55 – 64, 09 2014. ISSN 2176-6223. Disponível em: <http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S2176-62232014000300007&nrm=iso>.

10 MURUGESAN, A.; MANOHARAN, M. Dengue virus. In: **Emerging and Reemerging Viral Pathogens**. [S.l.]: Elsevier, 2020. p. 281–359.

11 VASILAKIS, N.; FOKAM, E. B.; HANSON, C. T.; WEINBERG, E.; SALL, A. A.; WHITEHEAD, S. S.; HANLEY, K. A.; WEAVER, S. C. Genetic and phenotypic characterization of sylvatic dengue virus type 2 strains. **Virology**, Elsevier, v. 377, n. 2, p. 296–307, 2008.

12 HOTTA, S. Experimental studies on dengue: I. isolation, identification and modification of the virus. **The Journal of infectious diseases**, JSTOR, p. 1–9, 1952.

13 SABIN, A. B.; SCHLESINGER, R. W. Production of immunity to dengue with virus modified by propagation in mice. **Science**, American Association for the Advancement of Science, v. 101, n. 2634, p. 640–642, 1945.

14 HAMMON, W. M.; RUNDNICK, A.; SATHER, G. Viruses associated with epidemic hemorrhagic fevers of the philippines and thailand. **Science**, American Association for the Advancement of Science, v. 131, n. 3407, p. 1102–1103, 1960.

15 DUARTE, H. H. P.; FRANÇA, E. B. Data quality of dengue epidemiological surveillance in belo horizonte, southeastern brazil. **Revista de Saúde Pública**, SciELO Brasil, v. 40, p. 134–142, 2006.

16 ZHAO, S.; MUSA, S. S.; FU, H.; HE, D.; QIN, J. Simple framework for real-time forecast in a data-limited situation: the zika virus (zikv) outbreaks in brazil from 2015 to 2016 as an example. **Parasites & vectors**, Springer, v. 12, n. 1, p. 1–13, 2019.

17 JUNIOR, V. L. P.; LUZ, K.; PARREIRA, R.; FERRINHO, P. Zika virus: a review to clinicians. **Acta medica portuguesa**, v. 28, n. 6, p. 760–765, 2015.

18 DUFFY, M. R.; CHEN, T.-H.; HANCOCK, W. T.; POWERS, A. M.; KOOL, J. L.; LANCIOTTI, R. S.; PRETRICK, M.; MARFEL, M.; HOLZBAUER, S.; DUBRAY, C. et al. Zika virus outbreak on yap island, federated states of micronesia. **New England Journal of Medicine**, Mass Medical Soc, v. 360, n. 24, p. 2536–2543, 2009.

19 SANTOS, T. D.; RODRIGUEZ, A.; ALMIRON, M.; SANHUEZA, A.; RAMON, P.; OLIVEIRA, W. K. de; COELHO, G. E.; BADARÓ, R.; CORTEZ, J.; OSPINA, M.; PIMENTEL, R.; MASIS, R.; HERNANDEZ, F.; LARA, B.; MONTOYA, R.; JUBITHANA, B.; MELCHOR, A.; ALVAREZ, A.; ALDIGHIERI, S.; DYE, C.; ESPINAL, M. A. Zika virus and the guillain-barré syndrome - case series from seven countries. **N. Engl. J. Med.**, v. 375, n. 16, p. 1598–1601, out. 2016.

20 FERREIRA, M. L. B.; BRITO, C. A. A. de; FRANÇA, R. F. de O.; MOREIRA, Á. J. P.; MACHADO, M. Í. de M.; MELO, R. da P.; MEDIALDEA-CARRERA, R.; MESQUITA, S. D.; SANTOS, M. L.; MEHTA, R. et al. Neurological disease in adults with zika and chikungunya virus infection in northeast brazil: a prospective observational study. **The Lancet Neurology**, Elsevier, v. 19, n. 10, p. 826–839, 2020.

21 ROSS, R. The newala epidemic: Iii. the virus: isolation, pathogenic properties and relationship to the epidemic. **Epidemiology & Infection**, Cambridge University Press, v. 54, n. 2, p. 177–191, 1956.

22 SIMON, F.; SAVINI, H.; PAROLA, P. Chikungunya: a paradigm of emergence and globalization of vector-borne diseases. **Medical Clinics of North America**, Elsevier, v. 92, n. 6, p. 1323–1343, 2008.

23 HOCHEDEZ, P.; HAUSFATER, P.; JAUREGUIBERRY, S.; GAY, F.; DATRY, A.; DANIS, M.; BRICAIRE, F.; BOSSI, P. Cases of chikungunya fever imported from the islands of the south west indian ocean to paris, france. **Eurosurveillance**, European Centre for Disease Prevention and Control, v. 12, n. 1, p. 13–14, 2007.

24 SIMON, F.; JAVELLE, E.; OLIVER, M.; LEPARC-GOFFART, I.; MARIMOUTOU, C. Chikungunya virus infection. **Current infectious disease reports**, Springer, v. 13, n. 3, p. 218–228, 2011.

25 SILVA, N. M. d.; TEIXEIRA, R. A. G.; CARDOSO, C. G.; JUNIOR, J. B. S.; COELHO, G. E.; OLIVEIRA, E. S. F. d. Chikungunya surveillance in brazil: challenges in the context of public health. **Epidemiol. Serv. Saude**, v. 27, n. 3, p. e2017127, set. 2018.

26 OLIVEIRA, R. Lourenço-de; BRAGA, I. A. et al. Updating the geographical distribution and frequency of aedes albopictus in brazil with remarks regarding its range in the americas. **Memórias do Instituto Oswaldo Cruz**, SciELO Brasil, v. 109, p. 787–796, 2014.

27 CHAVES, T. d. S. S.; PELLINI, A. C. G.; MASCHERETTI, M.; JAHNEL, M. T.; RIBEIRO, A. F.; RODRIGUES, S. G.; VASCONCELOS, P. F. da C.; BOULOS, M. Travelers as sentinels for chikungunya fever, brazil. **Emerging infectious diseases**, Centers for Disease Control and Prevention, v. 18, n. 3, p. 529, 2012.

28 MUSSO, D.; GUBLER, D. J. Zika virus. **Clin. Microbiol. Rev.**, American Society for Microbiology, v. 29, n. 3, p. 487–524, jul. 2016.

29 KARKHAH, A.; NOURI, H. R.; JAVANIAN, M.; KOPPOLU, V.; MASROUR-ROUDSARI, J.; KAZEMI, S.; EBRAHIMPOUR, S. Zika virus: epidemiology, clinical aspects, diagnosis, and control of infection. **Eur. J. Clin. Microbiol. Infect. Dis.**, Springer Science and Business Media LLC, v. 37, n. 11, p. 2035–2043, nov. 2018.

30 MEANEY-DELMAN, D.; ODUYEBO, T.; POLEN, K. N. D.; WHITE, J. L.; BINGHAM, A. M.; SLAVINSKI, S. A.; HEBERLEIN-LARSON, L.; GEORGE, K. S.; RAKEMAN, J. L.; HILLS, S.; OLSON, C. K.; ADAMSKI, A.; BARLOW, L. C.; LEE, E. H.; LIKOS, A. M.; MUÑOZ, J. L.; PETERSEN, E. E.; DUFORT, E. M.; DEAN, A. B.; CORTESE, M. M.; SANTIAGO, G. A.; BHATNAGAR, J.; POWERS, A. M.; ZAKI, S.; PETERSEN, L. R.; JAMIESON, D. J.; HONEIN, M. A.; for the U.S. Zika Pregnancy Registry Prolonged Viremia Working Group. Prolonged detection of zika virus RNA in pregnant women. **Obstet. Gynecol.**, Ovid Technologies (Wolters Kluwer Health), v. 128, n. 4, p. 724–730, out. 2016.

31 HTUN, T. P.; XIONG, Z.; PANG, J. Clinical signs and symptoms associated with WHO severe dengue classification: a systematic review and meta-analysis. **Emerg. Microbes Infect.**, Informa UK Limited, v. 10, n. 1, p. 1116–1128, dez. 2021.

32 CHAWLA, P.; YADAV, A.; CHAWLA, V. Clinical implications and treatment of dengue. **Asian Pac. J. Trop. Med.**, v. 7, n. 3, p. 169–178, mar. 2014.

33 QUEYRIAUX, B.; SIMON, F.; GRANDADAM, M.; MICHEL, R.; TOLOU, H.; BOUTIN, J.-P. Clinical burden of chikungunya virus infection. **The Lancet infectious diseases**, v. 1, n. 8, p. 2–3, 2008.

34 SIMON, F.; PAROLA, P.; GRANDADAM, M.; FOURCADE, S.; OLIVER, M.; BROUQUI, P.; HANCE, P.; KRAEMER, P.; MOHAMED, A. A.; LAMBALLERIE, X. de et al. Chikungunya infection: an emerging rheumatism among travelers returned from indian ocean islands. report of 47 cases. **Medicine**, LWW, v. 86, n. 3, p. 123–137, 2007.

35 NAQA, I. E.; MURPHY, M. J. What is machine learning? In: **Machine Learning in Radiation Oncology**. Cham: Springer International Publishing, 2015. p. 3–11.

36 DIŽO, J.; BLATNICKÝ, M.; MELNIK, R.; 0003-4677-2535), O. K. https://orcid.org/0000. A mathematical model of operation of a semi-trailer tractor powertrain. **Komunikácie**, University of Zilina, jun. 2022.

37 SARAVANAN, R.; SUJATHA, P. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In: IEEE. **2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)**. [S.l.], 2018. p. 945–949.

38 AMIT, Y.; GEMAN, D. Shape quantization and recognition with randomized trees. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 9, n. 7, p. 1545–1588, 1997.

39 CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random forests. In: **Ensemble machine learning**. [S.l.]: Springer, 2012. p. 157–175.

40 DEY, L.; MUKHOPADHYAY, A. Compact Genetic Algorithm-based Feature Selection for Sequence-based Prediction of Dengue-Human Protein Interactions. **IEEE/ACM Trans. Comput. Biol. and Bioinf.**, p. 1–1, 2021. ISSN 1545-5963, 1557-9964, 2374-0043. Disponível em: <https://ieeexplore.ieee.org/document/9380918/>.

41 JIANG, D.; HAO, M.; DING, F.; FU, J.; LI, M. Mapping the transmission risk of Zika virus using machine learning models. **Acta Tropica**, v. 185, p. 391–399, set. 2018. ISSN 0001706X. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0001706X18303619>.

42 MUDELE, O.; BAYER, F. M.; ZANANDREZ, L. F. R.; EIRAS, A. E.; GAMBA, P. Modeling the Temporal Population Distribution of $Ae.~aegypti$ Mosquito Using Big Earth Observation Data. **IEEE Access**, v. 8, p. 14182–14194, 2020. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/8957072/>.

43 YIN, M. S.; HADDAWY, P.; NIRANDMONGKOL, B.; KONGTHAWORN, T.; CHAI-SUMRITCHOKE, C.; SUPRATAK, A.; SA-NGAMUANG, C.; SRIWICHAI, P. A lightweight deep learning approach to mosquito classification from wingbeat sounds. In: **Proceedings of the Conference on Information Technology for Social Good**. New York, NY, USA: Association for Computing Machinery, 2021. (GoodIT '21), p. 37–42. ISBN 9781450384780. Disponível em: <https://doi.org/10.1145/3462203.3475908>.

44 HUANG, S.-W.; TSAI, H.-P.; HUNG, S.-J.; KO, W.-C.; WANG, J.-R. Assessing the risk of dengue severity using demographic information and laboratory test results with machine learning. **PLoS Negl Trop Dis**, v. 14, n. 12, p. e0008960, dez. 2020. ISSN 1935-2735. Disponível em: <https://dx.plos.org/10.1371/journal.pntd.0008960>.

45 FAHMI, A.; PURWITASARI, D.; SUMPENO, S.; PURNOMO, M. H. Performance Evaluation of Classifiers for Predicting Infection Cases of Dengue Virus Based on Clinical Diagnosis Criteria. In: **2020 International Electronics Symposium (IES)**. Surabaya, Indonesia: IEEE, 2020. p. 456–462. ISBN 978-1-72819-528-5 978-1-72819-530-8. Disponível em: <https://ieeexplore.ieee.org/document/9231728/>.

46 CLEMENTE, C. J.; JAAFAR, F.; MALIK, Y. Is Predicting Software Security Bugs Using Deep Learning Better Than the Traditional Machine Learning Algorithms? In: **2018 IEEE International Conference on Software Quality, Reliability and Security (QRS)**. Lisbon: IEEE, 2018. p. 95–102. ISBN 978-1-5386-7757-5. Disponível em: <https://ieeexplore.ieee.org/document/8424961/>.

47 ALENEZI, F.; TSOKOS, C. P. Machine Learning Approach to Predict Computer Operating Systems Vulnerabilities. In: **2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)**. Riyadh, Saudi Arabia: IEEE, 2020. p. 1–6. ISBN 978-1-72814-213-5. Disponível em: <https://ieeexplore.ieee.org/document/9096731/>.

48 PUSHPHAVATHI, T. P.; SUMA, V.; RAMASWAMY, V. A novel method for software defect prediction: Hybrid of FCM and random forest. In: **2014 International Conference on**

Electronics and Communication Systems (ICECS). Coimbatore: IEEE, 2014. p. 1–5. ISBN 978-1-4799-2320-5. Disponível em: <https://ieeexplore.ieee.org/document/6892743>.

49 HERVER, A.; CHATTERJEE, S. **Design Research in Information Systems. Theory and Practice**. [S.l.]: Integrated series in information systems, 2001.

50 PIMENTEL, M.; FILIPPO, D.; SANTORO, F. M. Design science research: fazendo pesquisas científicas rigorosas atreladas ao desenvolvimento de artefatos computacionais projetados para a educação. **Metodologia de Pesquisa em Informática na Educação: Concepção da Pesquisa. Porto Alegre: SBC**, 2019.

51 Thomás Tabosa. **Clinical cases of Dengue and Chikungunya**. Mendeley, 2021. Disponível em: <https://data.mendeley.com/datasets/bv26kznkjs/1>.

52 ALENEZI, F.; TSOKOS, C. P. Machine learning approach to predict computer operating systems vulnerabilities. In: IEEE. **2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)**. [S.l.], 2020. p. 1–6.

53 CROSS Validation Explained: Evaluating estimator performance. [S.l.]: Rahil Shaikh. <https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85>. Acessado em: 02 nov. 2022.

54 GRIDSEARCHCV. [S.l.]: Scikit Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>. Acessado em: 23 set. 2022.

APÊNDICES

APÊNDICE A – Funções utilitárias

Código-fonte 8 – Função para retornar os subconjuntos do conjunto de dados

```
def get_k_folds(ns = 5):
    from sklearn.model_selection import KFold
    return KFold(n_splits=ns, shuffle=True)
```

Código-fonte 9 – Função para visualizar o gráfico das *features* importantes

```
def show_feature_importances(model, features =
   DEFAULT_FEATURES):
    import matplotlib.pyplot as plt
    import pandas as pd
    import seaborn as sns

    importances = pd.Series(model.feature_importances_,
       index=features).sort_values(ascending=False)
    print(importances)

    sns.barplot(x=importances, y=importances.index)

    plt.xlabel('Pontuacao de importancia das
       caracteristicas')
    plt.ylabel('Caracteristicas')
    plt.title("Visualizando caracteristicas importantes")
    plt.show()
```

Código-fonte 10 – Função para visualizar o relatório das métricas de desempenho

```
def show_classification_report(y_test, y_pred):
```

```
    from sklearn.metrics import classification_report
    print("Relatorio de Classificacao", "\n")
    print(classification_report(y_test, y_pred))
```

Código-fonte 11 – Função para visualizar a matriz de confusão

```
def show_confusion_matrix(y_test, y_pred):
    from sklearn.metrics import confusion_matrix
    import matplotlib.pyplot as plt
    import seaborn as sns

    cm = confusion_matrix(y_test, y_pred)
    ax= plt.subplot()
    sns.heatmap(cm, annot=True, ax = ax, fmt='g'); #annot=
       True to annotate cells

    ax.set_xlabel('Rotulos previstos');ax.set_ylabel('
       Rotulos verdadeiros')
    ax.set_title('Matriz de confusao 2x2');
    ax.xaxis.set_ticklabels(['CHIKV', 'DENV']); ax.yaxis.
       set_ticklabels(['CHIKV', 'DENV'])

    plt.show()
```