DOI: 10.36558/rsc.v11i3.7475

Uso de *Machine Learning* e de *Deep Learning* para prever internações causadas por dengue na Paraíba

Use of Machine Learning and Deep Learning to predict hospitalizations caused by dengue in Paraíba

Ewerthon Dyego de Araújo Batista
Universidade Estadual da Paraíba
ewerthon.batista@aluno.uepb.edu.br

Wellington Candeia de Araújo
Universidade Estadual da Paraíba
wcandeia@uepb.edu.br

Romeryto Vieira Lira
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
romeryto.lira@academico.ifpb.edu.br

Laryssa Izabel de Araújo Batista
Universidade Federal da Paraíba
laryssa.izabel@gmail.com

ABSTRACT

Dengue is a public health problem in Brazil, and although it is not a new disease, there is still no vaccine, without restrictions on its use, that can be applied to the population. Complementing this, the vector of dengue finds in Brazil an ideal place for its proliferation. With that, the numbers of dengue continue to grow. According to the epidemiological bulletin of Paraíba, released in August 2021, there was an increase of 53% of cases compared to the previous year. Machine Learning (ML) and Deep Learning techniques are being used as tools for predicting the disease and helping to fight it. The objective of this work was, through the Random Forest (RF), Support Vector Regression (SVR), Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) techniques, and using epidemiological information, climate and health, create a system capable of forecasting hospitalizations caused by dengue in the cities of Bayeux, Cabedelo, Cajazeiras, João Pessoa, Patos and Santa Rita. At the end of the project, the system managed to make forecasts for Bayeux with an error rate of 0.529017139, while in Cabedelo the error was 0.927428107, for Cajazeiras 1,000751246, in João Pessoa 9.552880644, in Patos 0.422049567 and, finally, to Santa Rita, 0.745519952.

Keywords

Forecast; Dengue; Hospitalization; Machine Learning; Deep Learning

RESUMO

Dengue é um problema de saúde pública no Brasil, e, embora não seja uma doença nova, ainda não existe uma vacina sem restrições de uso que possa ser aplicada na população. Complementar a isso, o vetor da dengue encontra, no território brasileiro, locais favoráveis a sua proliferação, ocasionando um crescimento contínuo da doença. Conforme o boletim epidemiológico da Paraíba, divulgado em agosto de 2021, houve um aumento de 53% de casos em relação ao ano anterior. Técnicas de *Machine Learning* (ML) e de *Deep Learning* estão sendo utilizadas como ferramentas para a previsão da doença e auxiliando no seu combate. O objetivo deste trabalho foi, por meio das técnicas *Random Forest* (RF), *Support Vector Regression* (SVR), *Multilayer Perceptron* (MLP), *Long Short-Term Memory* (LSTM) e *Convolutional Neural Network* (CNN), e usando informações epidemiológicas, climáticas e sanitárias, criar um sistema capaz de realizar previsões de internações causadas por dengue para as cidades Bayeux, Cabedelo, Cajazeiras, João Pessoa, Patos e Santa Rita. Ao término do trabalho, o sistema conseguiu realizar previsões para Bayeux com taxa de erro 0,529017139, para Cabedelo o erro foi 0,927428107, 1,000751246 para Cajazeiras, 9,552880644 em João Pessoa, 0,422049567 em Patos e, finalmente, para Santa Rita, 0,745519952.

Palavras-Chave

Previsão; Dengue; Internação; Machine Learning; Deep Learning

CCS Concepts

• Applied computing→ Life and medical sciences→ Health care information systems

1. INTRODUÇÃO

Dengue é uma doença endêmica, causada pelo vírus DENV, sendo transmitida através do mosquito *Aedes aegypti*. Atualmente, existem quatro tipos sorológicos do vírus (1, 2, 3 e 4) em circulação no Brasil [26]. Embora não seja uma doença nova, ainda não existe vacina eficaz para a imunização da população contra todos os tipos do vírus. Uma vez infectado por uma sorologia, o paciente adquire imunidade a essa variação, porém, continua suscetível às demais [30]. Adicionalmente, de acordo com o sétimo boletim epidemiológico de arboviroses, houve um aumento de 53% de casos de dengue em comparação ao ano anterior [24].

Para combater a doença, os sistemas governamentais investem em campanhas de conscientização da população, sendo uma delas o correto descarte de pneus e recipientes, que estão a céu aberto. Isso acontece, visto que, eles podem acumular água e se tornar, futuramente, berço para proliferação do mosquito.

Além do problema de saúde, a dengue vem causando impacto financeiro aos cofres públicos. De acordo com informação fornecida pela secretaria de saúde da Paraíba[1], em 2021, já foram gastos R$ 936.130,00 em ações de combate e prevenção à dengue.

[1] Informação obtida através do protocolo de número 00099.002082/2021-2, no sicpb (https://sic.pb.gov.br/).

O uso da tecnologia da informação na saúde está cada vez mais constante. De acordo com Pinochet [10], os sistemas de informação vêm sendo utilizados no apoio à saúde, na prevenção de doenças, nas promoções de ações de saúde, no controle de doenças, mas também, na vigilância e no monitoramento de doenças. Para Longaray [19], a Tecnologia da Informação se tornou parte integral para todas as atividades relacionadas à prestação dos serviços de saúde.

Nesse contexto, as técnicas de ML e de DL estão sendo utilizadas como ferramentas de apoio ao combate da dengue [10] [19] [25]. Por meio da utilização de dados epidemiológicos, climáticos e fatores sociais, pesquisadores estão desenvolvendo modelos de previsão de casos da doença. Com base nas previsões, os governantes podem direcionar melhor os esforços e os recursos (financeiros, humanos e hospitalares) contra a doença [31].

Diante desse cenário, o objeto deste trabalho é, por meio das técnicas *Random Forest* (RF), *Support Vector Regression* (SVR), *Multilayer Perceptron* (MLP), *Long short-term memory* (LSTM) e *Convolutional Neural network* (CNN), criar e avaliar modelos para predição de casos internações causadas por dengue para as cidades: Bayeux, Cabedelo, Cajazeiras, João Pessoa, Patos e Santa Rita.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Dengue

A arbovirose dengue é causada pelo vírus DENV e transmitida através do mosquito *Aedes aegypti*. Os principais sintomas da dengue são: febre alta, dores musculares, mal-estar, falta de apetite e dores de cabeça. Em alguns casos mais graves, a dengue pode causar hemorragias e levar o paciente a óbito [9].

O ciclo da doença dengue inicia com o mosquito *Aedes aegypti* picando um humano infectado. Uma vez infectado, o mosquito é capaz de transmitir dengue até o fim da sua vida. O mosquito se reproduz através do depósito de seus ovos em águas paradas e encontra, em países de clima tropical, ambiente ideal para sua reprodução [20]. Estudos apontam que o vírus vem sofrendo mutações e, além de reproduzir em águas limpas, vem obtendo sucesso em ambientes com águas sujas, como, por exemplo, em esgotos [7].

As principais ações de combate à dengue se voltam contra a não proliferação do seu vetor. Para isso, há campanhas de conscientização da população solicitando o correto descarte de objetos, orientações sobre como armazenar água e a utilização de pesticidas [22].

2.2 Técnicas de *Machine Learning* e de *Deep Learning*

Random Forest (RF) é uma técnica de ML que, por meio da criação e do treinamento de diversas árvores de decisão, as previsões são feitas com base na média da previsão de cada árvore [28]. A Figura 1 ilustra o funcionamento básico da RF.

Support Vector Regression (SVR) é um método de regressão que utiliza vetores de suporte para encontrar uma função capaz de traçar um hiperplano contendo a maior parte dos dados de treinamento [2] [4]. A Figura 2 detalha a utilização dos vetores de suporte para encontrar o hiperplano.

A *Multilayer Perceptron* (MLP) é uma rede neural, do tipo *feedforward*, formada por neurônios [3]. A MLP é estruturada em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Com exceção da camada de saída, todos os nós estão totalmente conectados com os nós da camada seguinte [11] [14]. A arquitetura básica de uma MLP é demonstrada na Figura 3.

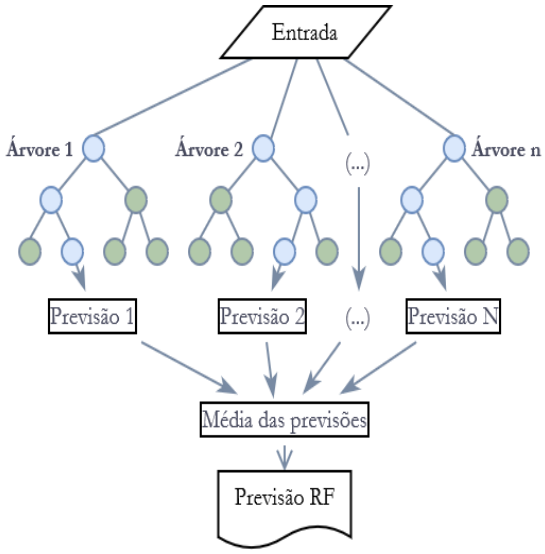

Figura 1: Funcionamento da RF Fonte: Autor

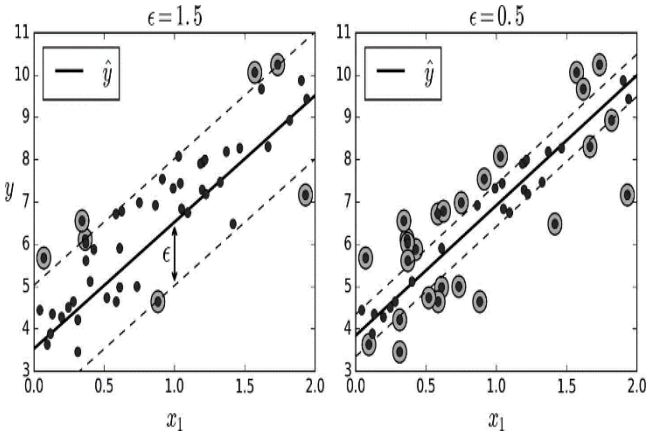

Figura 2: Hiperplano da SVR Fonte: Adaptado de [13]

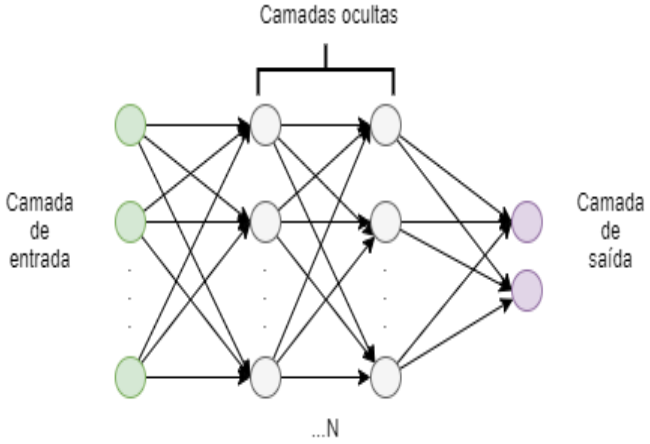

Figura 3: Arquitetura básica de uma MLP Fonte: Autor

Long Short-Term Memory (LSTM) faz parte das redes neurais recorrentes. A ideia por trás das LSTMs é a utilização de células capazes de decidir a curto e longo prazo quais dados devem ser incorporados ou esquecidos e resolver o problema de dependência de longo prazo [23]. O papel de decisão sobre a incorporação dos dados ou o seu descarte são, respectivamente, papéis do *input gate* e do *forget gate* [12] [16]. A representação de uma célula LSTM está presente na Figura 4.

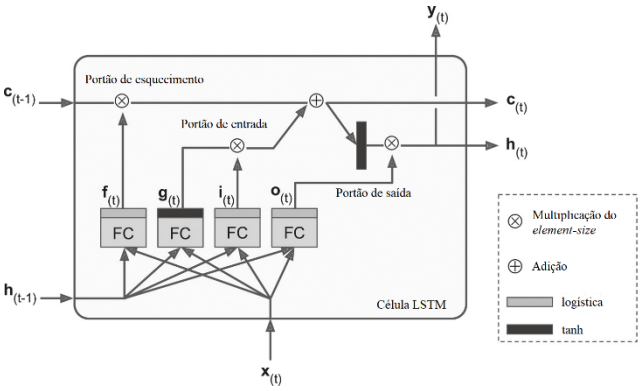

Figura 4: Funcionamento de uma célula LSTM Fonte: Adaptado de [13]

Convolutional Neural Network é uma técnica de DL muito empregada em processamento de imagens. Tradicionalmente, as redes neurais convolucionais são utilizadas para a extração de características importantes em imagens. Contudo, a convolução linear também vem sendo utilizada em previsões em razão da sua capacidade de capturar padrões nas séries temporais [33]. A Figura 5 relata o funcionamento de uma camada CNN aplicada a problemas de séries temporais.

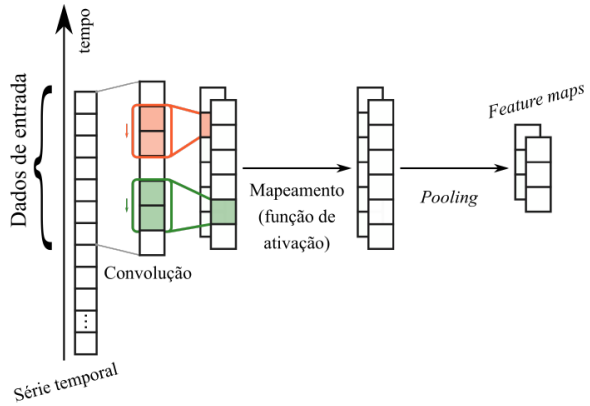

Figura 5: Camada CNN para séries temporais Fonte: [5]

3. TRABALHOS RELACIONADOS

A previsão de doenças não é uma tarefa fácil. Existem vários fatores influenciadores e impactantes durante a predição, como, por exemplo, fatores climáticos, fatores econômicos, fatores sociais, mobilidade urbana, entre outros [21]. Devido à complexidade, inúmeros trabalhos estão utilizando ML e DL durante a predição de doenças.

Doni and Sassipraba [12], na Índia, conduziram um estudo utilizando as técnicas LSTM, SVR, *Extreme Gradient Boosting* (XGboost), RF e *Generalized additive model* (GAM). Na criação dos modelos foram utilizados dados epidemiológicos e climáticos, fornecidos pelo governo, entre os anos de 2015 e 2018. A técnica LSTM obteve o melhor resultado com um *Root mean square error* (RMSE) de 42,00.

Xu et al. [32] realizaram predições de casos de dengue na China. Utilizando dados meteorológicos e as técnicas LSTM, *Back propagation neural network* (BPNN), GAM e SVR. A LSTM obteve a menor taxa de erro, RMSE 36,50.

Mussumeci and Codeço [21] propuseram a previsão de casos de dengue nos municípios do estado do Rio de Janeiro, utilizando as técnicas de LSTM, RF *e Least absolute shrinkage and selection operator* (LASSO), dados climáticos e históricos da doença. Como resultado, a técnica LSTM obteve uma taxa de erro de 0,45.

Carvajal et al. [8], nas Filipinas, utilizaram dados climáticos para a previsão de dengue. As técnicas utilizadas foram RF e GAM e a verificação dos valores ficou por meio do RMSE. RF obteve a menor taxa de erro: 0,29.

Guo et al. [15], utilizando dados históricos de dengue, entre 2011 e 2014, na China, conseguiram prever casos da doença com um RMSE 0,2861 através da SVR.

Appice et al. [2], no México, propuseram a criação de modelos de previsão de casos de dengue utilizando as técnicas *AUTOencoding based TIme series Clustering with Nearest Neighbour* (AutoTic-NN), *K-Nearest Neighbourhood* (KNN), SVR e *Autoregressive integrated moving average* (ARIMA). Para a produção dos seus modelos, foram utilizados dados históricos da doença e climáticos entre 1985 e 2010.O AutoTic-NN obteve a menor taxa de erro, RMSE 5,18.

4. METODOLOGIA

4.1 Base de dados

Para a criação dos modelos de predição deste estudo, foram utilizados os números mensais das internações causadas por dengue nos municípios. As internações tiveram como fonte o Sistema de Informações Hospitalares do SUS (SIH/SUS) entre os anos de 2010 e 2019. Além dos dados históricos da doença, para cada município, foi utilizado o valor mensal de precipitação. A pluviometria foi fornecida pela Agência Executiva de Gestão das Águas do Estado da Paraíba (AESA).

Do Sistema Nacional de Informações sobre Saneamento (SNIS), foram coletadas informações sobre água e esgoto entre os anos de 2010 e 2019 e incorporadas ao trabalho. Após a análise qualitativa, optou-se por utilizar os dados de índice de coleta de esgoto e o de tratamento de esgoto. Por fim, foi criado um banco de dados relacional, chamado *DashDengue*. A Figura 6 demonstra o modelo relacional do banco.

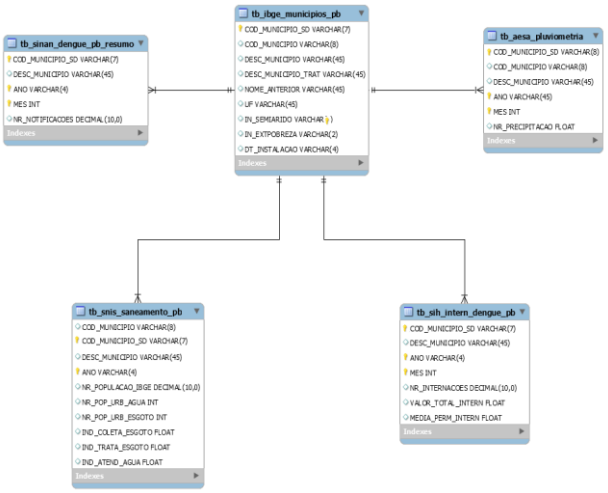

Figura 6 - Modelo relacional do banco *DashDengue* Fonte: Autor

4.2 Sistema para predição de casos de internação

O desenvolvimento do sistema de predição foi realizado utilizando a linguagem *Python*. Ademais, foram empregadas as bibliotecas *Scikit-learn* [25] e *TensorFlow* [1].

As fases de desenvolvimento foram divididas em: criação, treinamento e escolha dos modelos, ajustes nos hiperparâmetros, geração das previsões e, por fim, validação estatística dos resultados.

4.2.1 Criação, treinamento e escolha dos modelos

A primeira etapa do processo foi recuperar os dados da base *DashDengue* e normalizá-los. De posse das informações, foram criados quatros cenários de pesquisa (Tabela 1) variando a quantidade de anos dos dados e se haveria ou não tratamento de observações fora do comum (*outliers*).

Tabela 1: Cenários de pesquisa Fonte: Autor

Cenário	Período início	Período fim	Tratar *outlier*
Cenário 1	2010	2019	Não
Cenário 2	2015	2019	Não
Cenário 3	2010	2019	Sim
Cenário 4	2015	2019	Sim

Na sequência, foram propostos 8 modelos candidatos contendo as combinações dos atributos: número de internação, pluviometria mensal, índice de coleta de esgoto e índice de tratamento de esgoto. As variações estão dispostas na Tabela 2. Finalmente, para cada atributo, foram adicionadas de 1 a 4 *lags* (informações do passado para os atributos) com os dados dos últimos meses.

Tabela 2: Combinação dos atributos previsores para a previsão de internações Fonte: Autor

#	nr_internações	precipitação	coletaEsgoto	tratamentoEsgoto
1	Sim	Sim	Sim	Sim
2	Sim	Sim	Não	Não
3	Sim	Não	Sim	Não
4	Sim	Não	Não	Sim
5	Sim	Sim	Sim	Não
6	Sim	Sim	Não	Sim
7	Sim	Não	Sim	Sim
8	Sim	Não	Não	Não

Definidos os modelos, eles foram submetidos ao treinamento através das técnicas RF, SVR, MLP, LSTM e CNN. A escolha por essas técnicas foi feita conforme uma revisão sistemática realizada previamente [6]. Sobre as configurações, o RF foi executado com *n_estimators*=50. Para o SVR, foram utilizados os parâmetros padrões do *Scikit-learn*. Em relação à MLP foram adicionadas duas camadas ocultas com 250 neurônios. Parâmetros não citados seguiram os valores padrão do *Scikit-Learn*.

A técnica LSTM foi implementada com base na biblioteca *TensorFlow* e possui a seguinte característica: camada de entrada, camada LSTM com 50 neurônios e uma camada densa. A quantidade de épocas foi definida em 50 e o *batch_size* em 12.

Finalmente, a técnica CNN é a combinação de camadas convolucional e LSTM. A primeira camada é a de entrada seguida de uma camada Conv1D, com 16 filtros. Na sequência foi adicionada uma camada LSTM com 50 neurônios e, por fim, uma camada densa. A quantidade de épocas foi de 50 e o *batch_size*=12. Os Parâmetros não listados assumiram os valores padrão do *TensorFlow*.

Os dados foram separados entre treino e teste, seguindo a proporção de 80% para treino e 20% para teste. Finalizado o treinamento, foram geradas previsões e, utilizando a técnica RMSE [8], a taxa de erro foi mensurada. Sobre o RMSE, quanto mais baixo for o seu valor melhor é a previsão. O modelo e cenário com melhores resultados foram os escolhidos para seguir para a próxima etapa. A Figura 7 demonstra as atividades realizadas durante a etapa de criação e escolha dos modelos.

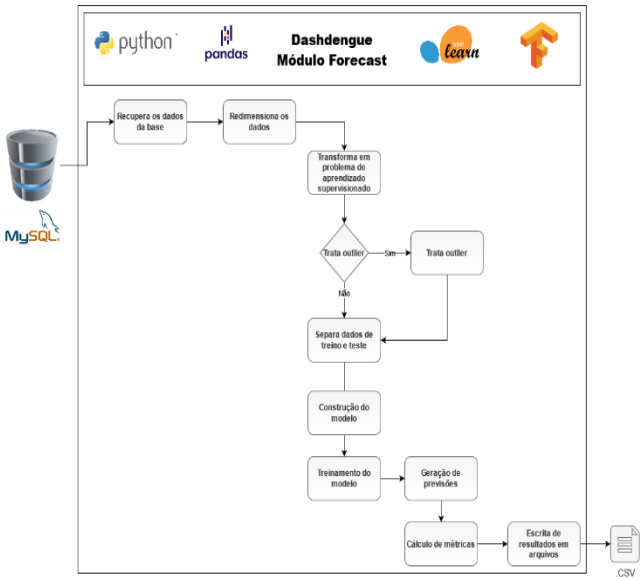

Figura 7: Etapas de criação e escolha dos modelos Fonte: Autor

4.2.2 Ajustes nos hiperparâmetros

Na fase de ajuste de parâmetros, os modelos vencedores foram submetidos a variações dos parâmetros. A Tabela 3 contém os parâmetros utilizados bem como os seus valores.

Tabela 3: Hiperparâmetros por técnica Fonte: Autor

Técnica	Parâmetro	Lista de valores
RF	*n_estimators*	{25,50,100,500,1000}
	min_samples_split	{2,5,10,20}
	min_samples_leaf	{1,2,5,10,20}
SVR	*kernel*	{*'rbf','poly','sigmoid'*}
	gamma	{*'scale','auto'*}
	C	{1.0,5.0,10.0,25.0}
	epsilon	{0.1,1.0,2.0}
MLP	*activation*	{*'logistic', 'tanh', 'relu'*}
	solver	{*'sgd', 'adam'*}
	batch_size	{'auto',12,24,48}
	learning_rate	{*'constant', 'adaptive'*}
	Número de camadas ocultas	{1,2}
	Número de neurônios camada 1	{25,50,100,250}
	Número de neurônios camada 2	{0,25,50,100,250}
LSTM	*neurons*	{25,50,100,250,500}
	activation	{*'tanh','softmax','relu'* }
	recurrent_activation	{*'relu','tanh','sigmoid'* }
	dropout	{0.0, 0.1, 0.2}
	batch_size	[*None*,12,24,48]
CNN	*filters*	{2,8,16}
	neurons	{25,50,100,250,500}
	activation	{*'tanh','softmax','relu'* }
	recurrent_activation	{*'relu','tanh','sigmoid'* }
	dropout	{0.0, 0.1, 0.2}
	batch_size	{*None*,12,24,48}

Os parâmetros e seus valores foram escolhidos após prévios testes. A configuração com melhor taxa de erro foi utilizada para a previsão dos casos de internação.

4.2.3 Geração de previsões, avaliação dos resultados e comprovações estatísticas

De posse da melhor configuração de modelo, cenário e parâmetros, os dados foram submetidos ao treinamento, à geração das previsões numéricas de internações e ao cálculo da taxa de erro. Por questões estocásticas de algumas técnicas, separados em quatros *rounds*, foram realizadas 100 execuções para cada uma das configurações. Com intuito de evitar o *overfitting*, a técnica de parada antecipada foi implementada.

Não foram encontrados na literatura estudos realizando previsões de internação para os municípios aqui listados. Com isso, adotou-se o procedimento de comparar os resultados com a abordagem *Naive Forecast* [18] [29] e ARIMA [2]. Em relação à validação estatística, foi verificada a normalidade dos resultados usando o teste de Shapiro-Wilk. Se os resultados obedecessem à curva normal, as diferenças estatísticas seriam computadas através dos testes Anova [27] e Tukey [17]. Caso contrário, foram usados Kruskall e Dunn.

5. RESULTADOS E DISCUSSÃO

O primeiro objetivo deste projeto foi verificar qual modelo e cenário produziu os melhores resultados para as cidades alvo. Após a execução das combinações, um *script* em python determinou qual ensaio obteve a menor taxa de erro (RMSE).

Os resultados da escolha dos parâmetros previsores, a quantidade anos a ser utilizada, se haverá ou não tratamento de *outliers* e a quantidade de *lags* estão representados na Tabela 4. Adicionalmente, a menor taxa de erro para a combinação também é demonstrada.

Tabela 4: Resultados contendo os modelos vencedores e suas configurações

Município	RMSE	Parâmetros	Anos	*Outlier*	*Lags*
Bayeux	0,629282246	Internações, precipitação e índice de coleta de esgoto	2010 - 2019	Não	4
Cabedelo	0,955240078	Internações, precipitação e índice de coleta de esgoto	2015 - 2019	Sim	4
Cajazeiras	1,342038155	Internações e precipitação	2015 - 2019	Sim	3
João Pessoa	10,10272226	Internações	2010 – 2019	Sim	3
Patos	0,454293686	Internações e precipitação	2015 - 2019	Sim	3
Santa Rita	1,081508751	Internações, precipitação e índice de coleta de esgoto	2015 - 2019	Sim	3

Como notado, para os municípios da Mata Paraibana: Bayeux, Cabedelo, João Pessoa e Santa Rita, a combinação de parâmetros de internações, precipitação e índice de coleta de esgoto obteve melhores resultados em três das quatro cidades. A não utilização de informações de esgoto, em João Pessoa, pode ser explicada devido à maior média de coleta de esgoto (69,37%) e de tratamento de esgoto (99,95%) dessa cidade comparada as outras três.

Os municípios do Sertão, Cajazeiras e Patos, utilizaram como atributos previsores os dados de internações e de pluviometria. Sobre o tratamento de *outliers*, o artifício foi utilizado em cinco das seis cidades do estudo. Em relação aos anos de dados

utilizados, em 66% das cidades usaram informações entre 2010 e 2015. Enfim, o número de *lags* variou entre 4 e 3.

A Tabela 5 ilustra os resultados obtidos após os ajustes de hiperparâmetros e qual técnica foi a vencedora em cada cidade.

Tabela 5: Resultados após ajustes de parâmetros e técnica vencedora

Município	Melhor RMSE	Técnica vencedora
Bayeux	0,529017139	LSTM
Cabedelo	0,927428107	LSTM
Cajazeiras	1,000751246	LSTM
João Pessoa	9,552880644	CNN
Patos	0,422049567	CNN
Santa Rita	0,745519952	RF

A estratégia de ajuste de parâmetros melhorou a taxa de erro para todas as cidades. A técnica LSTM venceu em 50% das cidades, com CNN vencendo em 33% e RF aparecendo como vencedora em uma cidade (17%). Durante a validação estatística, para todas as cidades, a curva normal foi seguida.

Para comprovar se há diferença estatística entre os RMSEs produzidos pelas técnicas, para cada cidade, foram realizados os testes ANOVA e Tukey considerando α=0,05. As seguintes hipóteses estatísticas foram consideradas:

- H_0: Estatisticamente os resultados produzidos são iguais se *p-value* > 0,05;
- H_1: Estatisticamente há diferença entre os resultados se *p-vaule* < 0,05.

A Figura 8 demonstra os resultados dos testes de Tukey para a cidade de Bayeux. Ao verificar apenas o RMSE a técnica LSTM é a mais indicada para realizar as previsões. Contudo, como pode ser observado, não há, estatisticamente, diferença entre os resultados obtidos pela técnica de menor taxa de erro (LSTM) em comparação com CNN. Isso ocorre, pois o *p-value* foi de 0,0647. Logo, a hipótese nula não pode ser rejeitada.

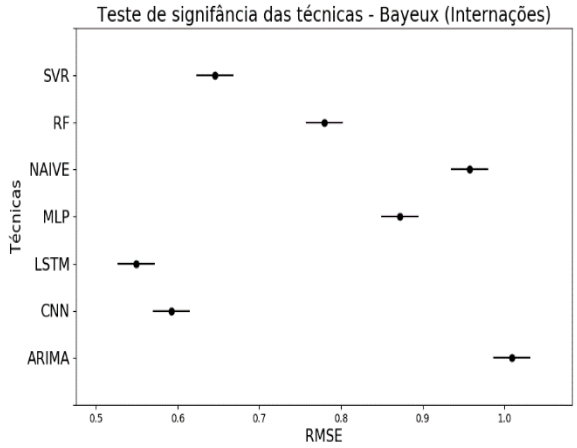

Figura 8: Testes de Tukey para a cidade de Bayeux Fonte: Autor

Esse fato pode ser corroborado ao analisar as curvas de previsões presentes na Figura 9.

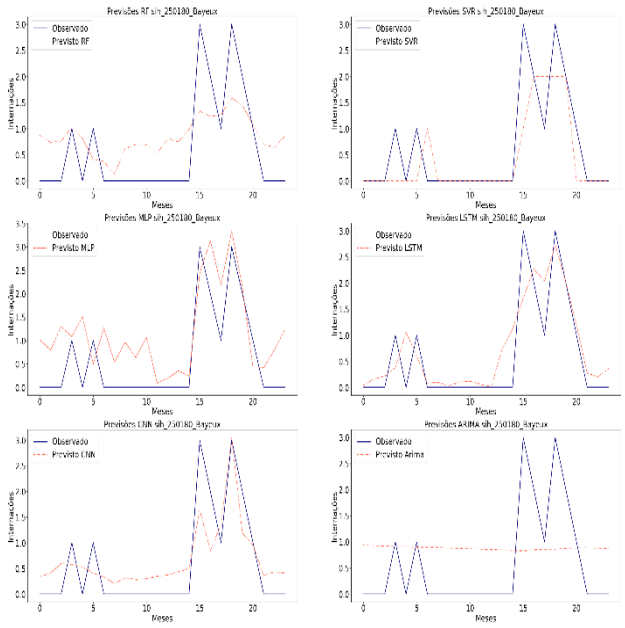

Figura 9: Previsões para a cidade de Bayeux Fonte: Autor

Os testes de Tukey para a cidade de Cabedelo são ilustrados na Figura 10. Para Cabedelo, LSTM obteve a menor taxa de erro e a sua superioridade ficou evidenciada frente as demais técnicas. Adicionalmente, foi comprovada a diferença estatística entre elas.

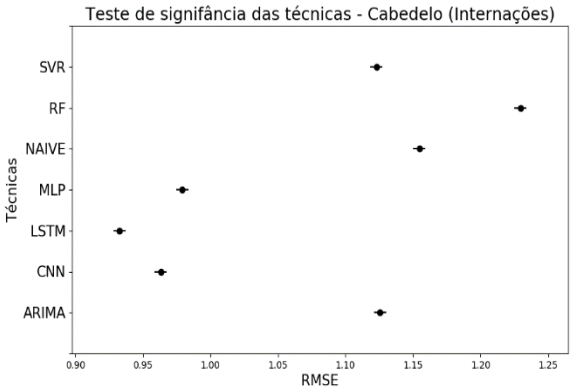

Figura 10: Testes de Tukey para a cidade de Cabedelo Fonte: Autor

A validação gráfica de Cabedelo, demonstrada na Figura 11, evidencia a melhor adequação das previsões produzidas pela LSTM.

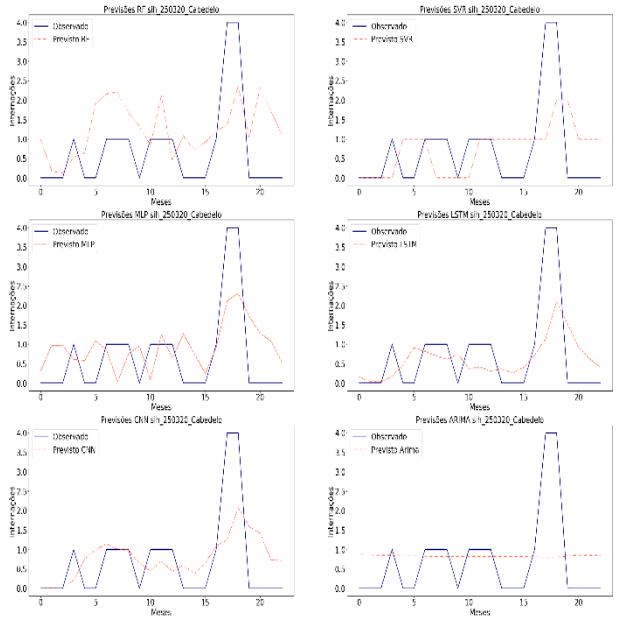

Figura 11: Previsões para a cidade de Cabedelo

Os resultados dos testes comparativos e a validação gráfica para a cidade de Cajazeiras estão presentes, respectivamente, nas figuras Figura 12 e Figura 13. A técnica LSTM conseguiu seguir a tendência de crescimento e redução dos casos de internações causadas por dengue para Cajazeiras e obteve o menor erro durante as suas previsões.

Adicionalmente, ficou comprovada a sua diferença estatística ao compara, par a par, os resultados com a CNN, RF, MLP, SVR, NAIVE e ARIMA.

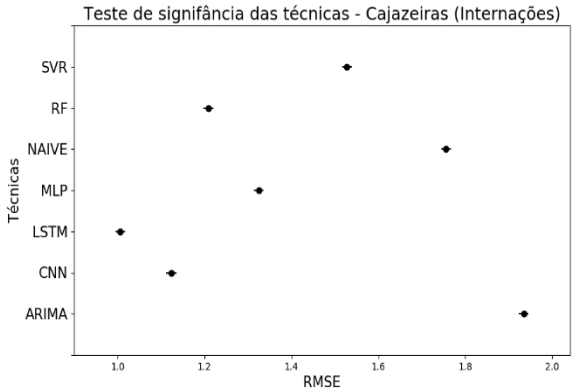

Figura 12: Testes de Tukey para a cidade de Cajazeiras Fonte: Autor

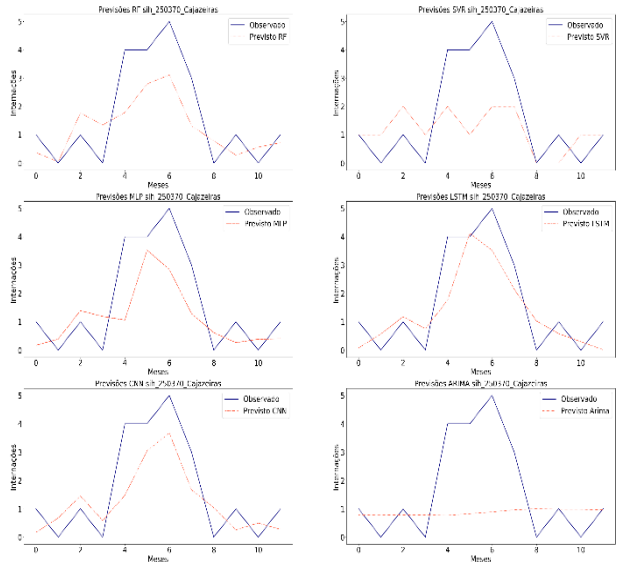

Figura 13: Previsões para a cidade de Cajazeiras Fonte: Autor

Os resultados para a cidade de João Pessoa informam a técnica CNN com menor taxa de erro.

A validação estatística, presente na Figura 14, relata que não há, estaticamente, diferença entre CNN e LSTM. O *p-value* da comparação da CNN com LSTM aponta um valor de 0,6386. Logo, a hipótese nula não pode ser rejeitada.

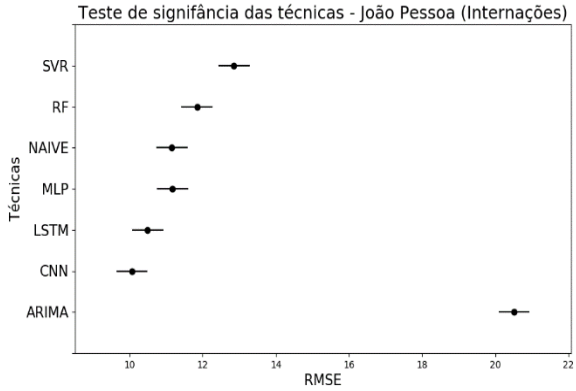

Figura 14: Testes de Tukey para a cidade de João Pessoa Fonte: Autor

A similaridade produzida entre as técnicas também pode ser observada através da validação gráfica presente na Figura 15.

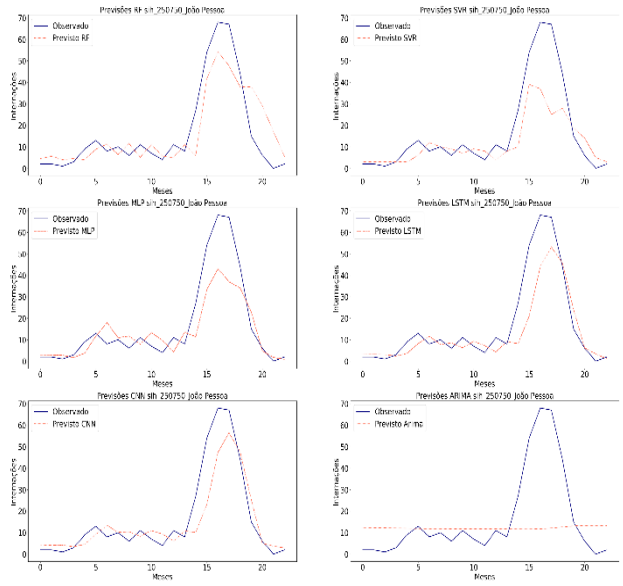

Figura 15: Previsões para a cidade de João Pessoa Fonte: Autor

Para a cidade de Patos, a técnica CNN obteve a menor taxa de erro durante as previsões e ficou comprovada a diferença estatística entre ela e as demais técnicas abordadas. Os resultados do teste de Tukey estão presentes na Figura 16.

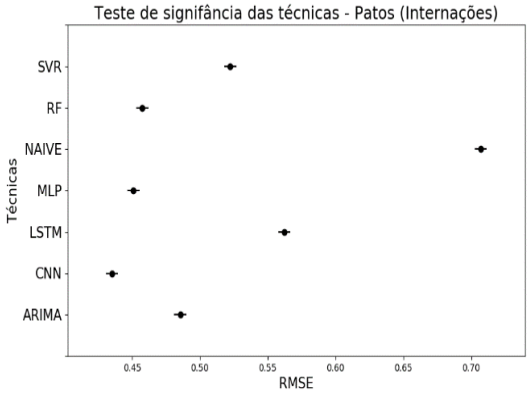

Figura 16: Testes de Tukey para a cidade de Patos Fonte: Autor

A validação gráfica (Figura 17) destaca a CNN como melhor técnica a seguir a linha de observações de casos de internações causadas por dengue.

O principal achado para as previsões de internações para Patos é a comprovação da eficácia da camada convolucional do modelo CNN. Ao observar os resultados produzidos pela LSTM, nota-se que a LSTM não conseguiu acompanhar a curva de internações. Contudo, a técnica CNN além de conseguir acompanhar a curva, obteve a menor taxa de erro durante as previsões para essa cidade.

Por fim, os testes de Tukey, presentes na Figura 18, informam superioridade e diferença estatística da técnica *Random Forest* durante as previsões na cidade de Santa Rita.

A validação gráfica para Santa Rita encontra-se presente na Figura 19. Como observado, a técnica RF conseguiu melhor adequação à curva casos de internações para essa cidade.

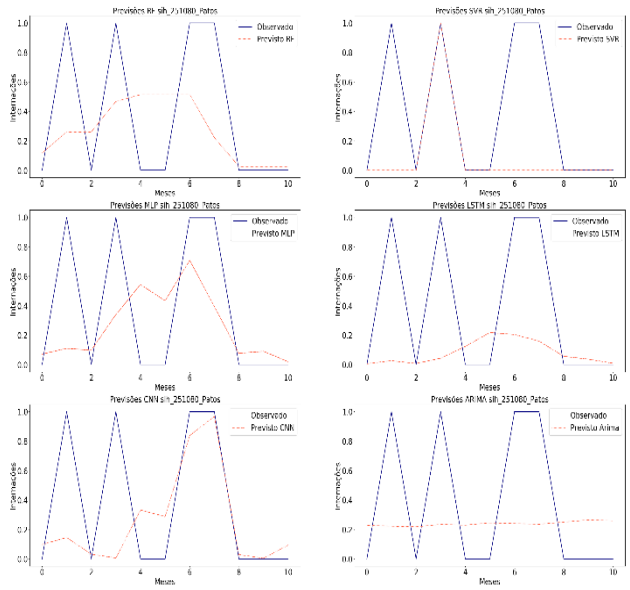

Figura 17: Previsões para a cidade de Patos Fonte: Autor

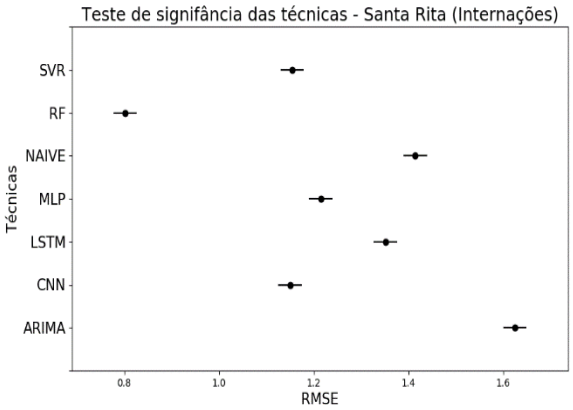

Figura 18: Testes de Tukey para a cidade de Santa Rita Fonte: Autor

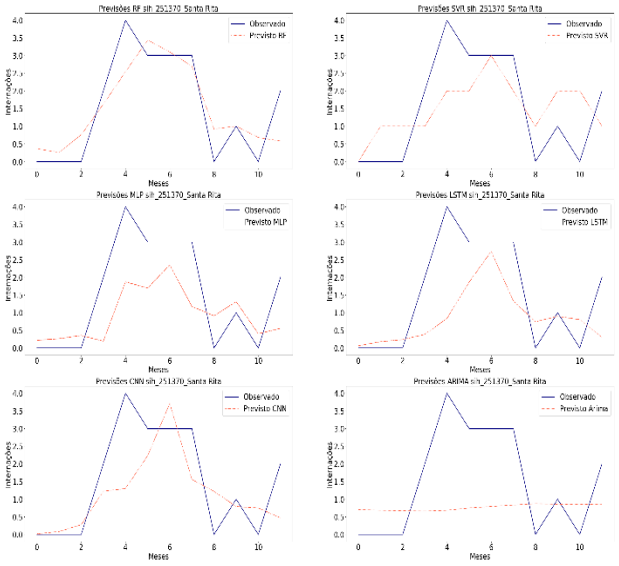

Figura 19: Previsões para a cidade de Santa Rita Fonte: Autor

6. CONSIDERAÇÕES FINAIS

Empregando dados epidemiológicos, climáticos e sanitários foi possível criar, avaliar e realizar previsões de casos internações causadas por dengue por meio de *Machine Learning* e de *Deep Learning*.

Foi constatado que as técnicas de *Deep Learning* (LSTM e CNN) saíram vencedoras em 83% das cidades. O recurso de tratamento para exclusão de *outilers* foi realizado em cinco das seis cidades. Logo, foi demonstrado que há necessidade de tratar os dados antes de realizar as previsões. Em relação à quantidade de anos, na maioria das escolhas, os anos de 2015 e de 2019 foram utilizados. Por fim, ficou demonstrada, estatisticamente, a diferença entre as técnicas e a superioridade das abordagens de DL frente a ML.

A escolha dos atributos previsores demonstra consistentes achados, pois, na grande maioria, municípios da mesma região geográfica utilizaram os mesmos atributos para realizar as previsões. Ademais, a relação de índice de coleta de esgoto, nos municípios da Mata Paraibana, demonstra uma oportunidade de pesquisa para investigar a influência da rede sanitária no número de casos de dengue.

Para trabalhos futuros, é sugerido a verificação de utilização de outros parâmetros para a criação dos modelos, tais como: índice de coleta de resíduos, indicadores de umidade e temperatura média. Além disso, é válido a utilização de outras técnicas como redes neurais recorrentes e novas variações das técnicas aqui estudadas. Como limitações deste trabalho, destaca-se a ausência de dados do ano 2020 e a falta de testes de homoscedasticidade antes de realizar os testes Anova e Tukey.

7. Referências

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-Scale Machine Learning. Retrieved July 21, 2021 from https://tensorflow.org.

[2] A. Appice, Y.R. Gel, I. Iliev, V. Lyubchich, and D. Malerba. 2020. A Multi-Stage Machine Learning Approach to Predict Dengue Incidence: A Case Study in Mexico. IEEE Access 8, (2020), 52713–52725. DOI:https://doi.org/10.1109/ACCESS.2020.2980634

[3] Rahul Awad, Mariette; Khanna. 2015. Efficient Learning Machines Theories, Concepts, and Applications for Engineers and System Designers. Springer nature.

[4] Mariette Awad and Rahul Khanna. 2015. Efficient learning machines: Theories, concepts, and applications for engineers and system designers. Apress Media LLC. DOI:https://doi.org/10.1007/978-1-4302-5990-9

[5] Felipe O Barino and Alexandre dos Santos Bessa. 2020. Rede Neural Convolucional 1D aplicada à previsão da vazão no Rio Madeira. (2020). DOI:https://doi.org/10.14209/SBRT.2020.1570640893

[6] Ewerthon Dyego de Araujo Batista, Wellington Candeia de Araújo, Romeryto Vieira Lira, and Laryssa Izabel de Araujo Batista. 2021. Previsão de casos de dengue através de Machine Learning e Deep Learning: uma revisão sistemática. Res. Soc. Dev. 10, 11 (August 2021), e33101119347. DOI:https://doi.org/10.33448/rsd-v10i11.19347

[7] Eduardo B. Beserra, Eraldo M. de Freitas, José T. de Souza, Carlos R. M. Fernandes, and Keliana D. Santos. 2009. Ciclo de vida de Aedes (Stegomyia) aegypti (Diptera, Culicidae) em águas com diferentes características. Iheringia. Série Zool. 99, 3 (2009), 281–285. DOI:https://doi.org/10.1590/S0073-47212009000300008

[8] T.M. Carvajal, K.M. Viacrusis, L.F.T. Hernandez, H.T. Ho, D.M. Amalin, and K. Watanabe. 2018. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. BMC Infect. Dis. 18, 1 (2018). DOI:https://doi.org/10.1186/s12879-018-3066-0

[9] W Carvalho, T. M., Tenório, G. L., Figueiredo, K., Vellasco, M., Caarls. 2019. Comparison of Machine Learning Models for Total Dengue Cases Prediction. (2019).

[10] Luis Hernan Contreras Pinochet. 2011. Tendências de Tecnologia de Informação na Gestão da Saúde. Mundo saúde (1995) (2011), [382-394]. Retrieved October 20, 2021 from http://bvsms.saude.gov.br/bvs/artigos/tendencias_tecnologia_informacao_gestao_saude.pdf

[11] Harvey DEITEL, Paul J.; DEITEL. 2019. Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and the Cloud. Pearson.

[12] A.R. Doni and T. Sasipraba. 2020. Lstm-Rnn Based Approach for Prediction of Dengue Cases in India. Ing. des Syst. d'Information 25, 3 (2020), 327–3355. DOI:https://doi.org/10.18280/isi.250306

[13] Aurélien Géron. 2019. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow. Alta Books, Rio de Janeiro.

[14] Aaron Goodfellow, Ian ; Bengio, Yoshua; Courviller. 2016. Deep learning. Mit press.

[15] Pi Guo, Tao Liu, Qin Zhang, Li Wang, Jianpeng Xiao, Qingying Zhang, Ganfeng Luo, Zhihao Li, Jianfeng He, Yonghui Zhang, and Wenjun Ma. 2017. Developing a dengue forecast model using machine learning: A case study in China. PLoS Negl. Trop. Dis. 11, 10 (2017). DOI:https://doi.org/10.1371/journal.pntd.0005973

[16] Matt Harisson. 2020. Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python. Novatec, São Paulo.

[17] Gopal Kanji. 2006. 100 Statistical Tests (3rd Editio ed.). Sage, London. DOI:https://doi.org/10.4135/9781849208499

[18] Simon Kirby, Kanya Paramaguru, and James Warren. 2015. The Accuracy of NIESR's GDP Growth Forecasts. Natl. Inst. Econ. Rev. 232, 1 (May 2015), 41–47. DOI:https://doi.org/10.1177/002795011523200114

[19] André Andrade Longaray and Tiago Machado Castelli. 2020. Avaliação do desempenho do uso da tecnologia da informação na saúde: revisão sistemática da literatura sobre o tema. Cien. Saude Colet. 25, 11 (November 2020), 4327–4338. DOI:https://doi.org/10.1590/1413-812320202511.26342018

[20] Cecilia de Almeida Marques-Toledo, Carolin Marlen Degener, Livia Vinhal, Giovanini Coelho, Wagner Meira, Claudia Torres Codeço, and Mauro Martins Teixeira. 2017. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. PLoS Negl. Trop. Dis. 11, 7 (July 2017). DOI:https://doi.org/10.1371/journal.pntd.0005729

[21] E. Mussumeci and F. Codeço Coelho. 2020. Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression. Spat. Spatiotemporal. Epidemiol. 35, (2020). DOI:https://doi.org/10.1016/j.sste.2020.100372

[22] R. Norrby. 2014. Outlook for a dengue vaccine. Clinical Microbiology and Infection 20, 92–94. DOI:https://doi.org/10.1111/1469-0691.12442

[23] Ilyas Ozer, Onursal Cetin, Kutlucan Gorur, and Feyzullah Temurtas. 2021. Improved machine learning performances with transfer learning to predicting need for hospitalization in arboviral infections against the small dataset. Neural Comput. Appl. (2021). DOI:https://doi.org/10.1007/s00521-021-06133-0

[24] Paraíba. 2021. Boletim Epidemiológico de Arbovirores - Julho - 2021. Retrieved August 24, 2021 from https://paraiba.pb.gov.br/diretas/saude/arquivos-1/vigilancia-em-saude/be_arbo_07_2021-2-6-2.pdf

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 85 (2011), 2825–2830. Retrieved July 21, 2021 from http://jmlr.org/papers/v12/pedregosa11a.html

[26] Duc Nghia Pham, Tarique Aziz, Ali Kohan, Syahrul Nellis, Juraina Binti Abd Jamil, Jing Jing Khoo, Dickson Lukose, Sazaly AbuBakar, Abdul Sattar, and Hong Hoe Ong. 2018. How to Efficiently Predict Dengue Incidence in Kuala Lumpur. In Proceedings - 2018 4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018. DOI:https://doi.org/10.1109/ICACCAF.2018.8776790

[27] Jinsheng Ren, Ke Qin, Ying Ma, and Guangchun Luo. 2014. On software defect prediction using machine learning. J. Appl. Math. 2014, (2014). DOI:https://doi.org/10.1155/2014/785435

[28] Peter Russell, Stuart; Norvig. 2013. Inteligência Artificial. Elsevier, Rio de Janeiro.

[29] Galit; SHMUELI and Kenneth C LICHTENDAHL JR. 2016. Practical time series forecasting with r: A hands-on guide (2nd ed.). Axelrod Schnall Publishers.

[30] Sathyamangalam Swaminathan and N. Khanna. 2019. Dengue vaccine development: Global and Indian scenarios. Int. J. Infect. Dis. 84, (July 2019), S80–S86. DOI:https://doi.org/10.1016/j.ijid.2019.01.029

[31] Vanessa Teich, Roberta Arinelli, and Lucas Fahham. 2017. Aedes aegypti e sociedade: o impacto econômico das arboviroses no Brasil. J. Bras. Econ. da Saúde 9, 3 (December 2017), 267–276. DOI:https://doi.org/10.21115/JBES.V9.N3.P267-76

[32] Jiucheng Xu, Keqiang Xu, Zhichao Li, Fengxia Meng, Taotian Tu, Lei Xu, and Qiyong Liu. 2020. Forecast of dengue cases in 20 chinese cities based on the deep learning method. Int. J. Environ. Res. Public Health 17, 2 (2020). DOI:https://doi.org/10.3390/ijerph17020453

[33] Kui Zhao and Can Wang. 2017. Sales Forecast in E-commerce using Convolutional Neural Network. (2017). Retrieved from http://arxiv.org/abs/1708.07946