UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I – CAMPINA GRANDE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E TECNOLOGIA EM SAÚDE
MESTRADO PROFISSIONAL EM CIÊNCIA E TECNOLOGIA EM SAÚDE

EWERTHON DYEGO DE ARAUJO BATISTA

UTILIZAÇÃO DE TÉCNICAS DE MACHINE LEARNING E DE DEEP LEARNING PARA A PREDIÇÃO DE CASOS DE DENGUE NOS MUNICÍPIOS DA PARAÍBA

CAMPINA GRANDE
2021

EWERTHON DYEGO DE ARAUJO BATISTA

UTILIZAÇÃO DE TÉCNICAS DE MACHINE LEARNING E DEEP LEARNING PARA A PREDIÇÃO DE CASOS DE DENGUE NOS MUNICÍPIOS DA PARAÍBA

Dissertação apresentada ao Programa de Pós-graduação em Ciência e Tecnologia em Saúde da Universidade Estadual da Paraíba como parte dos requisitos para obtenção do título de Mestre em Ciência e Tecnologia em Saúde.

Orientador: Prof. Dr. Wellington Candeia de Araújo

CAMPINA GRANDE
2021

B333u Batista, Ewerthon Dyego de Araujo.

Utilização de técnicas de *machine learning* e de *deep learning* para a predição de casos de dengue nos municípios da Paraíba [manuscrito] / Ewerthon Dyego de Araujo Batista. - 2021.

119 p.

Digitado.

Dissertação (Mestrado em Profissional em Ciência e Tecnologia em Saúde) - Universidade Estadual da Paraíba, Pró-Reitoria de Pós-Graduação e Pesquisa , 2021.

"Orientação : Prof. Dr. Wellington Candeia de Araújo , Coordenação do Curso de Computação - CCT."

1. Dengue. 2. Machine learning. 3. Deep learning. 4. Inteligência artificial. I. Título

21. ed. CDD 006.3

Elaborada por Lucas H. A. da Silva - CRB - 15/898 BC/UEPB

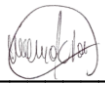
EWERTHON DYEGO DE ARAUJO BATISTA

UTILIZAÇÃO DE TÉCNICAS DE MACHINE LEARNING E DEEP LEARNING PARA A PREDIÇÃO DE CASOS DE DENGUE NOS MUNICÍPIOS DA PARAÍBA

Dissertação apresentada ao Programa de Pós-graduação em Ciência e Tecnologia em Saúde da Universidade Estadual da Paraíba como parte dos requisitos para obtenção do título de Mestre em Ciência e Tecnologia em Saúde.

Aprovada em: 07/10/2021.

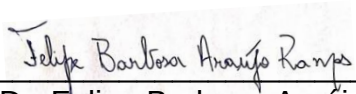
BANCA EXAMINADORA



Prof. Dr. Wellington Candeia de Araújo (Orientador)
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Frederico Moreira Bublitz
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Danilo de Almeida Vasconcelos
Universidade Estadual da Paraíba (UEPB)



Prof. Dr. Felipe Barbosa Araújo Ramos
Instituto Federal da Paraíba (IFPB)

Ao meu filho, Eliel Batista, DEDICO.

AGRADECIMENTOS

A DEUS, pela dádiva da vida, pela saúde, pelo discernimento e por todas as conquistas alcançadas em minha vida.

Aos meus pais, Edvandio e Nedja, que não mediram esforços e sacrifícios para a formação pessoal e educacional oferecida a mim e a minhas irmãs. Peço a Deus que eu possa ser para meu filho, ao menos, 10% dos pais que são para nós.

À minha esposa, Jessica Batista, pelo apoio incondicional, pelo companheirismo, por ser meu ponto de equilibro, por todas as mudanças positivas causadas em minha vida e, finalmente, porém, não exaustivamente, por proporcionar e por gerar um dos maiores sonhos em nossas vidas: Eliel.

À minha amada avó, Vó Rita, por ser uma segunda mãe, por ter ajudado na minha criação, por todo amor e carinho demonstrado por mim. Agradeço a Deus por sua longeva e iluminada vida, que assim permaneça por muitos anos.

À minha irmã, Edvandia, por todo apoio, incentivo e fraternidade.

À minha irmã, Laryssa, pelo estímulo, apoio, correções textuais e por compartilhar uma infância cheia de lembranças e de momentos felizes.

Aos meus sogros, Eudo e Solange, pelo incentivo, acolhimento, orações e por gerar a vida de Jéssica. Os três primeiros agradecimentos também são estendidos aos meus cunhados, Patrício e Artur.

Ao professor Wellington Candeia, meu orientador, por todos os ensinamentos, pelas orientações e pelas contribuições deste trabalho. Agradeço também pela ajuda na transição do mundo coorporativo para o mundo acadêmico. Com toda certeza, levarei e replicarei seu jeito de trabalhar com meus alunos.

A todos os professores que fizeram parte da minha formação (desde o jardim de infância até a pós-graduação).

“Caminhando e cantando e seguindo a canção. Somos todos iguais, braços dados ou não.”

Geraldo Vandré

RESUMO

Dengue é uma doença causada pelo vírus DENV e transmitida para o homem através do mosquito *Aedes aegypti*. Embora não seja uma doença nova, ainda não existe uma vacina regulamentada no Brasil que possa ser usada sem restrição na população. Logo, o combate contra a doença é feito através de ações para eliminação do mosquito transmissor. Os números da dengue voltaram a crescer no Brasil e na Paraíba. De acordo com o sétimo boletim epidemiológico de arbovirose da Paraíba, houve um acréscimo de 53% dos casos de dengue em relação aos casos do ano anterior. O objetivo deste trabalho foi criar um sistema capaz de realizar previsões de notificações e de internações causadas por dengue nos municípios da Paraíba. Por meio de técnicas de *Machine Learning* (*Random Forest* e *Support Vector Regression*) e de *Deep Learning* (*Multilayer Perceptron, Long Short-Term Memory e Convolutional Neural Network*) e utilizando dados epidemiológicos, climáticos e sanitários, entre os anos de 2010 e 2019, o sistema foi capaz de encontrar a melhor combinação de atributos previsores, os melhores parâmetros para as técnicas, realizar previsões de casos de internações e de notificações causadas por dengue para os municípios paraibanos Bayeux, Cabedelo, Cajazeiras, Campina Grande, Catolé do Rocha, João Pessoa, Monteiro, Patos e Santa Rita, determinar quais técnicas produzem melhores resultados por cidade e, finalmente, foi demonstrada a diferença estatística entre as abordagens. Os resultados produzidos demonstram a superioridade das técnicas de *Deep Learning* em comparação as técnicas de *Machine learing*. Durante a previsão de casos de notificações, a técnica *Long Short-Term Memory* (LSTM) obteve melhores resultados em 66,67% das cidades, *Convolutional Neural Network* (CNN) em 22,22% e *Multilayer Perceptron* (MLP) em 11,11%. Em relação às internações, LSTM obteve menor taxa de erro em 33,34% dos munícipios, CNN, MLP e *Random Forest* (RF) obtiveram, cada uma delas, melhores resultados em 22,22% das cidades.

Palavras-Chave: Dengue. Previsão. *Machine Learning. Deep Learning*.

ABSTRACT

Dengue is a disease caused by the DENV virus and transmitted to humans through the Aedes aegypti mosquito. Although it is not a new disease, there is still no regulated vaccine in Brazil that can be used without restriction in the population. Therefore, the fight against the disease is done through actions to eliminate the transmitting mosquito. Dengue numbers returned to grow in Brazil and Paraíba. According to the seventh epidemiological bulletin of arbovirus in Paraíba, there was an increase of 53% of dengue cases in relation to the cases of the previous year. The objective of this work was to create a system capable of forecasting notifications and hospitalizations caused by dengue in the municipalities of Paraíba. Through Machine Learning (Random Forest and Support Vector Regression) and Deep Learning (Multilayer Perceptron, Long Short-Term Memory and Convolutional Neural Network) techniques and using epidemiological, climatic and sanitary data, between 2010 and 2019, the system was able to find the best combination of predictive attributes, the best parameters for the techniques, make predictions of cases of hospitalizations and notifications caused by dengue for the municipalities of Paraíba Bayeux, Cabedelo, Cajazeiras, Campina Grande, Catolé do Rocha, João Pessoa, Monteiro, Patos and Santa Rita, determine which techniques produce better results per city and, finally, the statistical difference between the approaches was demonstrated. The results produced demonstrate the superiority of Deep Learning techniques in comparison to Machine learning techniques. During notification case forecasting, the Long Short-Term Memory (LSTM) technique obtained better results in 66.67% of cities, Convolutional Neural Network (CNN) in 22.22% and Multilayer Perceptron (MLP) in 11.11 %. Regarding hospitalizations, LSTM had the lowest error rate in 33.34% of the municipalities, CNN, MLP and Random Forest (RF) each obtained better results in 22.22% of the cities.

Keywords: Dengue. Forecast. Machine Learning. Deep Learning.

LISTA DE FIGURAS

Figura 1 – Ciclo de vida do mosquito *Aedes aegypti* ---------- 23
Figura 2 – Número de casos prováveis de dengue no Brasil entre os anos de 2010 e 2020 ---------- 25
Figura 3 – Mapa da Paraíba contendo o levantamento LIRAa. A cor verde indica baixo risco (<1%), a cor amarela risco moderado (>1% e <3,9%) e a vermelha alto risco (>= 4%). ---------- 26
Figura 4 – Funcionamento de um problema de classificação através de um filtro de spam ---------- 27
Figura 5 – Ilustração de um problema de regressão ---------- 28
Figura 6 – Relação entre Inteligência Artificial, *Machine Learning* e *Deep Learning* 29
Figura 7 – Estrutura de funcionamento de uma regressão feita por *Random Forest* 30
Figura 8 – Representação do funcionamento da regressão através de SVR ---------- 32
Figura 9 – Arquitetura de rede neural *Multilayer Perceptron* (MLP) ---------- 33
Figura 10 – Funcionamento da célula LSTM ---------- 35
Figura 11 – Funcionamento da adaptação de uma rede convolucional para o processamento de série temporal ---------- 37
Figura 12 - Diagrama entidade relacionamento para as tabelas do *schema* DashDengue ---------- 47
Figura 13 - Representação do módulo de carga de dados ---------- 48
Figura 14 - Atividades realizadas pelo módulo *forecast* ---------- 51

LISTA DE QUADROS

Quadro 1 – Descrição dos parâmetros utilizados pelo algoritmo *Random Forest* ----31

Quadro 2 – Descrição dos parâmetros utilizados pela classe SVR ----------32

Quadro 3 – Parâmetros utilizados pelo algoritmo de MLP ----------34

Quadro 4 – Listagem dos parâmetros utilizados pelo *Keras* para a criação de rede neural recorrente com LSTM ----------36

Quadro 5 – Parâmetros utilizados pela classe Conv1D do *Keras* ----------38

Quadro 6 - Arquitetura e configurações dos modelos vencedores para os trabalhos relacionados ----------42

Quadro 7 - Cenários de pesquisa propostos ----------49

Quadro 8 - Combinação dos atributos previsores para a previsão de internações---49

Quadro 9 - Combinação dos atributos previsores para a previsão de notificações --50

Quadro 10 - Arquitetura e configuração dos parâmetros utilizados no treinamento dos modelos ----------52

Quadro 11 - Variações de parâmetros e lista de valores candidatos por técnica de ML e DL ----------55

LISTA DE TABELAS

Tabela 1 - Taxa de erro calculada para previsões de internações com dados entre 2010 e 2019 e sem tratamento de *outlier* (cenário 1) ----------59

Tabela 2 - Taxa de erro calculada para previsões de internações com dados entre 2015 e 2019 e sem exclusão de *outliers* (cenário 2) ----------59

Tabela 3 - Taxa de erro calculada para previsões de internações com dados entre 2010 e 2019 e com exclusão de *outliers* (cenário 3) ----------60

Tabela 4 - Taxa de erro calculada para previsões de internações com dados entre 2015 e 2019 e com exclusão de *outliers* (cenário 4) ----------60

Tabela 5 - Resultados contendo a melhor combinação de parâmetros, período, se houve exclusão de *outliers* e a quantidade de *lags* para os casos de internações---61

Tabela 6 - Taxa de erro calculada para previsões de notificações com dados entre 2010 e 2019 e sem tratamento de *outliers* (cenário 1)----------62

Tabela 7 - Taxa de erro calculada para previsões de notificações com dados entre 2015 e 2019 e sem tratamento de *outliers* (cenário 2)----------63

Tabela 8 - Taxa de erro calculada para previsões de notificações com dados entre 2010 e 2019 e com tratamento de *outliers* (cenário 3)----------63

Tabela 9 - Taxa de erro calculada para previsões de notificações com dados entre 2015 e 2019 e com tratamento de *outliers* (cenário 4)----------64

Tabela 10 - Resultados contendo a melhor combinação de parâmetros, período, se houve exclusão de *outliers* e a quantidade de *lags* para os casos de notificações --64

Tabela 11 - Melhores configurações por cidade para previsão de internações através da técnica *Random Forest* ----------66

Tabela 12 - Melhores configurações por cidade para previsão de internações através da técnica *Support Vector Regression*----------67

Tabela 13 - Melhores configurações por cidade para previsão de internações através da técnica *Multilayer Perceptron* ----------67

Tabela 14 - Melhores configurações por cidade para previsão de internações através da técnica *Long short-Term memory* ----------68

Tabela 15 - Melhores configurações por cidade para previsão de internações através da técnica *Convolutional neural network* ----------68

Tabela 16 - Melhores configurações por cidade para previsão de notificações através da técnica *Random Forest* ----------69

Tabela 17 - Melhores configurações por cidade para previsão de notificações através da técnica *Support Vector Regression* ----------69

Tabela 18 - Melhores configurações por cidade para previsão de notificações através da técnica *Multilayer Perceptron* ----------70

Tabela 19 - Melhores configurações por cidade para previsão de notificações através da técnica *Long short-Term memory* ----------70

Tabela 20 - Melhores configurações por cidade para previsão de notificações através da técnica *Convolutional neural network* ----------71

Tabela 21 - Melhores resultados e técnica vencedora para a previsões de internações ----------72

Tabela 22 - Melhores resultados e técnica vencedora para a previsões de notificações ----------73

LISTA DE GRÁFICOS

Gráfico 1 - Análise de significância estatística para as previsões de internações para a cidade de Bayeux ----------74

Gráfico 2 - Previsões de internações por técnica para a cidade de Bayeux ----------75

Gráfico 3 - Análise de significância estatística para as previsões de internações para a cidade de Cabedelo ----------76

Gráfico 4 - Previsões de internações por técnica para a cidade de Cabedelo --------77

Gráfico 5 - Análise de significância estatística para as previsões de internações para a cidade de Cajazeiras ----------78

Gráfico 6 - Previsões de internações por técnica para a cidade de Cajazeiras -------78

Gráfico 7 - Análise de significância estatística para as previsões de internações para a cidade de Campina Grande ----------79

Gráfico 8 - Previsões de internações por técnica para a cidade de Campina Grande ----------80

Gráfico 9 - Análise de significância estatística para as previsões de internações para a cidade de Catolé do Rocha----------80

Gráfico 10 - Previsões de internações por técnica para a cidade de Catolé do Rocha ----------81

Gráfico 11 - Análise de significância estatística para as previsões de internações para a cidade de João Pessoa----------82

Gráfico 12 - Previsões de internações por técnica para a cidade de João Pessoa --82

Gráfico 13 - Análise de significância estatística para as previsões de internações para a cidade de Monteiro ----------83

Gráfico 14 - Previsões de internações por técnica para a cidade de Monteiro --------84

Gráfico 15 - Análise de significância estatística para as previsões de internações para a cidade de Patos ----------85

Gráfico 16 - Previsões de internações por técnica para a cidade de Patos ----------86

Gráfico 17 - Análise de significância estatística para as previsões de internações para a cidade de Santa Rita ----------86

Gráfico 18 - Previsões de internações por técnica para a cidade de Santa Rita------87

Gráfico 19 - Análise de significância estatística para as previsões de notificações para a cidade de Bayeux ----------88

Gráfico 20 - Previsões de notificações por técnica para a cidade de Bayeux ---------88

Gráfico 21 - Análise de significância estatística para as previsões de notificações para a cidade de Cabedelo ----------89

Gráfico 22 - Previsões de notificações por técnica para a cidade de Cabedelo ------89

Gráfico 23 - Análise de significância estatística para as previsões de notificações para a cidade de Cajazeiras----------90

Gráfico 24 - Previsões de notificações por técnica para a cidade de Cajazeiras -----91

Gráfico 25 - Análise de significância estatística para as previsões de notificações para a cidade de Campina Grande ----------91

Gráfico 26 - Previsões de notificações por técnica para a cidade de Campina Grande ----------92

Gráfico 27 - Análise de significância estatística para as previsões de notificações para a cidade de Catolé do Rocha----------93

Gráfico 28 - Previsões de notificações por técnica para a cidade de Catolé do Rocha ----------94

Gráfico 29 - Análise de significância estatística para as previsões de notificações para a cidade de João Pessoa----------95

Gráfico 30 - Previsões de notificações por técnica para a cidade de João Pessoa --96

Gráfico 31 - Análise de significância estatística para as previsões de notificações para a cidade de Monteiro ----------97

Gráfico 32 - Previsões de notificações por técnica para a cidade de Monteiro--------97

Gráfico 33 - Análise de significância estatística para as previsões de notificações para a cidade de Patos ----------98

Gráfico 34 - Previsões de notificações por técnica para a cidade de Patos----------99

Gráfico 35 - Análise de significância estatística para as previsões de notificações para a cidade de Santa Rita ---------- 99

Gráfico 36 - Previsões de notificações por técnica para a cidade de Santa Rita --- 100

LISTA DE ABREVIATURAS E SIGLAS

AESA	Agência Executiva de Gestão das Águas do Estado da Paraíba
ARIMA	*Autoregressive Integrated Moving Average*
AUTOTIC-NN	*AUTOencoding based Time series Clustering with Near- est Neighbour*
BPNN	*BackPropagation Neural Network*
CEP	Comitês de Ética em Pesquisa
CNN	*Convolutional neural network*
CONEP	Comissão Nacional de Ética em Pesquisa
COVID-19	Doença coronavírus
CSV	*Comma-separated values*
DENV	Vírus da dengue
DL	*Deep Learning*
ENET	*Elastic Net Regression*
GAM	*Generalized 17all-ba model*
GB	*Gradient Boosting*
GBM	*Generalized Boosting models*
IPP	Índice de Infestação Predial
KNN	*K-nearest neighbors*
LIRAA	Levantamento Rápido de Índices para *Aedes aegypti*
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolute Error*
MAPE	*Mean absolute percentage error*
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
PNCD	Programa Nacional de Combate à Dengue
POLY	*Polynomial kernel*
RBF	*Radial basis function*
RELU	*Rectified Linear Units*
RF	*Random Forest*
RMSE	*Root Mean Absolute Error*
SIH	Sistema de Informações hospitalares
SINAN	Sistema de Informação de Agravos de Notificação
SNIS	Sistema Nacional de Informações sobre Saneamento
SUS	Sistema Único de Saúde
SVR	*Support Vector Regression*
TANH	Tangente hiperbólica
WHO	*World Health Organization*
XGBOOST	*Extreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO ---- 18
2 REVISÃO DA LITERATURA ---- 22
2.1 Aspectos epidemiológicos, socioambientais e dengue. ---- 22
2.2 Técnicas de *Machine Learning* e de *Deep Learning* ---- 26
2.2.1 Random Forest ---- 29
2.2.2 Support Vector Regression ---- 31
2.2.3 Multilayer Perceptron ---- 32
2.2.4 Long Short-Term Memory ---- 34
2.2.5 Convolutional Neural Network ---- 36
2.3 Demais técnicas de previsão ---- 38
2.3.1 Arima ---- 38
2.3.2 Naive Forecast ---- 38
2.4 Técnicas avaliativa das previsões ---- 38
2.5 Teste de significância dos resultados ---- 39
2.6 Trabalhos relacionados ---- 39
3 MATERIAL E MÉTODOS ---- 44
3.1 Delineamento da pesquisa ---- 44
3.2 Fonte de dados ---- 44
3.3 Aspectos éticos ---- 45
3.4 Recursos de *software* ---- 45
3.5 Sistema para predição de casos de internações e notificações ---- 46
3.5.1 Banco de dados ---- 46
3.5.2 Módulo de carga ---- 47
3.5.3 Módulo de Previsões ---- 48
3.5.3.1 Escolha dos cenários e atributos previsores ---- 50
3.5.3.2 Ajuste nos hiperparâmetros ---- 54
3.5.3.3 Geração das previsões ---- 56
3.5.4 Módulo avaliativo ---- 57
3.5.4.1 Escolha da melhor técnica ---- 57
3.5.4.2 Validação estatística ---- 57
3.5.4.3 Geração de gráficos para validação visual ---- 57

4 RESULTADOS E DISCUSSÃO ---------- 58
4.1 Determinação do melhor cenário e combinação de atributos previsores ---------- 58
4.1.1 Internações ---------- 58
4.1.2 Notificações ---------- 62
4.2 Ajustes nos hiperparâmetros dos algoritmos ---------- 66
4.2.1 Internações ---------- 66
4.2.2 Notificações ---------- 68
4.3 Geração das previsões e determinação da melhor técnica por cidade 71
4.3.1 Internações ---------- 71
4.3.2 Notificações ---------- 72
4.4 Avaliação dos resultados produzidos ---------- 73
4.4.1 Internações ---------- 74
4.4.2 Notificações ---------- 87
5 CONSIDERAÇÕES FINAIS ---------- 101
REFERÊNCIAS BIBLIOGRÁFICAS ---------- 103
APÊNDICE A – REVISÃO SISTEMÁTICA ---------- 108

1 INTRODUÇÃO

Originário do continente africano, o mosquito *Aedes aegypti* é capaz de transmitir algumas doenças para o homem, como, por exemplo, dengue, febre amarela, *zika* e *chikungunya*. A dengue é causada pelo vírus DENV e, atualmente, existem quatro sorologias desse vírus capazes de infectar o homem: DENV-1, DENV-2, DENV-3 e DENV-4. (FARES et al., 2015).

No Brasil, os primeiros registros de dengue datam do período colonial. Segundo historiadores, a entrada da doença aconteceu por meio do comércio de escravos (MAIA et al., 2019). Em 1955, após um trabalho sanitário dos órgãos governamentais e colaboração da população, o país conseguiu erradicar a dengue. Entretanto, ações similares não foram realizadas em países vizinhos e, na década de 80, ocorreram várias epidemias da doença no Brasil, com destaque para as ocorridas em Roraima e no Rio de Janeiro. Desde então, a doença se espalhou em todos os estados da federação e continua a crescer nesse país (ZARA et al., 2016).

O vetor da dengue encontra em países de clima tropical, como é o caso do Brasil, combinações climáticas ideais para a sua reprodução: elevado número de precipitações, de umidade e de temperatura (MARQUES-TOLEDO et al., 2017). Além disso, os problemas sociais e de saneamento potencializam a capacidade de reprodução do mosquito e a perpetuação da doença. O indevido descarte de resíduos sólidos combinando com uma precária coleta desses resíduos, em adição ao lançamento indevido de esgotos a céu aberto e a incorreta armazenagem de água, elevam a possibilidade de criadouros para o mosquito descartar seus ovos (SOUZA; ALBUQUERQUE, 2018).

Em 2019, a Organização Mundial de Saúde, em inglês (*WHO*), contabilizou cerca de 4.2 milhões de manifestações de dengue em todo o planeta. Anteriormente, a *WHO* chegou a emitir um alerta e classificou a dengue como uma das principais doenças para o ano de 2019. O alerta emitido pela *WHO*, posteriormente, ficou comprovado através de números. Segundo o SINAN, houve 1.556.595 casos prováveis de dengue no Brasil. Esse número representa um aumento de 484% em relação às observações do ano de 2018. Na Paraíba, os números também seguiram a tendência de crescimento e tiveram um aumento de 72% (BRASIL, 2019; DE JESUS et al., 2020).

O Ministério da Saúde, em 2002, criou o Programa Nacional de Combate à Dengue (PNCD) com intuito de padronizar quais ações devem ser tomadas em todo o território brasileiro no combate à dengue. Campanhas de conscientização para o correto descarte de lixo, indicações de ações para o adequado armazenamento de água, ações de enfrentamento por meio de carros do tipo fumacê e visita de agentes sanitários às residências onde há possíveis focos de dengue, são exemplos de atos preconizados pelo programa. Ademais, o melhoramento na coleta de resíduos e a ampliação da rede de saneamento também são atitudes tomadas visando combater o vírus da dengue (ROSA; BRAIDO; CAPORLINGUA, 2020).

Além de causar problemas clínicos aos pacientes, a dengue também causa impactos econômicos e sociais para o país (LEITE, 2015). O estudo liderado por Teich, Arinelli e Fahham (2017), elencou os custos realizados no combate ao vetor, os custos médicos diretos e os custos indiretos causados pela dengue. Segundo os autores, em 2016, foram gastos cerca de R$ 1.470.990.760 no combate ao vetor. Em relação aos custos médicos, R$ 175.876.163 foi o valor custeado pelo Governo Federal para tratar os doentes acometidos por dengue. Por fim, a doença gerou um custo indireto de R$ 293.341.383. Ainda do estudo, para a Paraíba, os custos relacionados ao combate ao vetor, os custos médicos diretos e os custos indiretos foram de R$ 13.504.533, R$ 4.289.618 e R$ 7.187.529, respectivamente.

Conforme relatam os estudos de Souza e Albuquerque (2018), Barbosa et al. (2020) e Ribeiro et al. (2021), apesar dos investimentos e dos esforços, o Brasil ainda passa por surtos de epidemias e de mortes causadas por dengue. Assim, é necessário intensificar as ações atuais e buscar ações complementares e inovadoras no combate à doença. As informações são corroboradas ao analisar o aumento dos casos no Brasil e na Paraíba. Na Paraíba, o sétimo boletim epidemiológico de arboviroses da Paraíba aponta um crescimento de 53% de casos de dengue em 2021 em comparação com os dados do ano passado (PARAÍBA, 2021).

O uso da tecnologia da informação na saúde está cada vez mais constante. De acordo com PINOCHET (2011), os sistemas de informação vêm sendo utilizados no apoio à saúde, na prevenção de doenças, promoções de ações de saúde, no controle de doenças, mas também na vigilância e monitoramento de doenças. Para Longaray e Castelli (2020), a Tecnologia da Informação se tornou parte integral para todas as atividades relacionadas à prestação dos serviços de saúde.

Nesse contexto, técnicas de *Machine Learning* (ML) e de *Deep Learning* (DL) vêm sendo utilizadas com sucesso na tarefa de predição de casos de dengue e de internações causadas pela doença, como, por exemplo, nos trabalhos de Carvajal et al. (2018), de Doni e Sasipraba (2020), de Sippyd et al. (2020) e de Xu et al. (2020). Por meio dos resultados produzidos, os pesquisadores fornecem aos governantes e a população em geral informações sobre possíveis surtos de dengue, ajudam o combate com um controle estratégico da doença, contribuem para o uso racional de recursos humanos e financeiros e, por conseguinte, estão salvando mais vidas.

Diante desse cenário desafiador, o objetivo principal do trabalho é, por meio de técnicas de *Machine Learning (Random Forest* e *Support Vector Regression)* e de *Deep Learning (Multilayer Perceptron, Long Short-Term Memory e Convolutional Neural Network)* e utilizando dados epidemiológicos, climáticos e sanitários, criar um sistema capaz de gerar previsões de notificações e de internações causadas por dengue para os municípios da Paraíba.

Os objetivos específicos do trabalho são: 1 – Coletar os dados epidemiológicos, climáticos e sanitários para todos os 223 municípios da Paraíba; 2 – Criar e povoar um banco de dados, relacional, com os dados coletados; 3 – Criar e treinar modelos para encontrar a melhor combinação de atributos previsores para realizar a previsão de casos de cada cidade; 4 – Encontrar a melhor combinação de parâmetros para as técnicas de ML e de DL utilizadas no trabalho; 5 – Gerar as previsões de internações e casos de notificação de dengue; 6 – Determinar qual técnica gera as melhores previsões por município; 7 – Validar e evidenciar, estatisticamente, se há diferença entre as previsões geradas pelas técnicas.

Este trabalho foi estruturado em cinco capítulos, a saber: 1 – Introdução contendo a contextualização, problematização e justificativa da pesquisa; 2 – Revisão da literatura versando sobre dengue, apresentando as técnicas de *Machine Learning* e *Deep Learning*, as técnicas de previsões *Naive Forecast* e ARIMA, as técnicas avaliativas para modelos de previsões *Root Mean Square error* e *Mean Absolute error*, as técnicas estatísticas para avaliar a significância dos resultados (ANOVA e TUKEY) e, os trabalhos relacionados; 3 – Material e Métodos expondo o delineamento da pesquisa, as fontes de dados, os aspectos éticos, os recursos de *software* e o Sistema para predição de casos de internações e notificações de dengue; 4 – Resultados e discussão contendo os melhores atributos previsores, escolha dos hiperparâmetros

paras as técnicas de ML e DL, indicação da melhor técnica e a taxa de erro para cada cidade e as validações dos resultados produzidos; 5 – Considerações finais mostrando as conclusões da pesquisa, as limitações do trabalho, bem como sugestões para trabalhos futuros.

2 REVISÃO DA LITERATURA

2.1 Aspectos epidemiológicos, socioambientais e dengue.

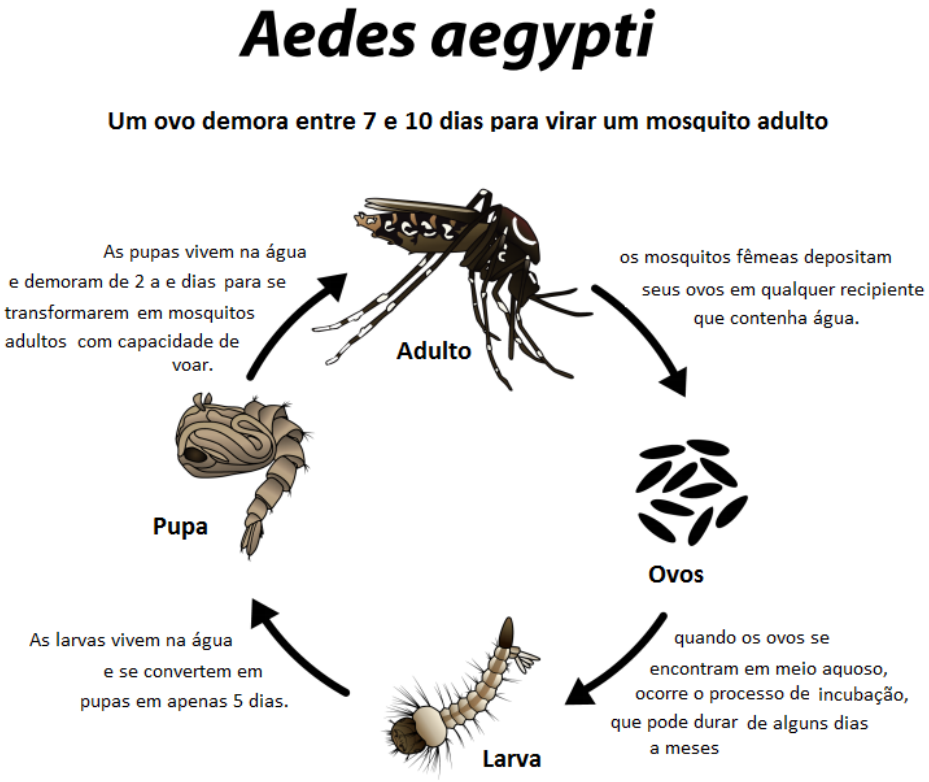
A dengue é uma arbovirose causada pelo vírus DENV e transmitida para os humanos através do mosquito *Aedes aegypti*. Pesquisadores relatam que o mosquito teve origem no Egito, África, e vem se espalhando para países tropicais e subtropicais dos demais continentes desde o século 16 (IOC/FIOCRUZ, 2011).

Para Zara et al. (2016), a dengue foi introduzida no Brasil durante o período colonial e o comércio de escravos teria sido a principal forma de entrada da doença no país. Após um hercúleo esforço, o Brasil conseguiu erradicar a doença do seu território em 1955. Contudo, não foram realizados trabalhos similares nos países vizinhos, principalmente, nos países da América do Sul e Central. Com isso, por meio de deslocamento de pessoas infectadas oriundas desses países e com o abandono das medidas preventivas no Brasil, nos anos 70, aconteceu a reentrada da doença (CÂMARA et al., 2007; FARES et al., 2015).

Os primeiros relatos laboratoriais e clínicos da dengue foram registrados pelo ministério da saúde no ano 1981, em Boa Vista, estado de Roraima. Após isso, em 1896, aconteceram epidemias no Rio de Janeiro e no Nordeste. Desde então, a dengue vem se espalhando e se tornou um dos maiores desafios na saúde pública do Brasil (STOLERMAN; MAIA; NATHAN KUTZ, 2019; VIANA; IGNOTTI, 2013).

O ciclo de vida do mosquito inicia com a fêmea fecundada do *Aedes aegypti* depositando os seus ovos em recipientes contendo água. Ao serem depositados, os ovos possuem capacidade de sobrevivência de até um ano e, ao encontrar em contato com água e em condições ambientais favoráveis, por volta de dois dias os ovos continuam seu ciclo e viram larvas. As larvas evoluem no ambiente aquático por cinco dias até se tornarem pupas. A fase final do ciclo se dá quando as pupas são maturadas e, em média de dois dias, viram um mosquito adulto (PARREIRA; ATOUGUIA; SOUSA, 2013). A Figura 1 apresenta o ciclo de vida do mosquito.

Figura 1 – Ciclo de vida do mosquito *Aedes aegypti*



Fonte: Ioc/Fiocruz (2019)

Em relação ao ciclo de transmissão da doença, ao picar um ser humano infectado pelo vírus DENV, haverá multiplicação desse vírus no mosquito. Uma vez infectado, o *Aedes aegypti* passará a ser um vetor transmissor da doença e será capaz de transmiti-la enquanto viver. Adicionalmente, alguns ovos do mosquito já podem carregar o vírus (PARREIRA; ATOUGUIA; SOUSA, 2013).

Ao entrar em contato com o corpo humano saudável, o DENV pode ficar incubado de 2 a 10 dias. Após esse tempo, iniciam os sintomas da doença: febre, mal-estar, dores de cabeça, dores musculares e falta de apetite. Nos casos mais graves da doença, o paciente pode necessitar de internação hospitalar e, inclusive, ser acometido de hemorragia. Nos piores casos leva o paciente a óbito (GRACIANO et al., 2017). Além disso, uma vez infectado por alguma das variantes, o ser humano adquire imunidade para esse tipo, contudo, continua suscetível às demais variantes (SWAMINATHAN; KHANNA, 2019).

Embora não seja uma doença nova, ainda não existe um medicamento adequado para combater o vírus. As ações de tratamento contra a dengue clássica e hemorrágica são focadas em amenizar os sintomas e os problemas causados pela doença (PARREIRA; ATOUGUIA; SOUSA, 2013).

Geralmente, as vacinas são utilizadas para combater viroses. Entretanto, para a dengue, não existe uma vacina sem restrição de uso capaz de combater todos os tipos de sorologia do DENV (SILVEIRA; TURA; SANTOS, 2019). No Brasil, existe a vacina Dengvaxia, regulamentada pela ANVISA contra a dengue. Todavia, ela é recomendada, exclusivamente, para pessoas que já foram acometidas por, pelo menos, alguma das variações do DENV. Caso seja aplicada na população sem contato anterior com o vírus, a vacina poderá potencializar a doença e gerar sérias complicações para o paciente (SILVEIRA; TURA; SANTOS, 2019).

Destarte, a principal arma contra a doença é combater a proliferação do mosquito. Os governos federais, estaduais e municipais investem em campanhas para conscientizar a população para o correto armazenamento de água, descarte de lixo ou de objetos que possam a vir se tornar um *habitat* do mosquito. A utilização de carros do tipo fumacê e a intervenção dos agentes de saúde por meio da utilização de produtos químicos em possíveis criadouros também são formas encontradas para combater a doença (NORRBY, 2014; SOUZA; ALBUQUERQUE, 2018).

No Brasil, atualmente, existem quatro tipos de sorologia do vírus DENV (DENV-1, DENV-2, DENV-3 e DENV-4) em circulação. O clima tropical do país possibilita um ambiente ideal para a proliferação do mosquito (PHAM et al., 2018). Adicionalmente, segundo Carvalho e Moreira (2017), problemas sociais (falta de moradia adequada, crescimento descontrolado de cidades) e sanitários (problemas de abastecimento, descarte de esgotos a céu aberto e problemas na de coleta de lixo) potencializam o problema da dengue nesse país.

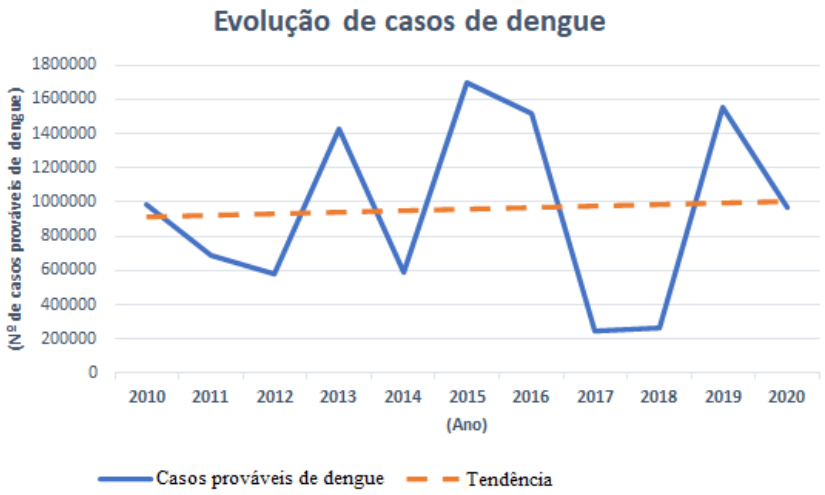
Desde a última erradicação da dengue no Brasil, o cenário socioambiental e o cenário climático sofreram alterações significativas e, consequentemente, dificultam o combate à dengue (MENDONÇA; SOUZA; DUTRA, 2009). Cada vez mais a população brasileira está concentrada em centros urbanos, onde o número de habitações com problemas de saneamento e de coleta de lixo aumenta a cada dia. Além do mais, há um notório incremento da produção e descarte incorreto de lixo não

orgânico, que não será decomposto pela natureza e, consequentemente, será potencial *habitat* do mosquito transmissor da dengue (RIBEIRO et al., 2021; SOBRAL; SOBRAL, 2019).

Outro complicador no combate da doença é a adaptabilidade do vetor. Anteriormente, era sabido que o *Aedes aegypti* utilizava locais com água parada e limpa durante o seu ciclo de reprodução. Contudo, pesquisadores demonstram a aptidão reprodutiva do mosquito em águas sujas e, até mesmo, em esgotos (ALMEIDA; COTA; RODRIGUES, 2020; BESERRA et al., 2009; CARVALHO; MOREIRA, 2017).

A dificuldade na luta contra a dengue, e, consequentemente, manutenção e crescimento da doença é demonstrado através dos números apresentados na Figura 2. Como pode ser observado, os casos prováveis de dengue caíram em 2017 e em 2018. Contudo, voltaram a aumentar em 2019 e estão em tendência de crescimento, ainda que, devido à COVID-19, trabalhos relatam indícios de subnotificação dos casos de dengue em 2020 (LEANDRO et al., 2020; NASCIMENTO et al., 2021).

Figura 2 – Número de casos prováveis de dengue no Brasil entre os anos de 2010 e 2020

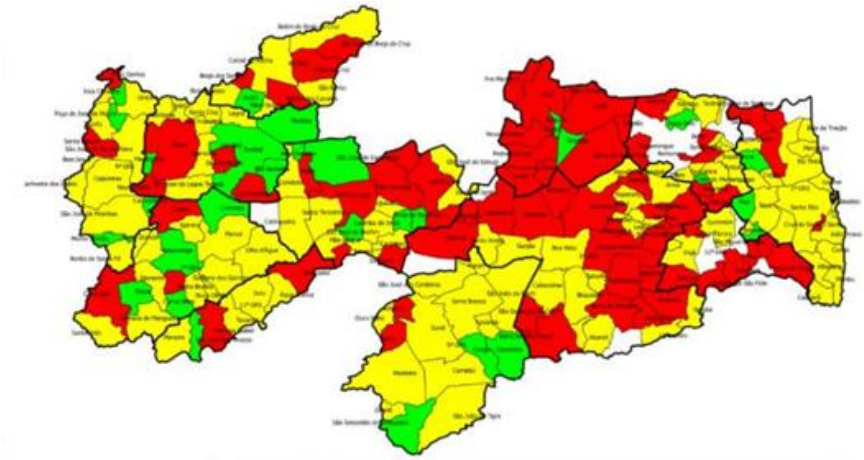


Fonte: Adaptado do SINAN

Na Paraíba, o cenário de manutenibilidade da doença também é demonstrado. Segundo o sétimo boletim epidemiológico de arboviroses da Paraíba, houve aumento de 53% de casos de dengue em relação ao ano de 2020. Além do mais, o Levantamento Rápido de Índices para *Aedes aegypti* – LIRAa aponta que 34,92% dos

municípios desse estado estão em situação de risco para surto de arboviroses, 51,19% estão em situação de alerta, 13,87% estão em situação satisfatória e, finalmente, 4,78% indicaram IPP zero (PARAÍBA, 2021). O mapa da Paraíba com as classificações está representado na Figura 3.

Figura 3 – Mapa da Paraíba contendo o levantamento LIRAa. A cor verde indica baixo risco (<1%), a cor amarela risco moderado (>1% e <3,9%) e a vermelha alto risco (>= 4%).



Fonte: PARAÍBA (2021)

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

Machine Learning é uma subárea da disciplina Inteligência Artificial. Diferentemente de sistemas ditos comuns, onde o programador deve, antecipadamente, definir e programar todas as regras e comportamentos do *software*, os sistemas de *Machine Learning* são capazes de aprender com os dados (GÉRON, 2019).

De acordo com Deitel e Deitel (2019), sistemas de *Machine Learning* são empregados em cenários em que as soluções de *softwares* tradicionais não conseguem resolver os problemas ou as suas soluções não são satisfatórias. Ainda em sua obra, os autores elencam as seguintes situações como mais populares durante o uso de ML: detecção de anomalias, problemas de classificação e de visão computacional, detecção de fraudes bancárias, exploração de dados, detecção de objetos em cena, sistemas de recomendações, processamento natural de linguagem, análise de sentimento, previsões de séries temporais, entre outros.

Conforme listado por Deitel e Deitel (2019), a aplicabilidade de ML é vasta e existem diversos tipos de sistemas de *Machine Learning*. Os sistemas de aprendizado supervisionado e sistemas de aprendizado não supervisionado são os principais. Para Russell, Stuart e Norvig (2013), a aprendizagem não supervisionada é caracterizada por aprender padrões com base na entrada (dados de treinamento) e não há um *feedback* explícito quanto a esse aprendizado. A atividade de agrupamento é a mais comum para esse tipo. Em contrapartida, no aprendizado supervisionado, são fornecidos dados de treinamento com as entradas e as saídas desejadas.

Ainda sobre aprendizado supervisionado, os problemas podem ser subdivididos em dois grupos de acordo com a saída produzida. Quando a saída do algoritmo de aprendizado supervisionado for um conjunto finito de valores, esse será chamado de problema de classificação. Caso o resultado de saída for um valor numérico, ele será chamado de problema de regressão. Os algoritmos mais importantes do aprendizado supervisionado são: regressão linear, regressão logística, máquina de vetores de suporte, árvores de decisão, florestas aleatórias e, finalmente, redes neurais (GÉRON, 2019; RUSSELL, STUART; NORVIG, 2013).

A Figura 4 ilustra o funcionamento de um filtro de *spam*. Com base em um conjunto de treinamento, o filtro deverá indicar se um novo e-mail pode ser um possível *spam*. Esse é um típico problema de aprendizado supervisionado de classificação, pois os rótulos são fornecidos e o algoritmo produzirá um resultado finito: marcar ou não o e-mail como *spam*.

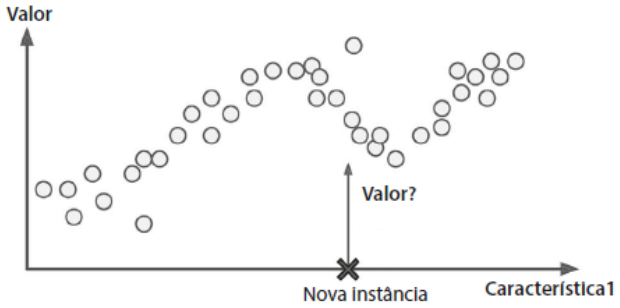
Figura 4 – Funcionamento de um problema de classificação através de um filtro de spam

Fonte: Géron (2019)

Os problemas de regressão consistem em prever um valor quantitativo por meio de atributos chamados previsores (IZBICKI, RAFAEL; SANTOS, 2020). São exemplos

de problemas de regressão: previsão da temperatura, previsão de casos de uma certa doença, previsão de valores de ações da bolsa de valores. A Figura 5 demonstra a problemática de prever o valor de um carro de acordo com os previsores quilometragem do veículo, idade e marca.

Figura 5 – Ilustração de um problema de regressão

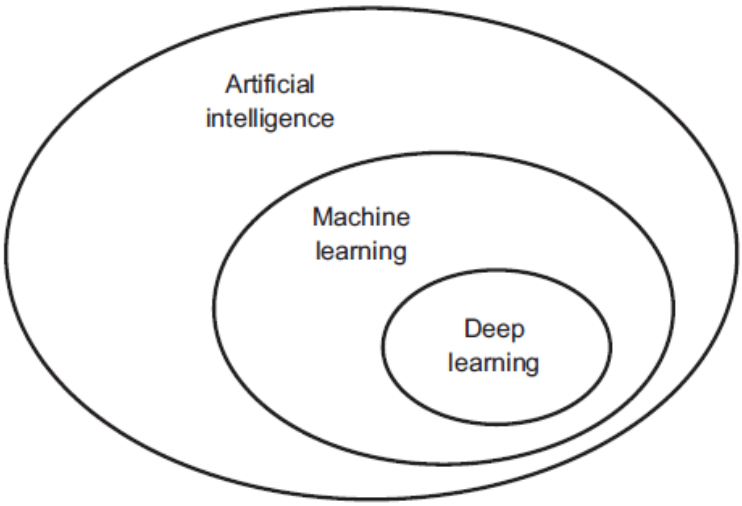


Fonte: Géron (2019)

Com o avançar da tecnologia, cada vez mais os problemas computacionais estão complexos, produzindo e necessitando de mais dados para o funcionamento. Além disso, houve um considerável aumento da capacidade de processamento e redução dos componentes dos computadores. Nesse contexto, os pesquisadores começaram a treinar modelos com mais dados, empilhar mais camadas sucessivas e a *Deep Learning* se popularizou (BONACCORSO, 2018; GOODFELLOW, IAN; BENGIO, YOSHUA; COURVILLER, 2016).

Chollet (2017), descreve *Deep Learning* como um subcampo da *Machine Learning* que está focando em novas formas de representação do aprendizado com base no uso de sucessivas e representativas camadas. A Figura 6 descreve os conceitos abordados por Chollet.

Figura 6 – Relação entre Inteligência Artificial, *Machine Learning* e *Deep Learning*



Fonte: Chollet (2017)

A *Deep Learning* vem sendo aplicada em problemas de classificação de imagens, condução de carros autônomos, rastreamento visual em tempo real, bioinformática. Embora boa parte desses problemas possa ser resolvido por técnicas clássicas ou de ML, na sua grande maioria, as técnicas de DL produzem resultados melhores (BONACCORSO, 2018).

Para Chollet (2017), o sucesso da *Deep Learning* está na habilidade de automatizar a engenharia de recursos. Isso é possível por meio da forma incremental, camada a camada, em que são desenvolvidas representações mais complexas e, ao mesmo tempo, aprendidas em conjunto.

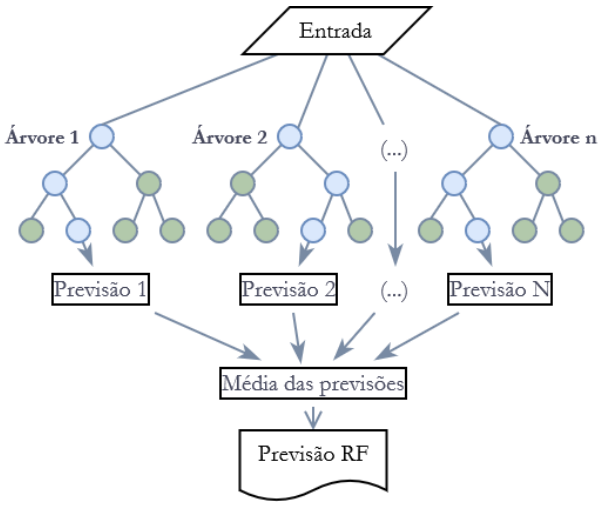
2.2.1 Random Forest

Random Forest é uma técnica de *Machine Learning* que, através da construção e treinamentos de árvores de decisão, é capaz de solucionar problemas de classificação e regressão (GÉRON, 2019; RUSSEL, STUART; NORVIG, 2013). Para Harisson (2020), o algoritmo *Random Forest* é eficiente na sua execução e, ao utilizar o artifício de adicionar mais árvores, diminui o problema de *overfitting* (super adequação) existente na estratégia de utilizar árvores de decisão individuais.

A Figura 7 detalha como é feita a previsão de valores após o treinamento das n árvores envolvidas no algoritmo de *Random Forest*. Ao submeter um valor de teste,

cada árvore irá produzir a sua previsão. Por fim, o algoritmo calcula a média das previsões e fornece o valor final.

Figura 7 – Estrutura de funcionamento de uma regressão feita por *Random Forest*



Fonte: Adaptado de Chollet (2017)

O algoritmo de regressão realizado por Random Forest é disponibilizado em Python através da classe RandomForestRegressor, presente na biblioteca Scikit-learn (PEDREGOSA et al., 2011). O Quadro 1 apresenta os principais parâmetros utilizados pelo algoritmo, a sua descrição e os valores padrões definidos pela biblioteca.

Quadro 1 – Descrição dos parâmetros utilizados pelo algoritmo *Random Forest*

Parâmetro	Descrição	Valor padrão
n_estimators	Número de árvores na floresta	100
max_depth	A profundidade máxima de uma árvore	*None*
min_samples_split	Número mínimo de amostras para dividir um nó.	2
min_samples_leaf	Número mínimo de registros presentes em um nó folha.	1
max_features	Número máximo de atributos para analisar a separação de nós.	*"auto"*
max_leaf_nodes	Número máximo de nós folhas.	*None*
bootstrap	Indicador se o *bootstrap* será utilizado para a criação das árvores	*True*

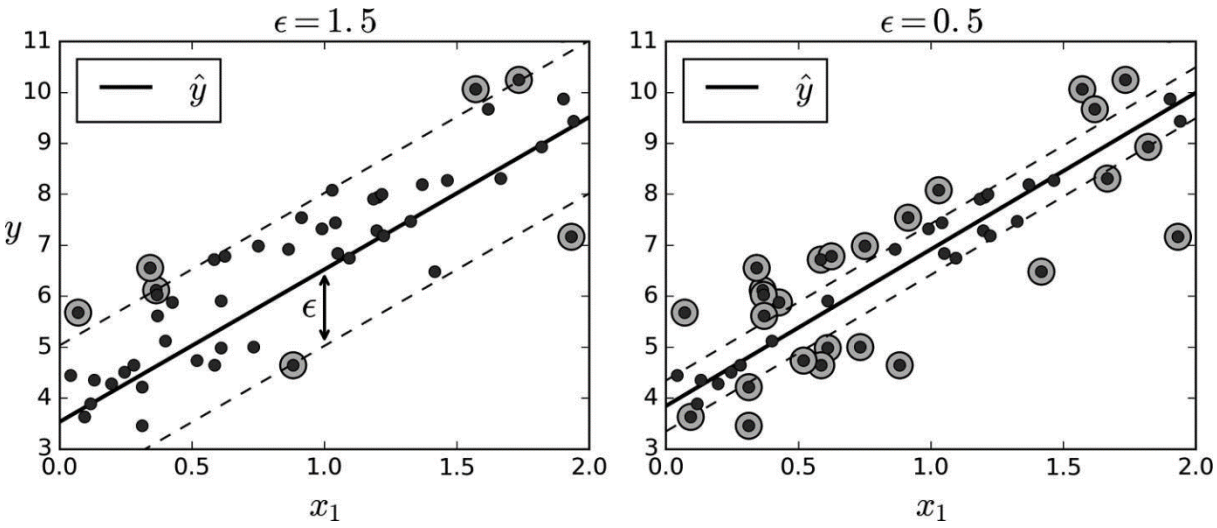
Fonte: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

2.2.2 Support Vector Regression

A técnica de *Machine Learning Support Vector Regression* consiste em criar uma função capaz de encontrar um hiperplano que abarca o número máximo de registros do conjunto de treinamento, respeitando as violações de margem (AWAD; KHANNA, 2015). A largura do hiperplano é controlada através de um parâmetro nomeado ε. A partir dos vetores de suportes e limites impostos por ε, a largura do hiperplano é determinada.

As representações de dois hiperplanos para um problema linear são demonstradas através da Figura 8. Contudo, conforme descrito por Awad e Khana (2015), a maioria dos problemas de regressão são de natureza não linear. Para a resolução desses, é utilizado o truque do *kernel.* Os *kernels* mais comuns são: linear, polinomial, RBF gaussiano e *sigmoide* (RUSSELL, STUART; NORVIG, 2013).

Figura 8 – Representação do funcionamento da regressão através de SVR



Fonte: Géron (2019)

A classe SVR, da biblioteca *Scikit-learn*, disponibiliza a implementação do algoritmo SVR. Os principais parâmetros utilizados pelo algoritmo são detalhados e explicados no Quadro 2.

Quadro 2 – Descrição dos parâmetros utilizados pela classe SVR

Parâmetro	**Descrição**	**Valor padrão**
kernel	Qual *kernel* será utilizado no algoritmo	*"rbf"*
degree	Grau da função polinomial (exclusivo para o *kernel* poli)	3
gamma	Coeficiente dos *kernels* 'rbf', 'poly' e 'sigmóide'	*"scale"*
C	Parâmetro de penalização para regularização.	*1.0*
epsilon	Parâmetro 32all-ba indicando o limite para penalização	*0.1*

Fonte: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

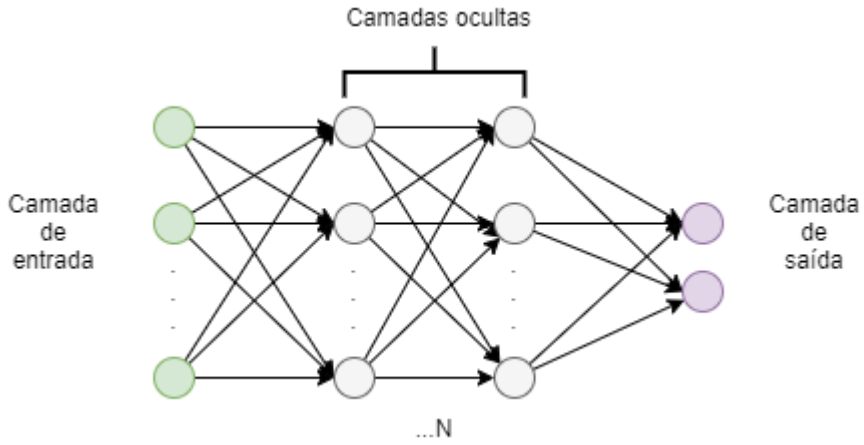
2.2.3 Multilayer Perceptron

Para Goodfellow, Bengio e Courviller (2016), a *Multilayer Perceptron* é uma rede neural, *feedforward,* composta por mais de um *perceptron.* A *Multilayer*

Perceptron (MLP) possui uma camada de entrada, uma camada de saída e uma ou mais camadas ocultas. Elas são classificadas como *feedforward*, pois a informação flui da camada de entrada em uma única direção: a camada de saída. O treinamento da MLP se dá através da técnica de aprendizado supervisionado *backpropagation,* onde os pesos das camadas ocultas vão sendo ajustados de modo que o erro produzido pela camada de saída seja reduzido (AWAD; KHANNA, 2015).

De acordo com Bonaccorso (2018) e Goodfellow, Bengio e Courviller (2016), a MLP é uma das principais e, talvez, a técnica que mais representa o conceito de aprendizado profundo. Isso ocorre devido ao empilhamento de camadas ocultas. Todos os neurônios de uma MLP estão totalmente conectados com os neurônios da camada seguinte com exceção da camada de saída. A Figura 9 representa a generalização da arquitetura de uma rede MLP.

Figura 9 – Arquitetura de rede neural *Multilayer Perceptron* (MLP)



Fonte: Adaptado de Géron (2019)

A *Scikit-learn* provê a implementação de regressão através de MLP por meio da classe MLPRegressor. Os principais parâmetros do algoritmo estão demonstrados no Quadro 3.

Quadro 3 – Parâmetros utilizados pelo algoritmo de MLP

Parâmetro	**Descrição**	**Valor padrão**
hidden_layer_sizes	Tupla que representa o número de camadas ocultas bem como o número de neurônios em cada uma delas.	(100,)
activation	Função de ativação das camadas ocultas. Valores possíveis: “identity”, “logistic”, “tanh” e “relu”	*“relu”*
Solver	Solucionador para a otimização dos pesos. Valores possíveis: *‘lbfgs’* é um otimizador da família quase-Newton, *‘sgd’* refere-se a uma descida do gradiente estocástico e ‘adam’ refere-se a um otimizador baseado em gradiente estocástico.	*“adam”*
batch_size	Tamanho do minilote. Quando não especificado, será o valor mínimo entre 200 e o número de entradas.	*“auto”*
learning_rate	Indicador da atualização da taxa de aprendizado. Valores possíveis: “*constant*”, “*adaptive*”	*“constant”*
early-stopping	*Flag* responsável por habilitar a parada antecipada do treinamento.	*“False”*

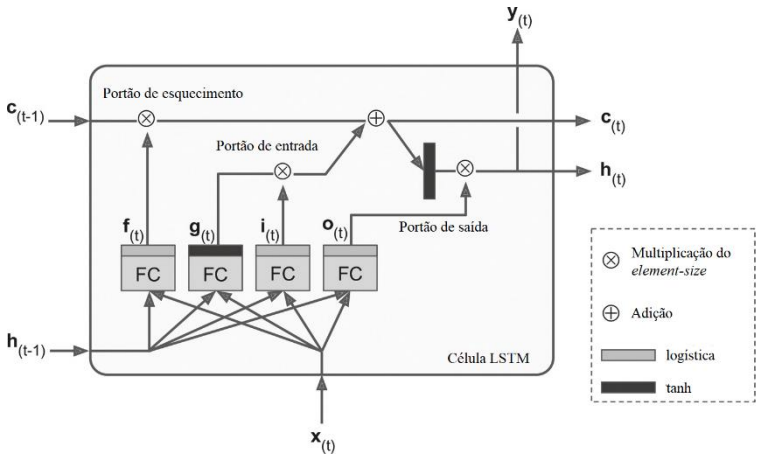
Fonte: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

2.2.4 Long Short-Term Memory

Long Short-Term Memory é um tipo de rede neural recorrente pertencente à abordagem de *Deep Learning.* Através das células LSTM, essa técnica é capaz de tratar problemas de dependência de dados a longo prazo (OZER et al., 2021).

A célula LSTM mantém um vetor com os dados da memória curta e outro vetor com as memórias antigas. Os portões de esquecimento, de entrada e de saída, representados na Figura 10, decidirão quais informações de longo prazo serão apagadas, quais informações de entrada serão incluídas no vetor de longo prazo e, finalmente, o portão de saída define quais informações do vetor de longo prazo serão utilizados para, junto com a memória de curto prazo, produzir a saída da célula (GÉRON, 2019).

Figura 10 – Funcionamento da célula LSTM



Fonte: Adaptado de Géron (2019)

A biblioteca *Keras* permite a utilização das soluções fornecidas pela plataforma de *Deep Learning* do Google, *TensorFlow*. Para criar uma rede recorrente com camadas LSTM através do *Keras* são utilizadas as classes *Sequential* e *LSTM*. Caso queira adicionar a parada antecipada (com intuito de evitar ou diminuir o overfitting), a classe *EarlyStopping* deve ser adicionada ao parâmetro *callback* da classe *Sequential*. O Quadro 4 lista e explica a utilização dos principais parâmetros utilizados por classe durante a criação de uma rede neural recorrente como camada LSTM e parada antecipada.

Quadro 4 – Listagem dos parâmetros utilizados pelo *Keras* para a criação de rede neural recorrente com LSTM

Classe	Parâmetro	Descrição	Valor padrão
LSTM	***batch_size***	Tamanho do minilote usado durante o treinamento.	-
	epochs	Número de épocas para treinamento do modelo.	-
	callbacks	Lista contendo quais *call-backs* serão aplicados no treinamento.	-
	units	Número de células LSTM presentes nas camadas ocultas.	-
Sequential	***activation***	Função de ativação.	*"tanh"*
	recurrent_activation	Função de ativação do passo recorrente.	*"sigmoid"*
	dropout	Taxa de exclusão de conexões das células durante o treinamento de uma rede.	*0.0*
EarlyStopping	***monitor***	Qual atributo será monitorado para determinar a parada.	*"val_loss"*
	patience	Limite de épocas sem melhorar o treinamento.	-

Fonte: https://keras.io/api/layers/recurrent_layers/lstm/

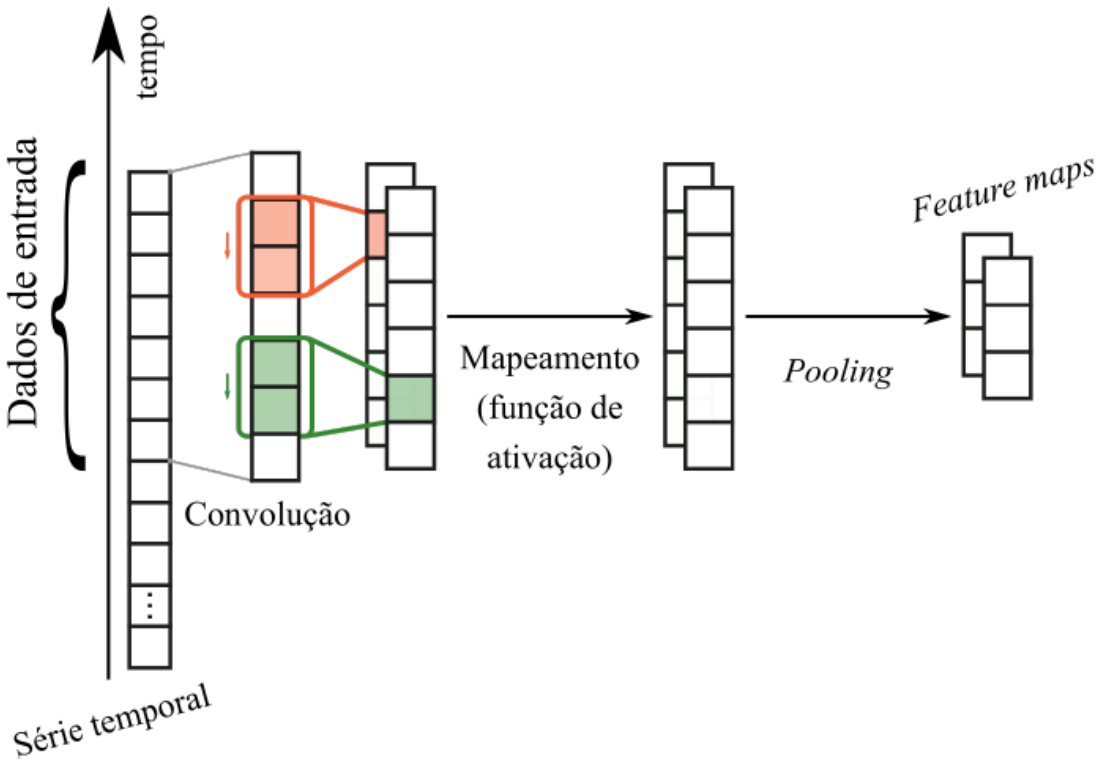
2.2.5 Convolutional Neural Network

A rede neural convolucional (CNN) é uma técnica de *Deep Learning* bastante utilizada na classificação e reconhecimento de padrões em imagens e vídeos. A ideia principal da CNN é, por meio de camadas de convolução e de *polling,* extrair as

características principais das imagens e reduzir o seu tamanho (DEITEL; DEITEL, 2019).

De acordo com (BARINO; BESSA, 2020), a utilização de redes CNNs pode ser adaptada a problemas de séries temporais ao utilizar o processo de convolução 1D. Ao aplicar tal técnica, podem ser descobertos importantes padrões na série. A Figura 11 traz a representação da aplicação da convolução 1D para uma série temporal.

Figura 11 – Funcionamento da adaptação de uma rede convolucional para o processamento de série temporal



Fonte: Barino e Bessa (2019)

Após a extração das informações mais importantes, os dados são submetidos a algum tipo de rede neural para que ocorra a precisão. Aqui, foi escolhida a rede LSTM demonstrada no tópico 2.2.4. A implementação de uma camada CNN com convolução de uma dimensão é feita no *Keras* através das classes *Conv1D*, *MaxPooling1D* e *Flatten*. Os principais atributos da classe Conv1D estão listados no Quadro 5. Os atributos de *MaxPooling1D* e *Flatten* não foram listados, pois não agregam relevância ao estudo.

Quadro 5 – Parâmetros utilizados pela classe Conv1D do *Keras*

Parâmetro	**Descrição**	**Valor padrão**
filters	Número de filtros utilizados na convolução.	-
kernel_size	Tamanho da janela de convolução.	-
activation	Função de ativação que será utilizada na convolução.	-

Fonte: https://keras.io/api/layers/convolution_layers/convolution1d/

2.3 Demais técnicas de previsão

2.3.1 Arima

Autoregressive Integrated Moving Average (ARIMA) é um modelo estatístico de previsão elaborado por Box e Jenkins em 1970. Muito utilizado em problemas de séries temporais, as previsões são realizadas por meio da combinação de modelo de autorregressão (AR), diferenciação (I) e modelo de média móvel (MA) (DUARTE; FAERMAN, 2019).

No P*ython*, o modelo ARIMA está disponível através da classe ARIMA, fornecido pela biblioteca *StatsModels*.

2.3.2 Naive Forecast

Naive forecast ou previsão ingênua, é um dos tipos de previsão mais simples. Para esse método, a previsão do próximo valor será exatamente igual ao valor atual (HYNDMAN; ATHANASOPOULOS, 2018). Embora não produza resultados tão efetivos, a técnica *Naive Forecast* é uma ótima abordagem para servir como *benchmark* durante a comparação de técnicas de previsão (SHMUELI; LICHTENDAHL JR, 2016).

2.4 Técnicas avaliativa das previsões

Conforme descrito por Kuhn e Johnson (2019), ao criar sistemas de previsão, o objetivo é gerar saídas próximas ao valor real, com menor taxa de erro, visto que, a probabilidade de prever valores iguais ao observado tende a 0.

As técnicas *Root Mean Absolute Error* (RMSE) e *Mean Absolute Error* (MAE) são as principais formas avaliativas dos modelos com *Machine Learning* e com *Deep Learning* (CARVAJAL et al., 2018).

O RMSE é calculado através da raiz quadrada da média da diferença entre o valor previsto (p) e o observado (0) elevada ao quadro. A equação 1 demonstra como é feito o cálculo do RMSE.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(p_t - o_t)^2}{T}} \quad (1.0)$$

A equação 2 demonstra como é calculado o erro médio absoluto (MAE).

$$MAE = \frac{\sum_{t=1}^{T}|p_t - o_t|}{T} \quad (2.0)$$

Ainda do trabalho Carvajal et al. (2018), o autor afirma que o RMSE é a melhor técnica para verificar a performance de modelos de predição, visto que, há maior penalização para discrepâncias maiores entre o valor previsto e observado.

2.5 Teste de significância dos resultados

Para verificar se há diferença estatística entre os resultados produzidos por modelos de regressão ou classificação de ML e quais resultados são diferentes, Ren et al. (2014) sugere realizar os testes estatísticos ANOVA e Tukey. Por meio da análise de variância, o ANOVA indica apenas se há ou não diferença entre as amostras. Contudo, para verificar quais amostras são diferentes, é necessário realizar, complementarmente, o teste de Tukey (KANJI, 2006).

2.6 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

(MUSSUMECI; CODEÇO COELHO, 2020). Devido à complexidade mencionada, inúmeros trabalhos estão utilizando ML e DL durante a predição de doenças.

Técnicas de *Machine Learning e de Deep Learning* vêm sendo aplicadas com sucesso na previsão de internações e de notificações de casos da doença dengue. Por meio desses resultados, os governantes e a população em geral estão podendo organizar campanhas de prevenção, direcionar melhor os esforços para combater a doença, dimensionar melhor leitos, recursos hospitalares e, portanto, salvar vidas (CHEN et al., 2018; GUO et al., 2017; XU et al., 2020).

SIPPYID et al. (2020), utilizando técnicas de *Machine Learning*, conduziram uma pesquisa capaz de prever casos de internações causadas por dengue. Para realizar as previsões, os autores utilizaram dados fornecidos pelo Ministério do Equador entre os anos de 2013 e 2017. Ainda sobre os dados, os autores criaram duas abordagens de utilização e as chamaram de S*everity Index for Suspected Arbovirus* (SISA) e *Suspected Arbovirus with Laboratory* (SISAL). Para o SISA, foram utilizados dados demográficos e sintomas dos pacientes. Em relação ao SISAL, além dos dados demográficos e sintomáticos, foram utilizadas informações laboratoriais como exames de sangue.

Para cada abordagem, os dados foram submetidos a treinamento e previsões foram geradas utilizando as técnicas: *K-nearest neighbors regression*, *Random forest*, *Elastic net regression*, *Generalized boosting models*, *Neural networks* e *Logistic regression*. *Generalized Boosting models (GBM)* obteve uma acurácia de precisão de 0,91 e foi a melhor técnica para SISA. Por fim, *Elastic Net Regression (ENET)* produziu previsões com 0,94 de acurácia e obteve o melhor resultado na abordagem SISAL.

O trabalho proposto por Silva (2019), objetivou prever casos de internações nos estados do Brasil por meio de análise de séries temporais. Em seu trabalho, foram utilizados dados de internações fornecidos pelo SINAN entre os anos de 2016 a 2019. As previsões foram feitas utilizando as técnicas de Suavização Exponencial, ARIMA, Redes Neurais Artificiais, Modelo por previsão de decomposição de série temporal (STL), Regressão linear e *Naive forecast.*

Para o estado da Paraíba, no melhor caso, Silva (2019) alcançou um *mean absolute percentage error* (MAPE) de 15,48%.

Na Índia, Doni e Sasipraba (2020) realizaram um estudo utilizando técnicas de *Deep Learning and Machine Learning* com intuito de prever casos de dengue. As previsões foram feitas por meio das técnicas LSTM, SVM, XGboost, RF e GAM. A avaliação dos resultados ficou a cargo do RMSE.

Em relação aos dados utilizados para treinamento dos modelos, foram utilizados dados pluviométricos, informações populacionais e registros de casos de dengue e mortes causadas por essa arbovirose. Todos os dados foram coletados do sistema do portal do governo Indiano e entre o período de 2015 e 2018. Sobre os resultados, a menor taxa de erro (42,00) foi obtida pelo LSTM.

A pesquisa de Xu et al. (2020), realizou previsões de casos de dengue, na China, por meio de dados climáticos, como, por exemplo, média de temperatura, média de precipitação e casos mensais de dengue. Os dados foram fornecidos pelo sistema de Vigilância de Doenças, Sistema de Meteorologia do país. Em relação ao período dos dados, eles foram entre 2005 e 2018.

As previsões foram feitas utilizando as técnicas LSTM, BPNN, GAM e SVR. Para a avaliação dos resultados foi utilizado o RMSE. Por fim, LSTM foi a técnica com menor taxa de erro de previsão: 36,50.

Carvajal et al. (2018) conduziram uma pesquisa com intuito de realizar previsão de casos de dengue em Manila, nas Filipinas, através de ML e de DL. Usando dados fornecidos pelo Governo entre os anos de 2009 e 2013, os modelos de ML foram treinados com informações meteorológicas e dados de notificação da dengue.

Para realizar as previsões, os autores utilizaram as técnicas RF, GAM e GB. A averiguação da performance foi feita através do RMSE e demonstrou como menor taxa de erro 0,29 para o RF.

Na China, Guo et al. (2017), através das técnicas SVR, LASSO, GAM e GBM, conseguiram realizar previsões com taxas de erro (RMSE) 0,2681, 2,0621, 4,4973 e 3,4529, para as respectivas técnicas. O SVR obteve o melhor resultado e os modelos de previsões foram criados utilizando dados epidemiológicos, oriundos do governo da China. Em relação ao período, os dados foram compreendidos entre os anos de 2011 e 2014.

Appice et al. (2020), no México, propuseram a criação de modelos de previsão de casos de dengue utilizando a técnica autoral AutoTic-NN (*AUTOencoding based*

TIme series Clustering with Near- est Neighbour) e as técnicas KNN, SVR e ARIMA. Foram utilizados dados históricos da doença e do clima entre 1985 e 2010 na criação dos modelos. Por fim, a técnica proposta pelos autores, AutoTic-NN, obteve a menor taxa de erro (RMSE= 5,18).

Os demais detalhes de implementação, bem como as arquiteturas dos modelos vencedores, estão demonstrados no Quadro 6.

Quadro 6 - Arquitetura e configurações dos modelos vencedores para os trabalhos relacionados

Trabalho	**Arquitetura da técnica vencedora**
SIPPYID et al. (2020)	O sistema foi desenvolvido em R, apoiado através da biblioteca *caret*. Não foi explicitado quais parâmetros foram utilizados no modelo vencedor. Contudo, a etapa de *tunning* de parâmetros foi realizada com as seguintes configurações: ● Algoritmo GMB: função *gbm* do r com os parâmetros: *interaction.depth* = c(1,5,9), *n.trees*=(1:5)*30, *shrinkage*=0.1, *n.minobsinnode*=10 ● Algoritmo ENET: função *enet* do r com os parâmetros: *alpha* = c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9) e *lambda* = c(0,0.001,0.01,0.1,0.25,0.5,1,2,2.5,3)
Silva (2019)	Utilizou a linguagem de programação R. O treinamento foi realizado através da função *stlf*(). As previsões foram realizadas através da chamada *forecast*(*STL*, 8) .
Doni e Sasipraba (2020)	Implementação realizada no google colab seguindo a arquitetura: camada de entrada, uma camada LSTM com 64 células e camada de saída. Os seguintes parâmetros foram alterados *activation=relu*, *dropout=0.5* e *ephocs=50.*
Xu et al. (2020)	Programação realizada por meio de P*ython* com apoio da biblioteca *TensorFlow*. Arquitetura da LSTM: camada de entrada, camada LSTM com 64 células e camada de saída. Os parâmetros *solver, dropout* e *epoch*s foram alterados para “adam”, 0,4 e 1000, respectivamente.

Continua...

Continuação

Trabalho	**Arquitetura da técnica vencedora**
Carvajal et al. (2018)	Os modelos foram criados utilizando a linguagem de programação R. As previsões foram realizadas através do pacote *randomForest*, versão 4.6.12, e utilizou 1000 árvores(*n.tree=1000*).
Guo et al. (2017)	Python foi a linguagem de programação utilizada. Contudo, não foi informado qual biblioteca de ML foi utilizada. O SVR foi configurado com o *kernel=linear*, e *C=0.005*.
Appice et al. (2020)	A previsão foi realizada através da técnica *nearest neighbour*, em java, com parâmetro *neighbors=12.*

Fonte: Autoria própria.

3 MATERIAL E MÉTODOS

3.1 Delineamento da pesquisa

Esta é uma pesquisa científica classificada como experimental. A partir de dados epidemiológicos, pluviométricos e sanitários, combinado com técnicas de *Machine Learning* e de *Deep Learning,* pretende-se criar e avaliar modelos para a previsão de internações e notificações causadas pela doença dengue para os municípios do estado da Paraíba.

A pesquisa foi desenvolvida no Núcleo de Tecnologias Estratégicas em Saúde (NUTES), unidade pertencente a Universidade Estadual da Paraíba (UEPB), na cidade de Campina Grande/PB. Devido à pandemia do COVID-19 e por se tratar de uma pesquisa exploratória utilizando *software* e dados abertos, todas as atividades foram desenvolvidas de forma remota.

O levantamento bibliográfico foi realizado através de uma revisão sistemática, presente no Apêndice A, e intitulada Previsão de casos de dengue através de *Machine Learning* e de *Deep Learning*: uma revisão sistemática (BATISTA et al., 2021). A revisão foi direcionada de acordo com o *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) (MOHER et al., 2009). O protocolo de estudo foi registrado no Open Science Framework (OSF) e está disponível através do link: https://osf.io/fqa57.

Para complementar o levantamento bibliográfico, foram realizadas pesquisas em bases eletrônicas das áreas de Medicina e de Ciência da Computação.

3.2 Fonte de dados

Foram utilizados dados mensais de internações e de notificações causadas por dengue para todos os 223 municípios da Paraíba entre os anos de 2010 e 2019. Os dados foram coletados através da ferramenta TABNET do DATASUS e exportados para arquivos textos no formato CSV (*Comma Separated Values*). As internações tiveram como fonte o Sistema de Informações Hospitalares do SUS (SIH/SUS). Em relação às notificações, a origem dos dados foi o Sistema de Informação de Agravos de Notificação (SINAN).

Além das informações sobre a doença, foram utilizados dados mensais sobre a pluviometria de cada município da Paraíba. Para esses, a fonte foi a Agência Executiva de Gestão das Águas do Estado da Paraíba (AESA). Os dados foram salvos em arquivo texto, no formato CSV, e compreenderam o período entre 2010 e 2019.

Complementar aos dados já listados, do Sistema Nacional de Informações sobre Saneamento (SNIS), para cada município, foram coletadas as informações anuais sobre água e esgoto entre os anos de 2010 e 2019. Adicionalmente, as informações de número da população também foram coletadas. Após a análise preliminar, optou-se por utilizar apenas os dados de índice de coleta de esgoto e índice de tratamento de esgoto para cada município paraibano.

De posse de todas as informações, foi criado um banco de dados relacional e as informações foram adicionadas nele.

3.3 Aspectos éticos

Todas as informações utilizadas neste projeto são de domínio público. Logo, de acordo com a resolução nº 510 de 7 de abril de 2016, não são necessários registros e aprovações junto aos Comitês de Ética em Pesquisa (CEP) e Comissão Nacional de Ética em Pesquisa (CONEP) (GUERRIERO, 2016).

3.4 Recursos de *software*

O banco de dados foi viabilizado por meio do sistema de gerenciamento de banco de dados MySql. A instância do banco foi hospedada na máquina do desenvolvedor e, semanalmente, foram realizadas rotinas de *backups* para preservação dos dados.

Em relação às demais atividades de programação, foi utilizada a linguagem de Python, versão 3.7.8. O sistema de previsão foi implementado utilizando técnicas de ML e de DL com base nas bibliotecas *Scikit-learn*, *Keras/TensorFlow* e apoiados pelo *framework* Pandas. As validações e testes estatísticos foram apoiados pelas bibliotecas *Scipy* e *Scikit*. Por fim, a geração de gráficos utilizou o *framework Matplotlib*.

Para o desenvolvimento da pesquisa foi utilizado um *notebook* pessoal com a seguinte configuração: sistema operacional Windows 10 pro, processador *Intel Core* i5 de 2,6 GHz, disco rígido SATA 512 GB e memória *RAM* de 16 GB (DDR3).

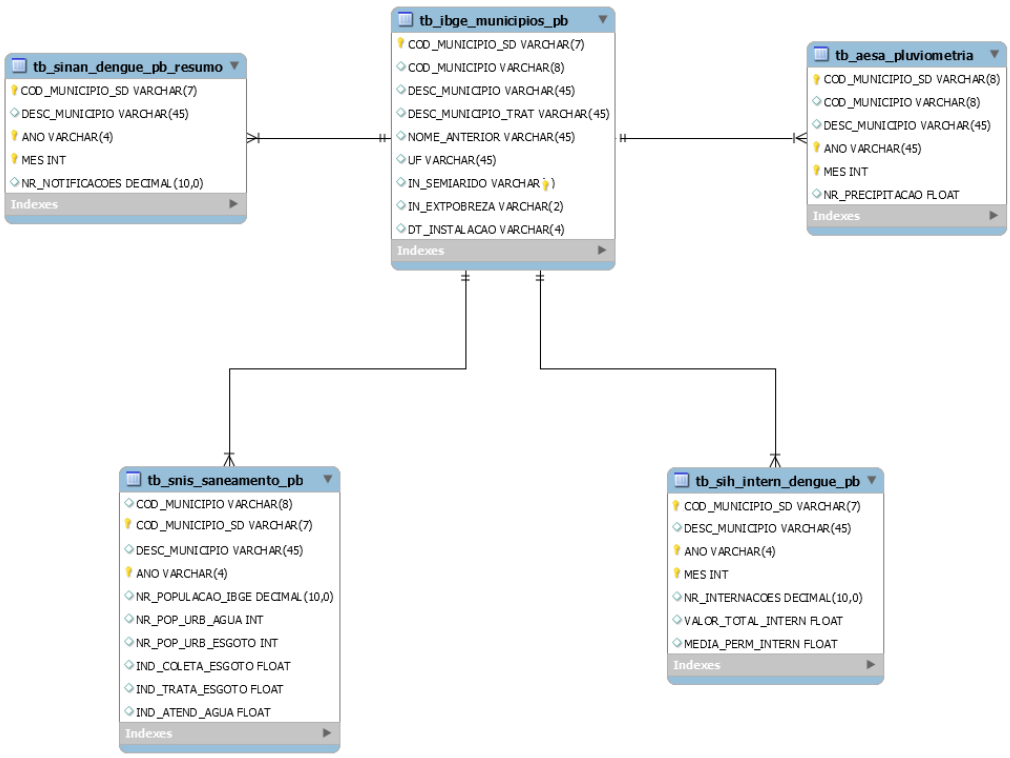
3.5 Sistema para predição de casos de internações e notificações

O sistema para predição de casos de internações e de notificações de dengue foi divido em quatro etapas: criação de banco de dados, criação de módulo de carga, criação de módulo de previsões e criação de módulo avaliativo. Os tópicos a seguir detalham as atividades desenvolvidas em cada uma das fases.

3.5.1 Banco de dados

Visando a continuidade do sistema e o reaproveitamento dos dados em outros estudos, foi criado um banco de dados, nomeado *DashDengue*, para armazenar os dados trabalhados. Após a criação da instância do banco e do *schema* de dados através da ferramenta *Workbench*, fornecida pelo MySql, foram criadas as tabelas responsáveis por representar e armazenar as informações de internações, notificações, pluviometria, saneamento e municípios. A Figura 12 representa o diagrama entidade-relacional para *schema DashDengue*.

Figura 12 - Diagrama entidade relacionamento para as tabelas do *schema* DashDengue



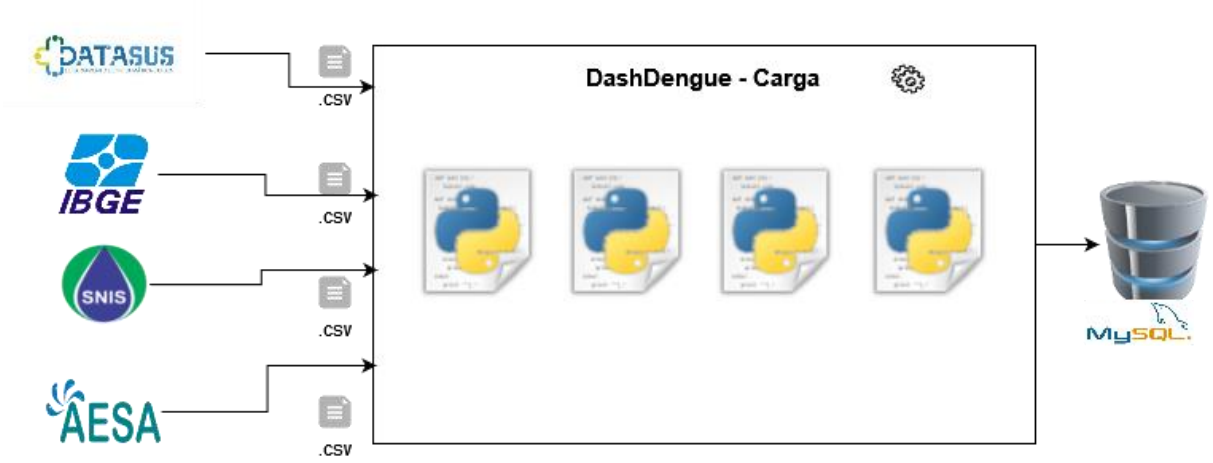
Fonte: Autoria própria

3.5.2 Módulo de carga

Ao analisar os arquivos .csv do SIH/SUS, SINAN, AESA e SNIS, foi verificado a falta de padronização e de interoperabilidade entre os sistemas. Com isso, através de *scripts* em P*ython*, todos os arquivos foram padronizados quanto ao *encoding,* à notação numérica e foram adicionados os códigos dos municípios e as descrições dos municípios fornecidos pelo IBGE.

A seguir e, mais uma vez utilizando P*ython*, foram criados *scripts* responsáveis por processar os arquivos .csv coletados e carregar as informações nas suas respectivas tabelas. O desenho da solução é demonstrado na Figura 13.

Figura 13 - Representação do módulo de carga de dados



Fonte: Autoria própria

3.5.3 Módulo de Previsões

Antes da execução do módulo de previsão, foi verificada a completude dos dados para os municípios da Paraíba. Feito isso, foi definido o critério para participação dessa pesquisa: caso o município contenha os dados de internações, de notificações, os dados pluviométricos e as informações sanitárias para o período entre 2010 e 2019 estará apto a participar da pesquisa.

Após a verificação, nove municípios possuem, na sua integralidade, os dados. São eles: Bayeux, Cabedelo, Cajazeiras, Campina Grande, Catolé do Rocha, João Pessoa, Monteiro, Patos e Santa Rita. Logo, esses serão os municípios presentes no trabalho.

Definidas as cidades, foram criados quatro cenários de pesquisa. Os cenários foram idealizados utilizando a variação de período entre 10 e 5 anos e se haveria ou não tratamento de *outliers.* De acordo com Géron (2019), a utilização de dados com *outliers* podem influenciar negativamente em modelos de previsão. O Quadro 7 ilustra o detalhe dos cenários propostos.

Quadro 7 - Cenários de pesquisa propostos

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

Fonte: Autoria própria

Outro desafio durante a previsão por meio de técnicas de *Machine Learning* e de *Deep Learning* é achar a combinação ideal dos atributos previsores (HARISSON, 2020). Sendo assim, foram geradas 8 combinações utilizando os atributos previsores número de internação/notificação, pluviometria mensal, índice de coleta de esgoto e, finalmente, índice de tratamento de esgoto. O Quadro 8 e Quadro 9 demonstram as variações propostas.

Quadro 8 - Combinação dos atributos previsores para a previsão de internações

Combinação	nr_internações	nr_precipitação	coleta_esgoto	tratamento_esgoto
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

Fonte: Autoria própria

Quadro 9 - Combinação dos atributos previsores para a previsão de notificações

Combinação	nr_notificações	nr_precipitação	coleta_esgoto	tratamento_esgoto
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

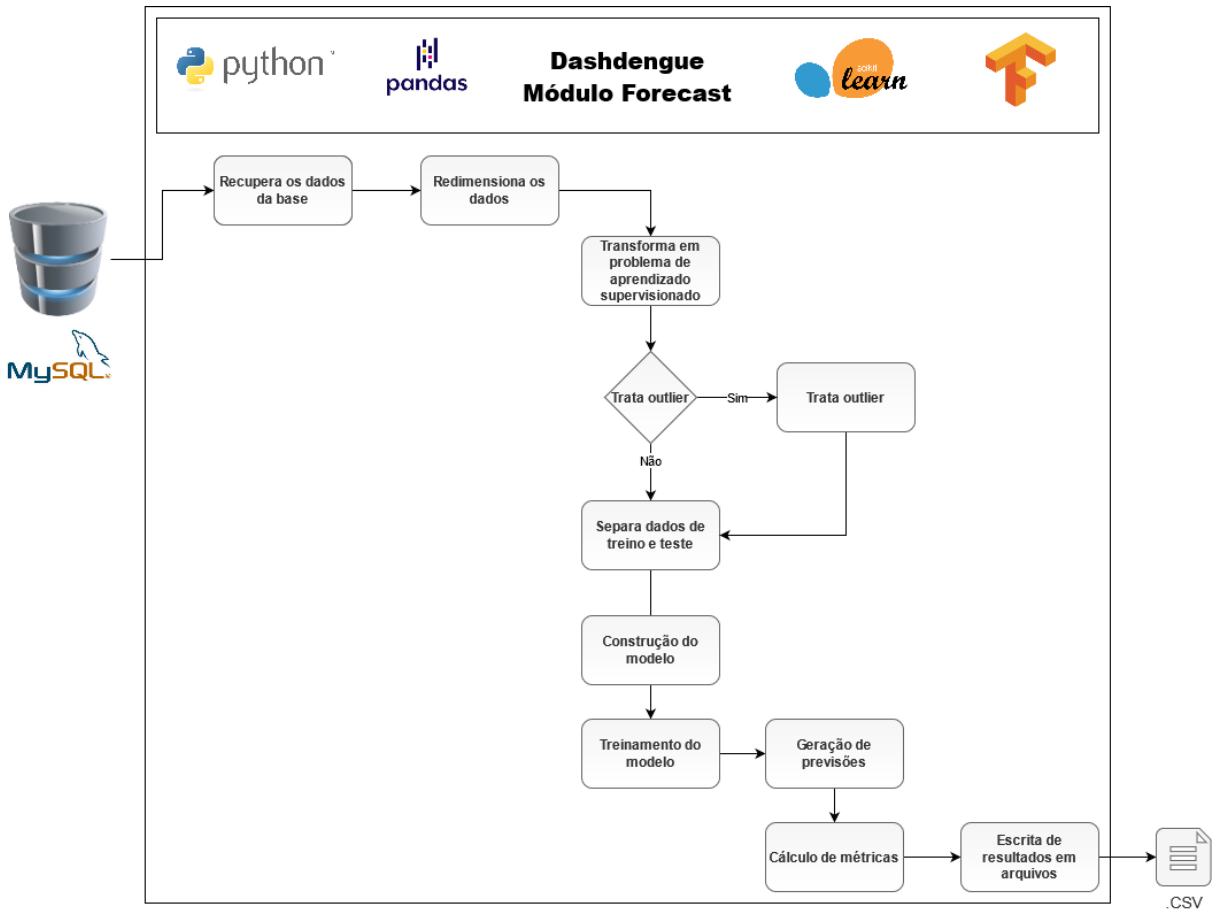
Fonte: Autoria própria

Por questões de performance e limitação de recursos computacionais, os processamentos de casos de internações e de notificações descritos nesta seção (3.4.3) foram realizados em momentos distintos.

3.5.3.1 Escolha dos cenários e atributos previsores

Para cada município elegível, foram treinados e testados modelos de acordo com os cenários propostos e as combinações dos atributos previsores. O objetivo desta etapa foi encontrar qual é a melhor configuração para cada cidade. O fluxo de atividades é descrito na Figura 3.

Figura 14 - Atividades realizadas pelo módulo *forecast*



Fonte: Autoria própria

De acordo com Kuhn, Max e Johnson (2019), as técnicas de ML e de DL são sensíveis à variação de escala dos dados previsores. Dessa forma, para evitar o problema, foi realizada a normalização dos dados através do método *StandardScaler* pertencente a biblioteca *Sklearn*. Como retorno, todos os dados foram normalizados entre o intervalo de 0 a 1.

A etapa seguinte foi transformar os dados em problema de aprendizado supervisionado. Aqui, utilizando o *framework* Pandas, além dos atributos previsores, foram adicionadas de 1 a 4 *lags* (informações passadas) para cada atributo das combinações. O número máximo de 4 *lags* foi definido para manter a correta proporção entre dados de treino e testes. O tratamento de *outliers,* quando aplicável, ocorreu através da biblioteca *Stats* do P*ython*. Caso a biblioteca apontasse o valor mensal de notificação ou internação como um *outlier* na amostra, todo aquele mês foi desconsiderado do conjunto de dados.

Na sequência, os dados foram separados entre 80% para treino e 20% para teste e treinados utilizando as técnicas RF, SVR, RN, LSTM e CNN. A escolha da porcentagem está de acordo com a literatura (GÉRON, 2019). Em relação às técnicas, essas foram escolhidas de acordo com a análise produzida pela revisão sistemática realizada e presente no Apêndice A. O Quadro 10 contém a arquitetura e configuração de parâmetros utilizada por cada uma das técnicas utilizadas. A escolha desses parâmetros foi realizada após a realização de testes experimentais.

Quadro 10 - Arquitetura e configuração dos parâmetros utilizados no treinamento dos modelos

Técnica	**Arquitetura**	**Configuração de parâmetros**
RF	Floresta com 500 árvores.	*n_estimators=500, criterion="mse", max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0, max_features="auto", max_leaf_nodes=None, min_impurity_decrease=0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, ccp_alpha=0, max_samples=None*
SVR	Não se aplica	*kernel='rbf', degree=3, gamma='scale', coef0=0, tol=0.001, C=1, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=- 1*

Continua...

Continuação

Técnica	Arquitetura	Configuração de parâmetros
LSTM	● Camada de entrada; ● Camada LSTM com 50 neurônios; ● Camada densa com um neurônio.	*units=50, activation='tanh', recurrent_activation='sigmoid', use_bias=True, kernel_initializer='glorot_uniform', recurrent_initializer='orthogonal', bias_initializer='zeros', unit_forget_bias=True, kernel_regularizer=None, recurrent_regularizer=None, bias_regularizer=None, activity_regularizer=None, kernel_constraint=None, recurrent_constraint=None, bias_constraint=None, dropout=0, recurrent_dropout=0, return_sequences=False, return_state=False, go_backwards=False, stateful=False, time_major=False, unroll=False,* *epochs=50, batch_size=12*
CNN	● Camada de entrada; ● Camada Conv1D com 16 filtros; ● Camada LSTM com 50 neurônios; ● Camada densa com 1 neurônio.	*Filters=16, kernel_size, strides=1, padding='valid', data_format='channels_last', dilation_rate=1, groups=1, activation=None, use_bias=True, kernel_initializer='glorot_uniform', bias_initializer='zeros', kernel_regularizer=None, bias_regularizer=None, activity_regularizer=None, kernel_constraint=None, bias_constraint=None, units=50, activation='tanh', recurrent_activation='sigmoid', use_bias=True, kernel_initializer='glorot_uniform', recurrent_initializer='orthogonal', bias_initializer='zeros', unit_forget_bias=True, kernel_regularizer=None,*

Continua...

Continuação

Técnica	Arquitetura	Configuração de parâmetros
CNN	● Camada de entrada; ● Camada Conv1D com 16 filtros; ● Camada LSTM com 50 neurônios; ● Camada densa com 1 neurônio.	*recurrent_regularizer=None, bias_regularizer=None, activity_regularizer=None, kernel_constraint=None, recurrent_constraint=None, bias_constraint=None, dropout=0, recurrent_dropout=0, return_sequences=False, return_state=False, go_backwards=False, stateful=False, time_major=False, unroll=False,* *epochs=50, batch_size=12*

Fonte: Autoria própria

Com o os modelos treinados, para cada cidade e combinação, foram geradas as previsões iniciais, desfeita a normalização e, por fim, foi calculado o RMSE entre o valor previsto pelos modelos e o valor observado (valores de testes). O resultado individual foi armazenado em arquivo texto *.csv*.

Com base na análise do menor RMSE, um *script* em *Python* analisou todos os resultados produzidos e indicou o melhor cenário, a melhor combinação de atributos previsores e a quantidade de *lags* para as cidades.

3.5.3.2 Ajuste nos hiperparâmetros

Conforme Deitel e Deitel (2019), uma das vantagens de utilizar técnicas de ML e de DL é a quantidade de ajustes e de combinações possíveis. Com isso, definido o cenário, a combinação de atributos previsores e a quantidade de *lags* para cada cidade, a próxima etapa foi submeter as técnicas de ML e de DL a variações de seus parâmetros.

O Quadro 11 contém a listagem dos parâmetros e os valores assumidos por cada atributo. A lista final desses parâmetros e os valores foram definidos após testes experimentais.

Quadro 11 - Variações de parâmetros e lista de valores candidatos por técnica de ML e DL

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]

Continua...

Continuação

Técnica	Parâmetro	Lista de valores
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}

Fonte: Autoria própria

Cada configuração descrita no quadro acima foi submetida a novo treinamento, teste e avaliação de erro através do RMSE. A configuração de parâmetros que produziu menor taxa de erro foi utilizada na etapa final de geração das previsões. A porcentagem de treino e teste seguem, respectivamente, 80% e 20%.

3.5.3.3 Geração das previsões

Após encontrar o melhor cenário, combinação de atributos previsores e configuração de cada técnica, para cada cidade, os modelos propostos foram submetidos a treinamento final. Visando evitar *o overfitting,* foram habilitadas as paradas antecipadas fornecidas pelas bibliotecas *Sklearn* e *Keras*/*TensorFlow*. Finalmente, foram geradas as previsões e calculadas as taxas de erro (RMSE). As previsões foram geradas de acordo com o cenário escolhido e salvas em arquivo texto *.csv*. Logo, devido à porcentagem e montagem de cenários de teste, as previsões das cidades foram feitas de forma mensal entre 12 e 24 meses. A porcentagem de treino e testes, assim como nos outros passos, seguiu a divisão de 80% de dados para treino e 20% de dados para teste.

Por questões estocásticas de algumas técnicas, separados em quatros *rounds*, foram realizadas 100 execuções para cada uma das configurações finais. A média de cada *round* foi calculada para as cidades e serviu como parâmetro para escolha da melhor técnica.

3.5.4 Módulo avaliativo

3.5.4.1 Escolha da melhor técnica

Não foram encontrados na literatura estudos realizando previsões de internações e notificações para os municípios aqui listados. De acordo com (KIRBY; PARAMAGURU; WARREN, 2015), para cenários como esses, é recomendado realizar a comparação com a técnica *Naive Forecast*. Caso a abordagem da pesquisa tenha resultados melhores que o *Naive Forecast,* afirma-se que houve sucesso em realizar as previsões. Por fim, os resultados obtidos também foram comparados com a técnica ARIMA (APPICE et al., 2020).

3.5.4.2 Validação estatística

Logo, após determinar a melhor técnica para cada cidade, os resultados foram submetidos a testes estatísticos para ratificar os achados desta pesquisa. Inicialmente, foi testada a normalidade dos resultados de cada *round.* Para isso, o teste de Shapiro-Wilk, fornecido pela biblioteca *Scipy*, foi utilizado. Em caso de seguir a distribuição normal e, utilizando o teste ANOVA, foram realizados os testes para verificar se há diferença estatística entre os resultados produzidos pelas técnicas. Nos casos de diferença apontada, o teste de Tukey demonstrou, numericamente e graficamente, a diferença, par a par, entre as técnicas.

3.5.4.3 Geração de gráficos para validação visual

A etapa final foi gerar os gráficos das previsões obtidas por cada técnica versus o valor real de internações e notificações de dengue. Os gráficos foram gerados com apoio da biblioteca *Maplotib*, do P*ython*.

4 RESULTADOS E DISCUSSÃO

Nesta seção serão apresentados os resultados e as discussões produzidos. Os resultados seguirão a ordem de apresentação:

1. Determinação do melhor cenário e combinação de atributos previsores;
2. Ajuste nos hiperparâmetros dos algoritmos;
3. Geração das previsões e determinação da melhor técnica por cidade;
4. Avaliação dos resultados produzidos.

Por questões didáticas, cada tópico anterior foi dividido entre os resultados para os casos de internação e os resultados para os casos de notificação.

4.1 Determinação do melhor cenário e combinação de atributos previsores

A primeira etapa do sistema de previsão foi determinar quantos anos de dados, se os dados considerados *outliers* seriam excluídos, quais as combinações de atributos e a quantidade de *lags* produziriam o melhor resultado para as cidades de Bayeux, Cabedelo, Cajazeiras, Campina Grande, Catolé do Rocha, João Pessoa, Monteiro, Patos e Santa Rita. A avaliação dos resultados foi feita por meio da análise do RMSE e as previsões com menor taxa de erro RMSE foram consideradas as melhores (CARVAJAL et al., 2018).

Para cada variação desses fatores, foram gerados modelos, realizados treinamentos, testes e avaliações das previsões. Os treinamentos e testes foram realizados utilizando cada uma das técnicas abordadas por este trabalho: RF, SVR, MLP, LSTM e CNN.

Os tópicos seguintes demonstram os melhores resultados por cenário e a melhor combinação de atributos previsores para cada uma das cidades.

4.1.1 Internações

O cenário 1 considerou a utilização de dados entre os anos de 2010 e 2019 e não houve exclusão de *outliers*. A Tabela 1 contém os menores valores de RMSE por combinação de atributos previsores para cada cidade. A escala de cores demonstra a

relação dos resultados, sendo, em verde, para melhor e, em vermelho, para o pior resultado.

Tabela 1 - Taxa de erro calculada para previsões de internações com dados entre 2010 e 2019 e sem tratamento de *outlier* (cenário 1)

Município	Combinação 1	Combinação 2	Combinação 3	Combinação 4	Combinação 5	Combinação 6	Combinação 7	Combinação 8
Bayeux	0,629282246	0,745255992	0,817199893	0,809463945	0,742640206	0,763919607	0,803127424	0,858739708
Cabedelo	2,26201145	2,547808115	2,529639646	2,697450172	2,402894754	2,68808105	2,534445072	2,58633309
Cajazeiras	1,391392303	1,363364235	1,396474234	1,372412394	1,372171708	1,368029873	1,41044091	1,379444605
Campina Grande	40,794585	34,36516611	37,69701711	38,53594935	37,69340199	36,82040274	40,24292856	35,99825863
Catolé do Rocha	0,676490506	0,640474269	0,467516056	0,502644701	0,657002175	0,588892252	0,472378017	0,548596368
João Pessoa	11,12981377	10,11886389	12,31868588	12,65871925	11,92016448	11,70162493	13,45346562	11,38536242
Monteiro	1,092520342	1,078429675	1,066173628	1,063290833	1,088327801	1,061740066	1,092630551	1,061655456
Patos	0,65707014	0,85370814	0,900998342	0,89963724	0,97807702	0,853693138	0,950582716	0,899605078
Santa Rita	1,244421164	1,208046093	1,106140514	1,279022621	1,181781403	1,250355844	1,149943578	1,22382296

Fonte: Autoria própria

Para o cenário 2, foram utilizados dados entre os anos de 2015 e 2019 e não houve exclusão de registros considerados *outliers*. Os melhores resultados das combinações de atributos são demonstrados na Tabela 2.

Tabela 2 - Taxa de erro calculada para previsões de internações com dados entre 2015 e 2019 e sem exclusão de *outliers* (cenário 2)

Município	Combinação 1	Combinação 2	Combinação 3	Combinação 4	Combinação 5	Combinação 6	Combinação 7	Combinação 8
Bayeux	0,77063502	0,923259985	1,028755559	1,014510728	0,749214033	0,926770378	1,032504237	1,072293359
Cabedelo	4,09641835	4,010448324	3,905085445	4,620438707	4,189366388	4,312065143	4,190581721	4.,42685298
Cajazeiras	1,931951345	1,859515032	2,25030147	2,082782263	1,891843633	1,902214078	2,06877221	2,040535953
Campina Grande	4,973578606	4,325556391	3,461959869	3,041739027	3,547390673	4,767319679	4,158600735	2,781317923
Catolé do Rocha	0,481654786	0,474148832	0,501682313	0,478717174	0,494723155	0,472629318	0,474731201	0,468952451
João Pessoa	17,80916702	18,45225271	19,66129195	19,34475682	18,73528947	17,12518945	19,27236011	19,81908927
Monteiro	1,575306425	1,592319163	1,65018325	1,651486803	1,577750191	1,590295177	1,653445142	1,607425234
Patos	1,243322163	1,180746417	1,207995349	1,236234297	1,230541209	1,192278864	1,192817757	1,232852759
Santa Rita	1,126177754	1,301260818	1,352428558	1,640842923	1,130180664	1,352682766	1,72168467	1,576015712

Fonte: Autoria própria

O cenário de pesquisa 3 realizou treinamento e testes com as combinações utilizando o período de dados entre 2010 e 2019. Adicionalmente, houve a exclusão dos registros considerados *outliers*. Os resultados são apresentados na Tabela 3.

Tabela 3 - Taxa de erro calculada para previsões de internações com dados entre 2010 e 2019 e com exclusão de *outliers* (cenário 3)

Município	Combinação 1	Combinação 2	Combinação 3	Combinação 4	Combinação 5	Combinação 6	Combinação 7	Combinação 8
Bayeux	0,689157314	0,785114694	0,809743272	0,789549511	0,776811673	0,737001743	0,803513002	0,846605426
Cabedelo	0,954583611	0,981526678	0,996024906	0,999540947	0,962365429	0,970518515	0,986743613	0,997949301
Cajazeiras	1,526062224	1,503062337	1,609739966	1,650906579	1,523334672	1,478871313	1,52071152	1,649308482
Campina Grande	11,61077231	11,08867909	11,04893735	11,14821455	11,65818779	11,41597985	11,07101867	10,98186141
Catolé do Rocha	0,616060725	0,598881144	0,480130818	0,51282386	0,6189917	0,600854003	0,490883548	0,574776171
João Pessoa	10,47772156	10,44645941	10,30173218	11,11756644	10,78606822	10,22137025	11,60765941	10,10272226
Monteiro	1,083435814	1,09455427	1,064673285	1,067142602	1,085869109	1,087233074	1,082108993	1,067426236
Patos	0,95060389	0,852690268	0,931471185	0,935269583	0,927681412	0,852932463	0.,44767181	0,9254248
Santa Rita	**1,164378319**	**1,127796332**	**1,098166468**	**1,273602295**	**1,144109603**	**1,230481407**	**1,150766752**	**1,230678024**

Fonte: Autoria própria

Os resultados para o cenário 4 estão presentes na Tabela 4. Aqui, foram considerados os dados entre os anos de 2015 e 2019 e houve a exclusão de registros considerados como *outliers*.

Tabela 4 - Taxa de erro calculada para previsões de internações com dados entre 2015 e 2019 e com exclusão de *outliers* (cenário 4)

Município	Combinação 1	Combinação 2	Combinação 3	Combinação 4	Combinação 5	Combinação 6	Combinação 7	Combinação 8
Bayeux	0,897446243	1,021830863	0,956586392	1,040827136	0,894713734	0,905860903	1,016367354	1,097805506
Cabedelo	2,420292438	2,398553714	2,440260895	2,342165087	2,200801867	2,335448739	2,541845706	2,386125653
Cajazeiras	1,446091629	1,342038155	1,69761261	1,650940705	1,470093421	1,366936986	1,733275163	1,635838364
Campina Grande	5,225328902	3,192438253	3,667968968	2,664844792	4,724848662	3,889911607	3,685344206	2,343211284
Catolé do Rocha	0,403745053	0,381533663	0,396157274	0,406100176	0,397525081	0,384095489	0,397630075	0,407838494
João Pessoa	14,40648549	14,80097908	13,69736167	10,82774107	14,57160493	12,49004911	13,55689816	12,33829671
Monteiro	1,065433329	1,08081476	1,08772746	1,065365102	1,071692196	1,066259149	1,081329452	1,059195869
Patos	0,456793479	0,454293686	0,458202652	0,71266238	0,455704826	0,461922112	0,457878914	0,47243351
Santa Rita	1,081508751	1,143400779	1,176559841	1,466532146	1,088953898	1,158309112	1,281316904	1,293986403

Fonte: Autoria própria

Após o processamento de todos os cenários de pesquisa, foi verificado qual cenário, combinação de atributos e quantidade de *lags* obtiveram o menor valor de RMSE. A Tabela 5 contém a consolidação dos resultados.

Tabela 5 - Resultados contendo a melhor combinação de parâmetros, período, se houve exclusão de *outliers* e a quantidade de *lags* para os casos de internações

Município	Melhor Rmse	Parâmetros previsores	Período	Exclusão de *outliers*	Lags
Bayeux	0,629282246	nr_internações, nr_precipitação, coleta_esgoto e tratamento_esgoto	2010 – 2019	Não	4
Cabedelo	0,954583611	nr_internações, nr_precipitação, coleta_esgoto e tratamento_esgoto	2010 – 2019	Sim	4
Cajazeiras	1,342038155	nr_internações e nr_precipitação	2015 – 2019	Sim	3
Campina Grande	2,343211284	nr_internações	2015 – 2019	Sim	2
Catolé do Rocha	0,381533663	nr_internações e nr_precipitação	2015 – 2019	Sim	4
João Pessoa	10,10272226	nr_internações	2010 – 2019	Sim	4
Monteiro	1,059195869	nr_internações	2015 – 2019	Sim	1
Patos	0,454293686	nr_internaçõe e nr_precipitação	2015 – 2019	Sim	3
Santa Rita	1,081508751	nr_internações, nr_precipitação, coleta_esgoto e tratamento_esgoto	2015 – 2019	Sim	3

Fonte: Autoria própria

De acordo com critérios econômicos, sociais e políticos, o IBGE divide a Paraíba em quatro mesorregiões: Mata Paraibana, Agreste Paraibano, Borborema e Sertão Paraibano. Conforme a classificação, Bayeux, Cabedelo e João Pessoa são classificados como Mata Paraibana. As cidades de Cajazeiras, Catolé do Rocha e Patos estão presentes no Sertão Paraibano. Por fim, Campina Grande e Monteiro pertencem, respectivamente, ao Agreste Paraibano e Borborema (PARAÍBA, 2016).

Ao analisar os resultados, verificou-se que os municípios da mesma mesorregião, na maioria dos casos, utilizaram os mesmos atributos previsores. As cidades de Bayeux, de Cabedelo e de Santa Rita, pertencentes à Mata Paraibana, obtiveram melhores resultados ao utilizar a combinação 1 de parâmetros (nr_internações, nr_precipitação, coleta_esgoto e tratamento_esgoto). A exceção ocorreu para João Pessoa, pois apenas os números históricos de internações foram utilizados durante as previsões. Esse fato pode ter relação devido a João Pessoa possuir, em média, o maior índice de coleta de esgoto (69,37%) em comparação a Bayeux (14,52%), a Cabedelo (19,11%) e a Santa Rita (7,45%).

Em relação aos municípios Cajazeiras, Catolé do Rocha e Patos, classificados como Sertão Paraibano, os atributos número de internações e pluviometria mensal obtiveram melhores resultados nessas cidades. Além das similaridades sociais e econômicas, entre os anos de 2010 e 2019, os municípios registaram valores similares de pluviometria (média de 68 mm) por mês. Para Campina Grande e para Monteiro, a utilização de informações históricas da doença foi a melhor combinação de parâmetros previsores. Para Monteiro esse fato acontece, provavelmente, por ela

estar entre os municípios estudados com a menor média pluviométrica (48 mm). Ademais, Monteiro possui o maior índice de coleta de esgoto (90,84%).

Sobre a utilização do período de anos, os dados entre 2015 e 2019 foram utilizados em 66,66% das cidades aqui trabalhadas. A estratégia de excluir os dados considerados como *outliers* mostrou relevância, já que o tratamento foi utilizado em oito das noves cidades.

Por fim, em relação à adição de *lags*, o acréscimo de 4 *lags* obteve melhores resultados em 44,44% das cidades, o acréscimo de 3 *lags* foi utilizado em 33,33% das cidades e, finalmente, 2 *lags* e 1 *lag* representam uma cidade cada (11,11%). Na pesquisa de Ribeiro et al. (2006), os autores afirmam que o reflexo do aumento de chuvas e temperaturas implicam no crescimento dos números de dengue a partir do segundo mês e se estendem até o quarto mês desde as primeiras observações das alterações climáticas. Logo, a utilização das *lags* estão condizentes aos achados do estudo de Ribeiro.

4.1.2 Notificações

Os resultados produzidos para previsões de notificações no cenário 1 de pesquisa estão na Tabela 6. Nesse cenário, foram utilizados dados entre os anos de 2010 e 2019 e não houve exclusão de registros considerados *outliers*.

Tabela 6 - Taxa de erro calculada para previsões de notificações com dados entre 2010 e 2019 e sem tratamento de *outliers* (cenário 1)

Município	Combinação 1	Combinação 2	Combinação 3	Combinação 4	Combinação 5	Combinação 6	Combinação 7	Combinação 8
Bayeux	2,892719943	2,809548746	3,072395263	3,108623509	2,723301562	2,918210154	2,836833089	2,960813541
Cabedelo	6,089371485	6,709214509	6,07920243	5,984814561	7,018555067	6,248661868	6,035178503	5,948505016
Cajazeiras	13,48157191	12,3118508	14,51680215	13,3339248	13,48509424	12,31193882	14,51461671	13,33914034
Campina Grande	25,67857307	27,61365127	31,2148628	31,74545358	26,82539728	27,33564924	30,97384608	32,17710163
Catolé do Rocha	4,70743428	5,237411988	4,705116011	4,71419508	4,414971668	5,093951986	4,70494225	4,979698337
João Pessoa	245,6248252	228,7207009	277,8162993	298,0990523	247,5211237	288,1984577	305,2510503	292,3013343
Monteiro	12,1477142	10,88405193	11,74259452	10,74769056	11,75281617	11,18620395	10,89417333	10,6605251
Patos	4,319883614	3,750077816	4,60536446	4,551718428	4,290272708	3,995099845	4,614936228	4,760298542
Santa Rita	18,6713354	17,61868456	17,31061755	17,76717014	17,8890661	15,6550903	18,81290382	17,04504799

Fonte: Autoria própria

As taxas de erros para as combinações de atributos no cenário 2 (período entre 2015 e 2019 e sem exclusão de *outliers* são demonstradas na Tabela 7.

Tabela 7 - Taxa de erro calculada para previsões de notificações com dados entre 2015 e 2019 e sem tratamento de *outliers* (cenário 2)

Município	Combinação 1	Combinação 2	Combinação 3	Combinação 4	Combinação 5	Combinação 6	Combinação 7	Combinação 8
Bayeux	3,622670136	2,531602327	3,255936302	3,581610901	2,708971554	3,482470494	3,601121472	3,465371693
Cabedelo	7,235588841	8,078261261	7,849933241	7,577414181	6,518404348	7,972027212	7,781714138	7,576616647
Cajazeiras	17,45863472	15,1443923	18,48323667	16,98355917	17,45660833	15,5346792	17,42991134	16,99522949
Campina Grande	10,80180214	11,45989153	9,271424283	9,293506347	8,505701618	9,888701946	9,266240641	9,71140323
Catolé do Rocha	2,99518566	2,111677517	2,755449682	2,418546369	2,903082788	2,168784201	2,767761758	2,433310318
João Pessoa	334,0369565	309,4484143	379,4846485	371,0733768	319,5873223	341,3206099	392,3497598	384,7177518
Monteiro	15,55520365	15,44716383	14,56675363	13,80149454	15,68664074	14,87537938	14,46975433	14,62025597
Patos	4,975387272	5,240662871	5,671670557	5,176358223	4,901163095	5,19209349	5,462804031	5,175659873
Santa Rita	20,04282399	20,06927886	17,68889167	18,39983522	18,34406617	18,20867829	19,93472653	20,52962663

Fonte: Autoria própria

O cenário 3 de pesquisa utilizou dados entre os anos de 2010 e 2019 e realizou a exclusão de registros considerados *outliers*. A Tabela 8 relata os resultados produzidos por combinação para cada uma das cidades.

Tabela 8 - Taxa de erro calculada para previsões de notificações com dados entre 2010 e 2019 e com tratamento de *outliers* (cenário 3)

Município	Combinação 1	Combinação 2	Combinação 3	Combinação 4	Combinação 5	Combinação 6	Combinação 7	Combinação 8
Bayeux	2.861946167	2.752980901	2.751332103	2.806418728	2.768536489	2.808596409	2.849240989	2.76506387
Cabedelo	6.267904677	6.658108988	6.196167872	6.204819534	6.969907802	6.232276033	6.263578585	6.228636868
Cajazeiras	13.35170754	12.80326694	14.53493474	13.79389126	13.30983278	12.80717362	14.45037805	12.89480798
Campina Grande	24.11827649	24.82322607	28.33032376	27.53926928	25.58035754	25.30867162	26.79551256	28.66907566
Catolé do Rocha	4.538761854	4.653640819	4.421732891	4.6181003	4.440532588	4.535127902	4.670280985	4.640071716
João Pessoa	193.6886452	180.7634559	208.4803755	210.9471016	192.2969825	211.2222583	203.4315858	186.4814392
Monteiro	12.4430359	10.9991988	12.89107757	11.05839902	11.64079867	11.17648432	10.89821426	10.67968096
Patos	4.422575623	3.695983662	5.184594099	5.106094251	4.311543959	3.994045611	5.174333244	5.191116735
Santa Rita	15.40856938	15.65595541	15.84633036	16.10025689	13.66087647	15.78442805	15.7009558	15.95503789

Fonte: Autoria própria

Os resultados produzidos para o cenário 4 de pesquisa, os dados entre 2015 e 2019 e com a exclusão de *outliers* estão presentes na Tabela 9.

Tabela 9 - Taxa de erro calculada para previsões de notificações com dados entre 2015 e 2019 e com tratamento de *outliers* (cenário 4)

Município	Combinação 1	Combinação 2	Combinação 3	Combinação 4	Combinação 5	Combinação 6	Combinação 7	Combinação 8
Bayeux	3,296591206	3,396288991	3,020115586	3,433055419	3,133514557	3,007131751	3,564660831	2,915104832
Cabedelo	6,321484978	6,659841779	6,983761919	7,091830602	6,115048122	6,218981742	7,340682084	7,134096738
Cajazeiras	18,04870375	13,58747474	17,58022842	16,81749862	15,53572879	13,59757922	17,89038485	17,29823051
Campina Grande	9,217484349	10,22428563	9,007925134	9,262316562	8,265547507	9,031923039	9,078234115	8,542571022
Catolé do Rocha	2,56683883	2,432091031	2,492851983	2,586844565	2,685475771	2,441012351	2,512209708	2,606848693
João Pessoa	310,6178576	306,143709	400,0859824	363,5092889	335,0406992	343,6138657	371,6273807	406,5514773
Monteiro	15,48873505	15,09633641	14,35621332	13,83402593	15,58120921	14,8411901	14,01089607	15,97886261
Patos	5,375867437	3,251664197	5,515450915	5,12283976	5,14535538	3,257364704	5,516405629	5,149704905
Santa Rita	17,95521457	16,77086878	19,43623302	19,17136196	13,40766199	18,46659063	19,16595206	18,8055212

Fonte: Autoria própria

Após as gerações de todos os resultados paras os quatro cenários de pesquisa, um *script* em python identificou, com base na análise do menor RMSE, qual é a melhor configuração de anos, a combinação de atributos previsores, se houve exclusão de *outliers* e quantas *lags* foram utilizadas para cada cidade. A Tabela 10 resume os melhores resultados alcançados.

Tabela 10 - Resultados contendo a melhor combinação de parâmetros, período, se houve exclusão de *outliers* e a quantidade de *lags* para os casos de notificações

Município	Melhor Rmse	Parâmetros previsores	Período	Exclusão de *outliers*	Lags
Bayeux	2,531602327	nr_notificações e nr_precipitação	2015-2019	Não	3
Cabedelo	5,948505016	nr_notificações	2010-2019	Não	3
Cajazeiras	12,3118508	nr_notificações e nr_precipitação	2010-2019	Não	2
Campina Grande	8,265547507	nr_notificações, nr_precipitação e coleta_esgoto	2015-2019	Sim	2
Catolé do Rocha	2,111677517	nr_notificações e nr_precipitação	2015-2019	Não	2
João Pessoa	180,7634559	nr_notificações e nr_precipitação	2010-2019	Sim	2
Monteiro	10,6605251	nr_notificações	2010-2019	Não	2
Patos	3,251664197	nr_notificações e nr_precipitação	2015-2019	Sim	2
Santa Rita	13,40766199	nr_notificações, nr_precipitação e coleta_esgoto	2015-2019	Sim	1

Fonte: Autoria própria

Ao analisar os resultados, os municípios de Cajazeiras, de Catolé do Rocha e de Patos, pertencentes ao Sertão Paraibano, obtiveram as menores taxas de erro utilizando a combinação de parâmetros: número de notificações e valor mensal de pluviometria. No caso das previsões de internações, para esses mesmos municípios, os atributos números de internações mais o valor mensal de pluviometria produziram os melhores resultados. Logo, diante desse cenário, podemos concluir que existem

um padrão para melhor prever casos de internação e de notificação de dengue para as cidades do Sertão: utilizar dados epidemiológicos e a observar a pluviometria.

Em relação à mesorregião da Borborema, representada por Monteiro, para as previsões de notificações, os melhores resultados foram obtidos utilizando apenas os dados epidemiológicos da doença. O mesmo comportamento foi observado durante a busca de parâmetros para as previsões de internações. Com isso, podemos afirmar que para prever casos de internações e de notificações de dengue para Monteiro, a utilização dos dados epidemiológicos é a melhor alternativa.

A utilização do número de notificações, da pluviometria mensal e do índice de coleta de esgoto mostrou maior relevância para a previsão de notificações em Campina Grande. Anteriormente, durante as previsões de internações, foram utilizados apenas os dados epidemiológicos.

Finalizando a análise dos atributos previsores, ao comparar as escolhas de parâmetros para internações e para notificações, houve alterações dos atributos para as cidades da Mata Paraibana (Bayeux, Cabedelo, João Pessoa e Santa Rita). Para Bayeux e para João Pessoa, a melhor combinação foi utilizar os números de notificações e de pluviometria mensal. Nos casos de Cabedelo os melhores resultados foram produzidos utilizando os números de e notificações e, para Santa Rita, foi por meio dos números de notificações mais pluviometria e índice de coleta de esgoto.

A respeito da utilização do período dos dados, a escolha pelo período entre 2015 e 2019 foi empregada em cinco dos noves municípios durante as previsões de notificações e, assim como nas previsões de internações, foi a maioria entre os municípios estudados.

Acerca das exclusões dos *outliers,* o número de cidades que fizeram o uso desse artifício nas previsões de notificações foi de quatro em detrimento a oito cidades nas previsões de internações. Os números podem apontar uma melhor qualidade nos dados de notificações fornecidos pelo SINAN em detrimento aos dados de internações fornecidos pelo SIH. Contudo, não é escopo deste projeto se aprofundar sobre este tema.

Por fim, em relação ao uso de *lags*, houve aumento da utilização de *lags* com período de 2 meses em relação às previsões de internações. O percentual (11%) de uso de apenas uma *lag* foi o mesmo no caso de internações.

4.2 Ajustes nos hiperparâmetros dos algoritmos

Definidas a melhor combinação de parâmetros previsores, o período de utilização dos dados, se haveria exclusão de *outliers* e a quantidade de *lags* a serem adicionadas, para cada cidade, foram realizados ajustes de parâmetros nas técnicas RF, SVR, MLP, LSTM e CNN. As melhores configurações são demonstradas a seguir.

4.2.1 Internações

A Tabela 11 demonstra a melhor configuração para a técnica *Random Forest.*

Tabela 11 - Melhores configurações por cidade para previsão de internações através da técnica *Random Forest*

Município	*n_estimators*	*min_samples_split*	*min_samples_leaf*
Bayeux	25	10	1
Cabedelo	25	2	1
Cajazeiras	25	5	1
Campina Grande	25	20	5
Catolé do Rocha	50	2	1
João Pessoa	25	10	2
Monteiro	50	2	1
Patos	25	2	1
Santa Rita	50	2	1

Fonte: Autoria própria

As melhores configurações para as técnicas SVR e MLP estão listadas nas tabelas Tabela 12 e Tabela 13.

Tabela 12 - Melhores configurações por cidade para previsão de internações através da técnica *Support Vector Regression*

Município	*Kernel*	*Gamma*	*C*	*epsilon*
Bayeux	*"poly"*	*"scale"*	*1.0*	*0.1*
Cabedelo	*"rbf"*	*"auto"*	*1.0*	*0.1*
Cajazeiras	*"sigmoid"*	*"auto"*	*1.0*	*1.0*
Campina Grande	*"rbf"*	*"auto"*	*10.0*	*0.1*
Catolé do Rocha	*"rbf"*	*"auto"*	*5.0*	*0.1*
João Pessoa	*"rbf"*	*"scale"*	*25.0*	*0.1*
Monteiro	*"poly"*	*"scale"*	*25.0*	*1.0*
Patos	*"rbf"*	*"scale"*	*5.0*	*0.1*
Santa Rita	*"poly"*	*"auto"*	*25.0*	*1.0*

Fonte: Autoria própria

Tabela 13 - Melhores configurações por cidade para previsão de internações através da técnica *Multilayer Perceptron*

Município	*activation*	*solver*	*batch_size*	*learning_rate*	*neurônios_layer1*	*neurônios_layer1*
Bayeux	*"tanh"*	*"adam"*	*24*	*"constant"*	50	25
Cabedelo	*"tanh"*	*"adam"*	*'auto'*	*"adaptive"*	50	25
Cajazeiras	*"relu"*	*"sgd"*	*'auto'*	*"constant"*	250	-
Campina Grande	*"relu"*	*"adam"*	*12*	*"constant"*	25	25
Catolé do Rocha	*"tanh"*	*"sgd"*	*24*	*"adaptive"*	100	25
João Pessoa	*"tanh"*	*"sgd"*	*12*	*"constant"*	100	50
Monteiro	*"relu"*	*"adam"*	*'auto'*	*"adaptive"*	50	25
Patos	*"relu"*	*"sgd"*	*'auto'*	*"constant"*	50	25
Santa Rita	*"tanh"*	*"adam"*	*12*	*"adaptive"*	50	25

Fonte: Autoria própria

Por fim, as configurações para as técnicas LSTM e CNN estão presentes nas tabelas Tabela 14 e Tabela 15.

Tabela 14 - Melhores configurações por cidade para previsão de internações através da técnica *Long short-Term memory*

Município	*neurons*	*activation*	*recurrent_activation*	*dropout*	*batch_size*
Bayeux	50	*"relu"*	*"sigmoid"*	*0.0*	*None*
Cabedelo	100	*"tanh"*	*"relu"*	*0.0*	12
Cajazeiras	250	*"tanh"*	*"tanh"*	*0.2*	12
Campina Grande	25	*"relu"*	*"sigmoid"*	*0.0*	12
Catolé do Rocha	25	*"relu"*	*"sigmoid"*	*0.0*	*None*
João Pessoa	25	*"tanh"*	*"sigmoid"*	*0.0*	12
Monteiro	40	*"tanh"*	*"sigmoid"*	*0.0*	*None*
Patos	25	*"tanh"*	*"sigmoid"*	*0.0*	*None*
Santa Rita	25	*"tanh"*	*"sigmoid"*	*0.0*	*None*

Fonte: Autoria própria

Tabela 15 - Melhores configurações por cidade para previsão de internações através da técnica *Convolutional neural network*

Município	*filters*	*neurons*	*activation*	*recurrent_activation*	*dropout*	*batch_size*
Bayeux	8	25	*"relu"*	*"relu"*	*0.0*	12
Cabedelo	8	250	*"relu"*	*"sigmoid"*	*0.2*	12
Cajazeiras	16	50	*"relu"*	*"sigmoid"*	*0.0*	*None*
Campina Grande	8	50	*"tanh"*	*"sigmoid"*	*0.0*	*None*
Catolé do Rocha	16	50	*"relu"*	*"sigmoid"*	*0.1*	*None*
João Pessoa	16	50	*"tanh"*	*"sigmoid"*	*0.0*	12
Monteiro	16	100	*"relu"*	*"tanh"*	*0.0*	*12*
Patos	16	25	*"tanh"*	*"sigmoid"*	*0.2*	*None*
Santa Rita	16	25	*"tanh"*	*"sigmoid"*	*0.0*	12

Fonte: Autoria própria

4.2.2 Notificações

As melhores configurações para as previsões de notificações através da técnica *Random Forest* estão demonstradas na Tabela 16.

Tabela 16 - Melhores configurações por cidade para previsão de notificações através da técnica *Random Forest*

Município	*n_estimators*	*min_samples_split*	*min_samples_leaf*
Bayeux	50	2	2
Cabedelo	50	10	10
Cajazeiras	50	6	5
Campina Grande	100	20	1
Catolé do Rocha	50	20	1
João Pessoa	25	20	10
Monteiro	50	10	2
Patos	25	10	20
Santa Rita	25	2	1

Fonte: Autoria própria

A Tabela 17 ilustra o resultado dos melhores parâmetros a serem utilizados para a previsão de casos de notificações.

Tabela 17 - Melhores configurações por cidade para previsão de notificações através da técnica *Support Vector Regression*

Município	*Kernel*	*Gamma*	*C*	*epsilon*
Bayeux	*"poly"*	*"auto"*	*10.0*	*0.1*
Cabedelo	*"rbf"*	*"auto"*	*1.0*	*0.1*
Cajazeiras	*"rbf"*	*"auto"*	*1.0*	*1.0*
Campina Grande	*"poly"*	*"auto"*	*1.0*	*0.1*
Catolé do Rocha	*"poly"*	*"scale"*	*1.0*	*0.1*
João Pessoa	*"rbf"*	*"auto"*	*1.0*	*0.1*
Monteiro	*"rbf"*	*"scale"*	*25.0*	*1.1*
Patos	*"rbf"*	*"auto"*	*1.0*	*0.1*
Santa Rita	*"rbf"*	*"scale"*	*10.0*	*1.0*

Fonte: Autoria própria

Em relação aos ajustes do MLP, a Tabela 18 descreve as melhores configurações encontradas.

Tabela 18 - Melhores configurações por cidade para previsão de notificações através da técnica *Multilayer Perceptron*

Município	*activation*	*solver*	*batch_size*	*learning_rate*	*neurônios_layer1*	*neurônios_layer1*
Bayeux	*"relu"*	*"adam"*	*"auto"*	*"constant"*	250	250
Cabedelo	*"logistic"*	*"adam"*	*12*	*"adaptive"*	250	250
Cajazeiras	*"logistic"*	*"sgd"*	*12*	*"constant"*	100	50
Campina Grande	*"relu"*	*"adam"*	*"auto"*	*"adaptive"*	50	25
Catolé do Rocha	*"logistic"*	*"adam"*	*24*	*"adaptive"*	250	25
João Pessoa	*"relu"*	*"sgd"*	*12*	*"adaptive"*	50	25
Monteiro	*"logistic"*	*"adam"*	*12*	*"adaptive"*	250	25
Patos	*"logistic"*	*"adam"*	*12*	*"constant"*	250	250
Santa Rita	*"relu"*	*"adam"*	*12*	*"adaptive"*	250	250

Fonte: Autoria própria

As configurações para as técnicas LSTM e CNN estão presentes, respectivamente, nas tabelas Tabela 19 e Tabela 20.

Tabela 19 - Melhores configurações por cidade para previsão de notificações através da técnica *Long short-Term memory*

Município	*neurons*	*activation*	*recurrent_activation*	*dropout*	*batch_size*
Bayeux	500	*"relu"*	*"sigmoid"*	*0.1*	*12*
Cabedelo	50	*"tanh"*	*"sigmoid"*	*0.0*	12
Cajazeiras	50	*"relu"*	*"sigmoid"*	*0.0*	*None*
Campina Grande	250	*"relu"*	*"sigmoid"*	*0.2*	12
Catolé do Rocha	500	*"relu"*	*"relu"*	*0.0*	*12*
João Pessoa	100	*"tanh"*	*"sigmoid"*	*0.1*	12
Monteiro	250	*"tanh"*	*"sigmoid"*	*0.0*	*None*
Patos	500	*"relu"*	*"tanh"*	*0.0*	12
Santa Rita	25	*"relu"*	*"sigmoid"*	*0.0*	12

Fonte: Autoria própria

Tabela 20 - Melhores configurações por cidade para previsão de notificações através da técnica *Convolutional neural network*

Município	*filters*	*neurons*	*activation*	*recurrent_activation*	*dropout*	*batch_size*
Bayeux	16	25	*"relu"*	*"sigmoid"*	*0.0*	*None*
Cabedelo	2	500	*"relu"*	*"relu"*	*0.2*	12
Cajazeiras	2	500	*"relu"*	*"relu"*	*0.1*	*12*
Campina Grande	8	50	*"relu"*	*"relu"*	*0.2*	*None*
Catolé do Rocha	8	500	*"relu"*	*"sigmoid"*	*0.1*	*None*
João Pessoa	2	250	*"relu"*	*"sigmoid"*	*0.0*	*None*
Monteiro	16	250	*"relu"*	*"sigmoid"*	*0.2*	*12*
Patos	2	50	*"relu"*	*"relu"*	*0.1*	*None*
Santa Rita	16	50	*"relu"*	*"tanh"*	*0.2*	12

Fonte: Autoria própria

4.3 Geração das previsões e determinação da melhor técnica por cidade

Feitos os ajustes de parâmetros e por questões estocásticas de alguns algoritmos, foram realizados quatro *rounds* de execuções. Para cada cidade e técnica, foram feitas 100 execuções em cada *round* e os resultados foram utilizados na definição da melhor técnica e na análise de diferença estatística. Por fim, as previsões finais para cada cidade foram geradas nessa etapa.

4.3.1 Internações

A Tabela 21 demonstra a menor taxa de erro por cidade durante a previsão de casos de internações e qual técnica foi a responsável por isso.

Tabela 21 - Melhores resultados e técnica vencedora para a previsões de internações

Município	Menor RMSE	Técnica vencedora
Bayeux	0,529017139	LSTM
Cabedelo	0,927428107	LSTM
Cajazeiras	1,000751246	LSTM
Campina Grande	2,193690519	MLP
Catolé do Rocha	0,365512405	MLP
João Pessoa	9,552880644	CNN
Monteiro	1,02320815	RF
Patos	0,422049567	CNN
Santa Rita	0,745519953	RF

Fonte: Autoria própria

Ao analisar os resultados, fica demonstrado que a técnica LSTM produziu os melhores resultados em três das nove cidades, representando um percentual de 33,33%. Na sequência, a técnica CNN foi vencedora em 22,22% dos casos. A MLP obteve menor taxa de erro para as cidades de Campina Grande e Catolé do Rocha, representando 22,22%. Por fim, RF também conseguiu prever internações com menor erro para Monteiro e Santa Rita (22% das cidades estudadas). O SVR não obteve a menor taxa de erro em nenhum município.

Os números demonstram a superioridade das técnicas de *Deep Learning* em comparação às técnicas de *Machine Learning*. LSTM, CNN e MPL conseguiram produzir as melhores previsões em 7 das 9 cidades estudadas.

4.3.2 Notificações

Os resultados das previsões de notificações e as técnicas com os melhores resultados por cidade são demonstrados através da Tabela 22.

Tabela 22 - Melhores resultados e técnica vencedora para a previsões de notificações

Município	Menor RMSE	Técnica vencedora
Bayeux	2,110531124	CNN
Cabedelo	5,939260648	LSTM
Cajazeiras	11,92052734	LSTM
Campina Grande	8,003158932	LSTM
Catolé do Rocha	1,541425279	LSTM
João Pessoa	168,4912073	CNN
Monteiro	10,56430032	LSTM
Patos	2,869606677	LSTM
Santa Rita	13,40766199	MLP

Fonte: Autoria própria

Durante a previsão de notificações, a técnica LSTM obteve os melhores resultados em seis das noves cidades estudadas. A técnica combinada CNN demonstrou a menor taxa de erro em duas cidades e, enfim, MLP performou melhor em uma cidade.

Para as previsões de casos de notificações, as técnicas de *Deep Learning* demonstraram superioridades em todos os municípios abordados.

4.4 Avaliação dos resultados produzidos

Não foram encontrados na literatura trabalhos com previsões de internações e de notificações de casos de dengue para as cidades aqui trabalhadas. De acordo com Shmueli e Lichtendahl jr (2016), uma abordagem para comparar os resultados de uma pesquisa em cenários como esse é verificar se o erro produzido é melhor que o gerado pela técnica *Naive Forecast*. Adicionalmente, os resultados foram comparados com a técnica estatística ARIMA.

Para comprovar se há diferença estatística entre os RMSEs produzidos pelas técnicas para cada cidade e, separadamente, para internações e para notificações, foram realizados os testes ANOVA e Tukey com $\alpha = 0,05$. As seguintes hipóteses estatísticas foram consideradas:

H_0: Estatisticamente os resultados produzidos são iguais (se *p-value* > 0,05) e

H_1: Estatisticamente há diferença entre os resultados (se *p-vaule* < 0,05)

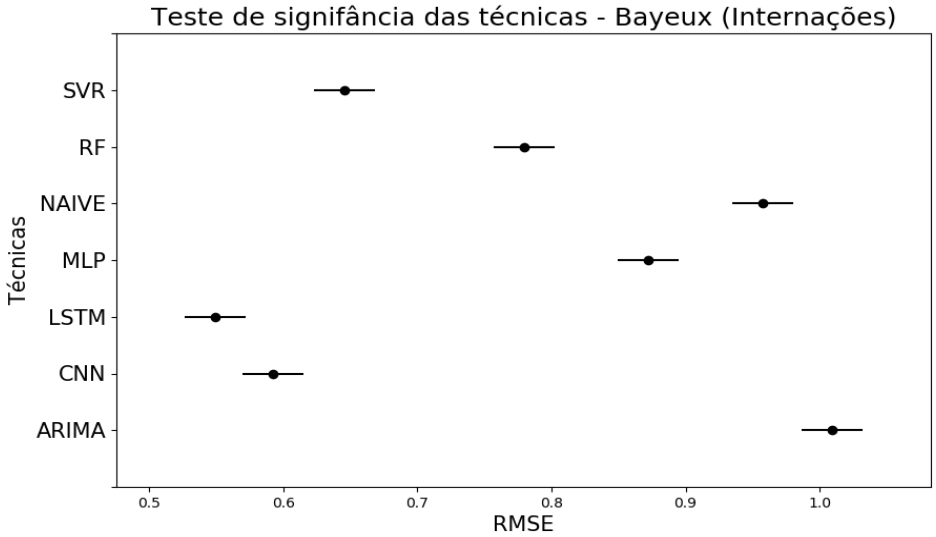
Adicionalmente, o teste ANOVA e o teste Tukey possuem como pressuposto a adequação dos resultados à curva normal. Sendo assim, foi realizado o teste de normalidade Shapiro-Wilk com α = 0,05 antes da execução do ANOVA e Tukey.

Todos os resultados obtidos para previsões de internações e de notificações seguiram a curva normal e o teste ANOVA demonstrou existir diferença entre os resultados. As saídas dos testes de Tukey são demonstrados na sequência. Por fim, além dos testes supracitados e com intuito de analisar visualmente os resultados, foram gerados gráficos com as previsões de internações e notificações.

4.4.1 Internações

O Gráfico 1 ilustra a saída dos testes de Tukey para a cidade de Bayeux. Como pode ser observado, LSTM obteve a menor taxa de erro e, ao analisar apenas o RMSE, ela pode ser indicada como a melhor técnica de previsão. Contudo, estatisticamente, não há diferença estatística entre os resultados produzidos por LSTM e CNN, pois o *p-value* do teste foi de 0,0647 (> 0,05).

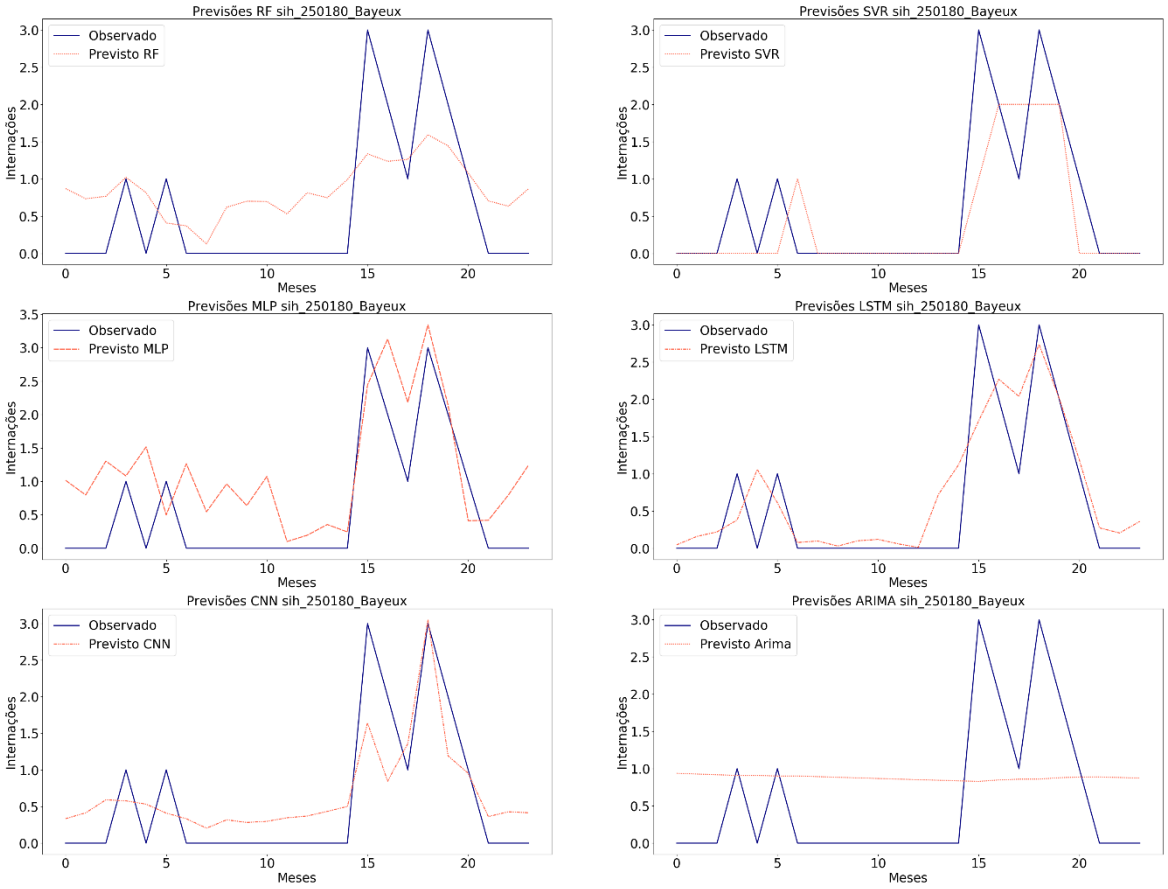
Gráfico 1 - Análise de significância estatística para as previsões de internações para a cidade de Bayeux



Fonte: Autoria própria

As previsões de internações para Bayeux estão presentes no Gráfico 2. Como o período de utilização de dados escolhido foi entre 2010 e 2019 e, respeitando o percentual de 20% dos dados para testes, as previsões foram feitas para dois anos.

Gráfico 2 - Previsões de internações por técnica para a cidade de Bayeux

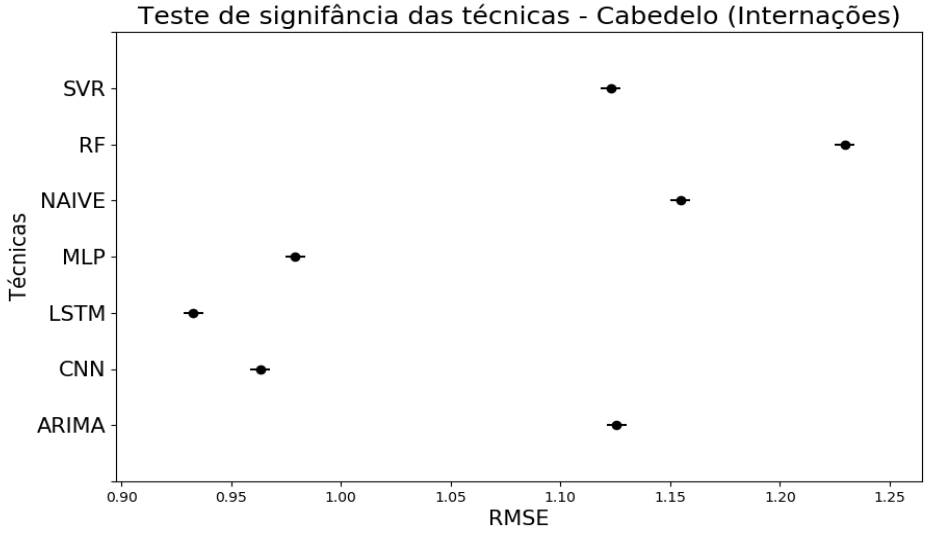


Fonte: Autoria própria

Analisando os gráficos, podemos observar a melhor adequação das previsões fornecidas pela LSTM para a linha de internações observadas. Adicionalmente, conforme demonstrado nos testes estáticos, os resultados produzidos pela CNN são bem similares aos produzidos por LSTM e são refletidos em suas linhas de resultado.

Para a cidade de Cabedelo a técnica LSTM obteve a menor taxa de erro e, de acordo com o teste de Tukey, presentes no Gráfico 3, há diferença entre os seus resultados e os resultados de todas as demais técnicas.

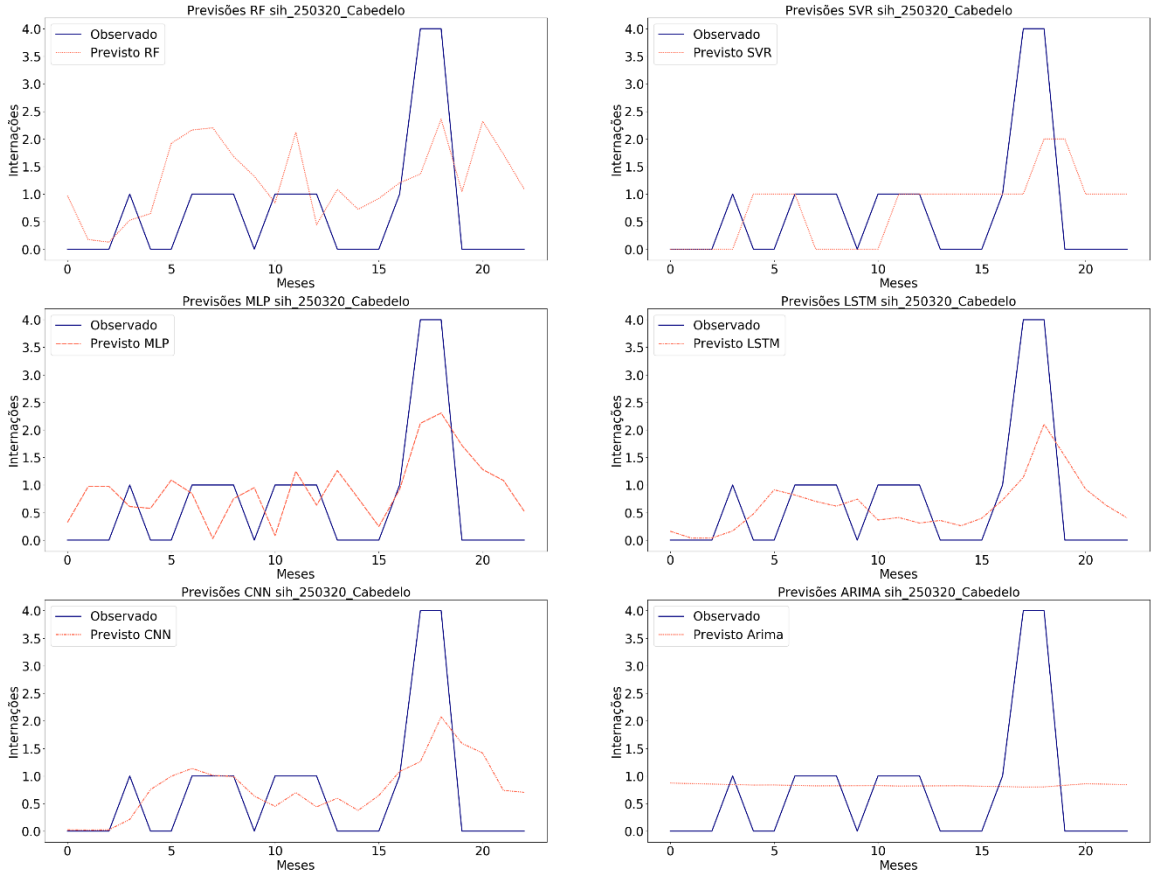
Gráfico 3 - Análise de significância estatística para as previsões de internações para a cidade de Cabedelo



Fonte: Autoria própria

Similarmente a Bayeux, foram realizados dois anos de previsões de internações para Cabedelo. Os números estão demonstrados no Gráfico 4.

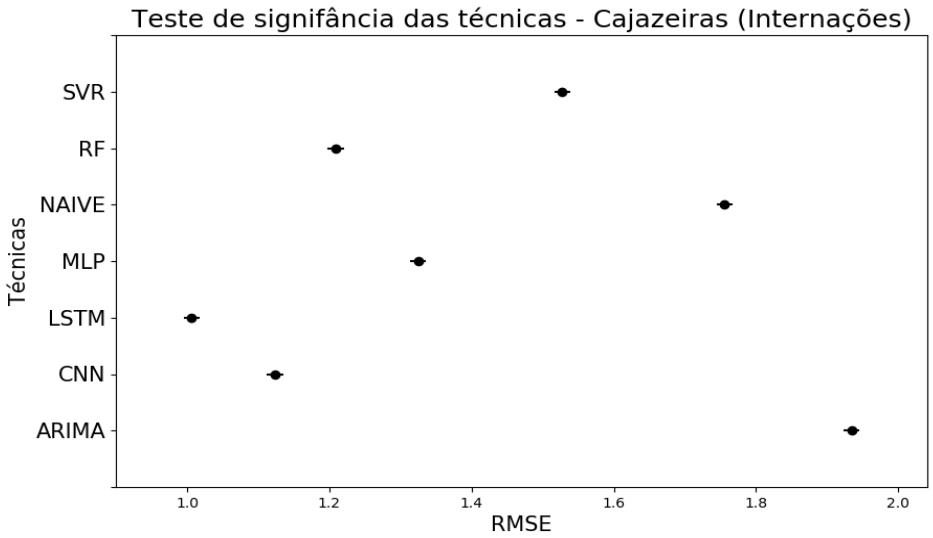
Gráfico 4 - Previsões de internações por técnica para a cidade de Cabedelo



Fonte: Autoria própria

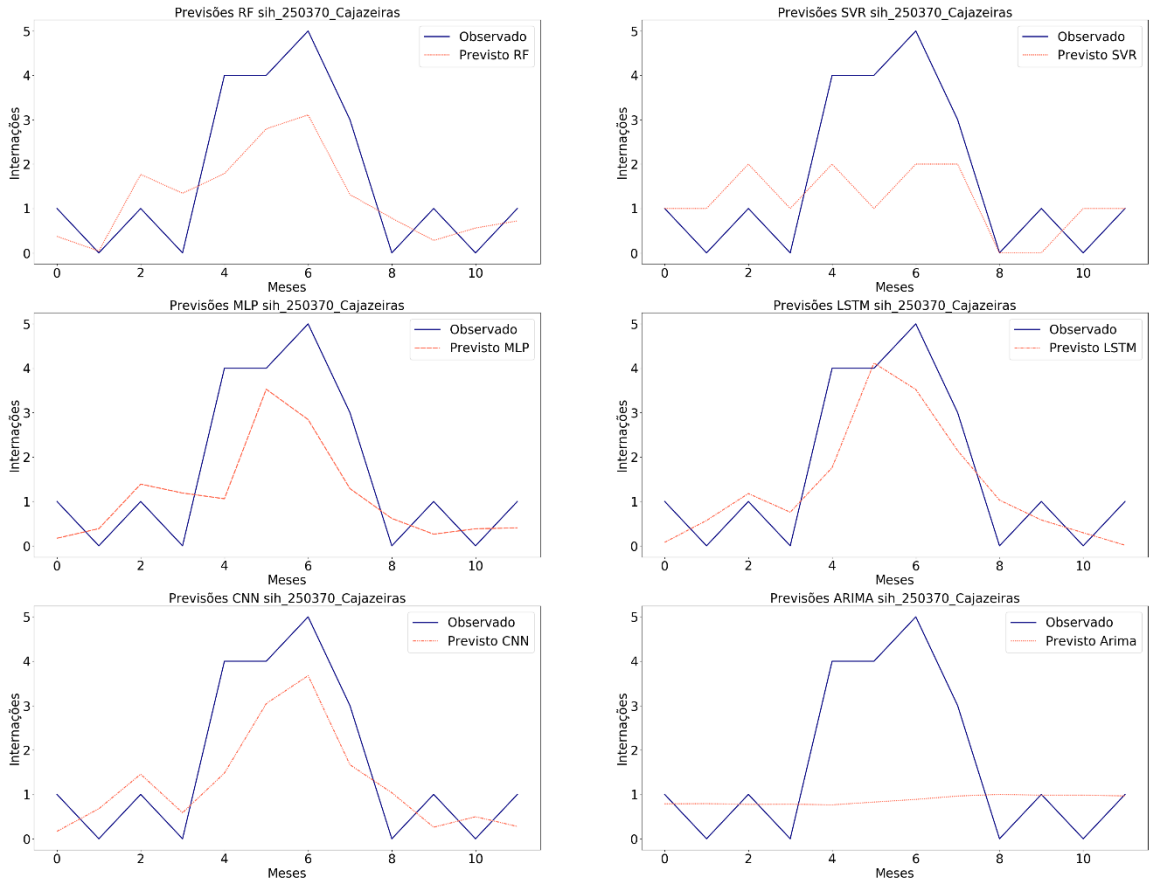
Os gráficos Gráfico 5 e Gráfico 6, demonstram, respectivamente, os resultados dos testes de Tukey e as previsões geradas para a cidade de Cajazeiras.

Gráfico 5 - Análise de significância estatística para as previsões de internações para a cidade de Cajazeiras



Fonte: Autoria própria

Gráfico 6 - Previsões de internações por técnica para a cidade de Cajazeiras

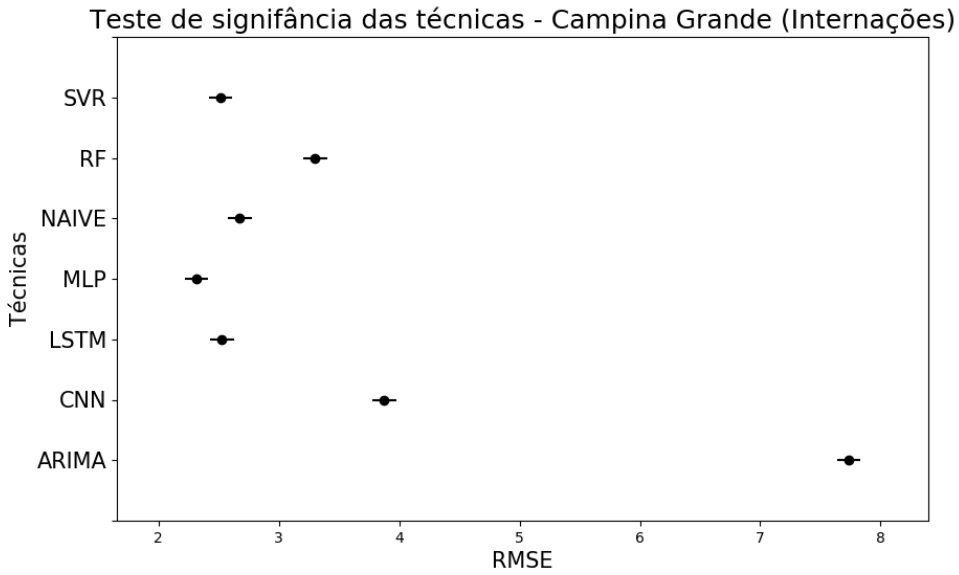


Fonte: Autoria própria

Como observado, LSTM obteve os melhores resultados e todos são diferentes segundo o teste de Tukey. Adicionalmente, a técnica LSTM obteve a melhor adequação à curva de internações observadas. Para Cajazeiras, seguindo a regra dos 20% dos dados para teste, foram realizadas previsões para 12 meses.

Os testes de Tukey referentes à cidade de Campina Grande estão presentes no Gráfico 7. A técnica MLP obteve a menor taxa de erro e seus resultados são, significante diferentes da SVR e da LSTM, segunda e terceiras colocadas, já que o *p-value* da comparação de MLP com SVR foi 0,0455, enquanto o *p-value* de MLP com LSTM foi 0,032.

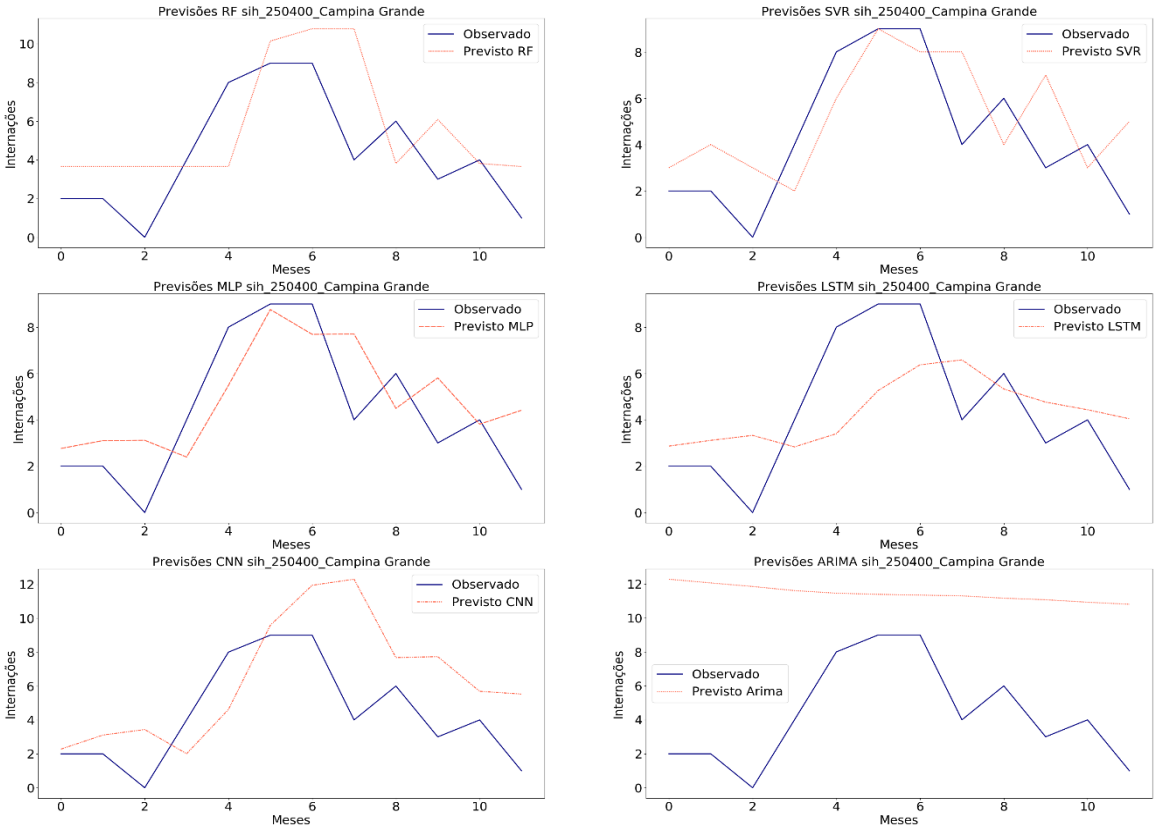
Gráfico 7 - Análise de significância estatística para as previsões de internações para a cidade de Campina Grande



Fonte: Autoria própria

As previsões de internações em Campina Grande foram feitas com um ano. A técnica *Multilayer perceptron* conseguiu gerar resultados de internações bem similares aos valores observados nessa cidade. Embora SVR não tenha produzido menores resultados para nenhuma cidade deste estudo, com a segunda menor taxa de erro para Campina Grande, SVR obteve a sua melhor colocação de previsão durante o estudo das internações.

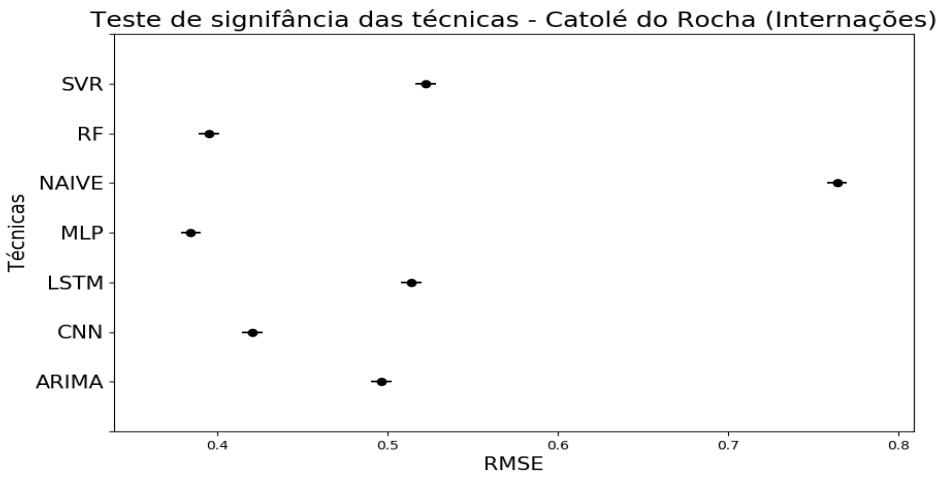
Gráfico 8 - Previsões de internações por técnica para a cidade de Campina Grande



Fonte: Autoria própria

Os resultados do teste de significância para a cidade de Catolé do Rocha estão evidenciados no Gráfico 9.

Gráfico 9 - Análise de significância estatística para as previsões de internações para a cidade de Catolé do Rocha

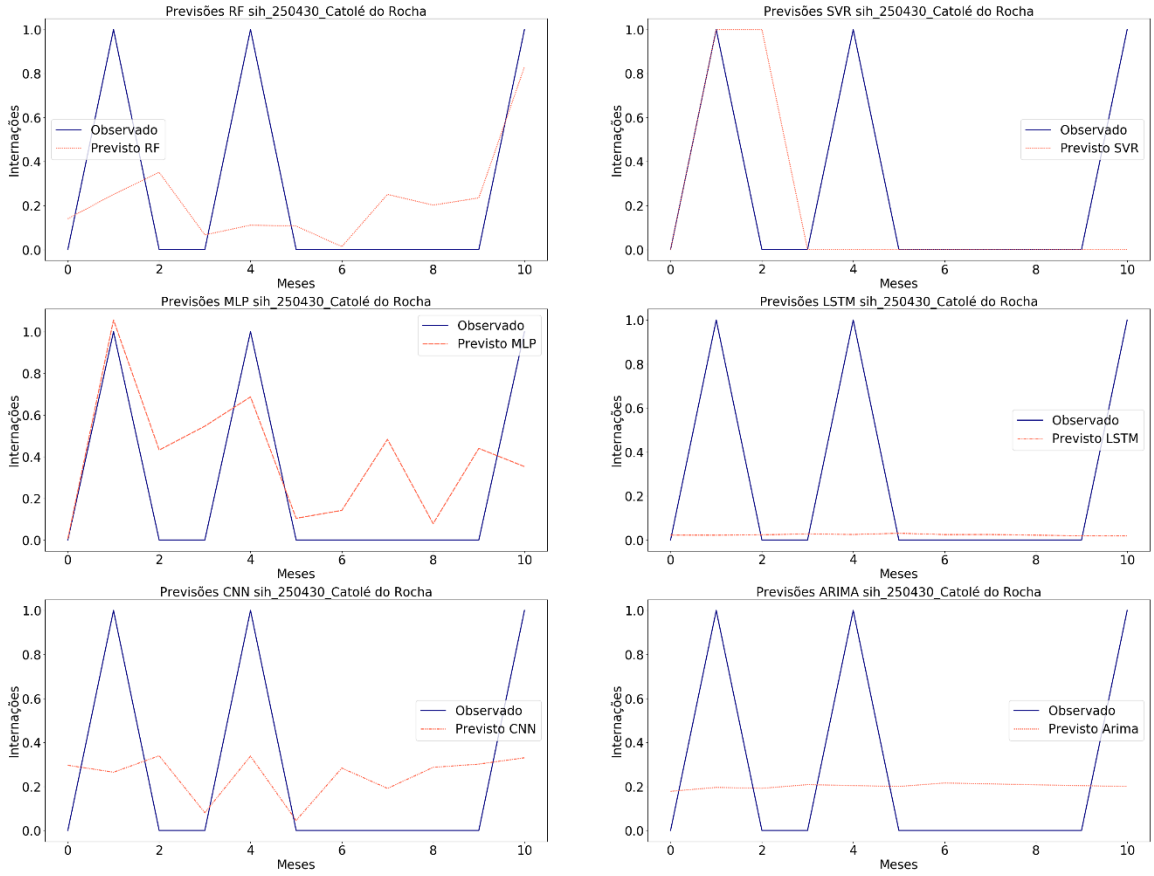


Fonte: Autoria própria

A técnica MLP obteve o menor RMSE. Entretanto, o resultado do teste de Tukey ao comparar MLP com RF retornou um *p-value* 0,097. Logo, não há comprovação estatística de diferenças entre essas técnicas.

O Gráfico 10 contém as previsões geradas para Catolé do Rocha. Conforme notado, MLP conseguiu fazer as melhores previsões nos 6 primeiros meses e RF performou melhor nos últimos 6. Esse fato pode justificar a similaridade do teste de Tukey para essas duas técnicas.

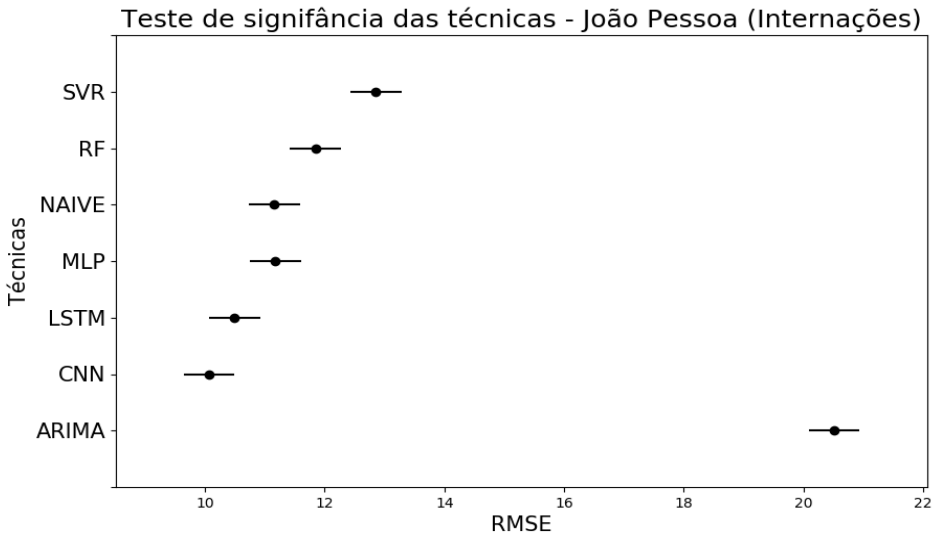
Gráfico 10 - Previsões de internações por técnica para a cidade de Catolé do Rocha



Fonte: Autoria própria

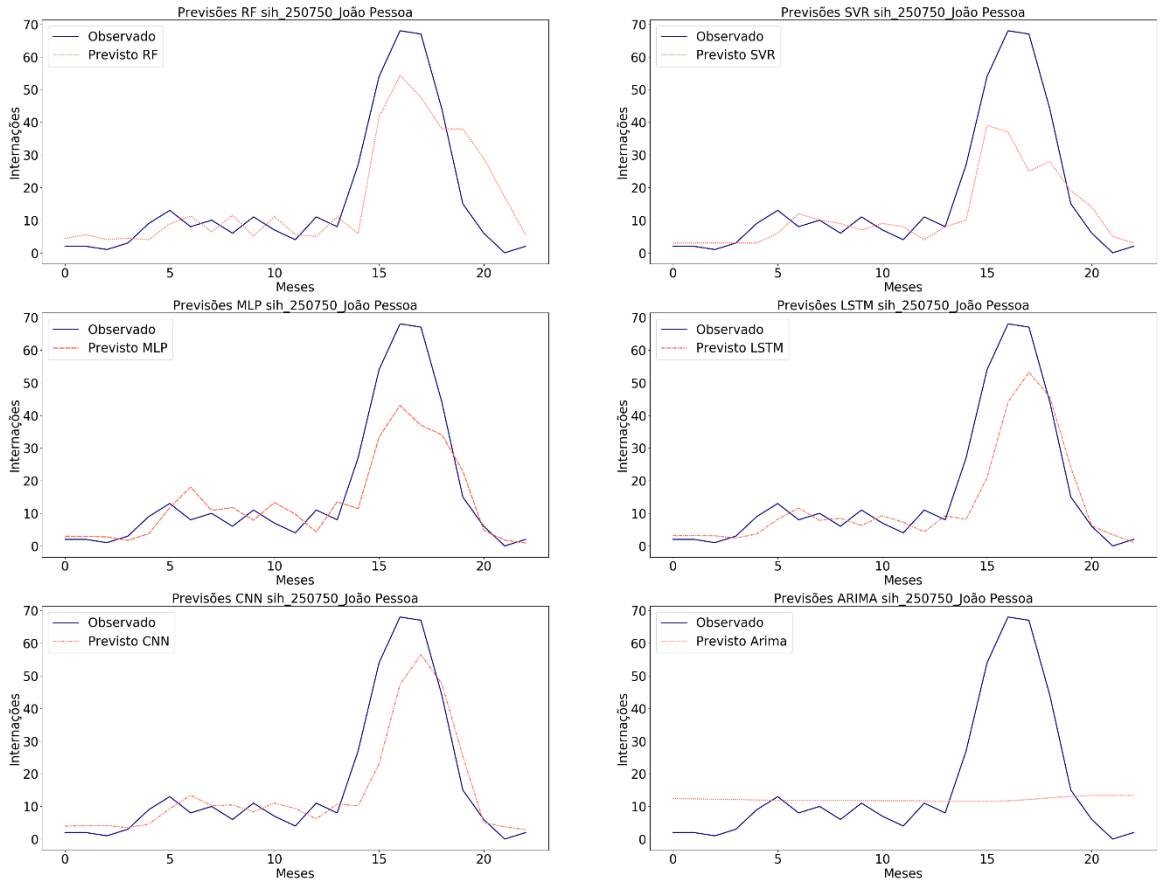
Os testes de Tukey para João Pessoa estão ilustrados no Gráfico 11. CNN foi a técnica com menor resultado, seguido por LSTM. Ao verificar o *p-value* da comparação par a par entre elas, o resultado obtido foi de 0,6386. Portanto, não há argumento estatístico para afirmar a diferença entre CNN e LSTM. Esse fato também pode ser observado ao verificar a similaridade entre as curvas de previsões entre CNN e LSTM presentes no Gráfico 12.

Gráfico 11 - Análise de significância estatística para as previsões de internações para a cidade de João Pessoa



Fonte: Autoria própria

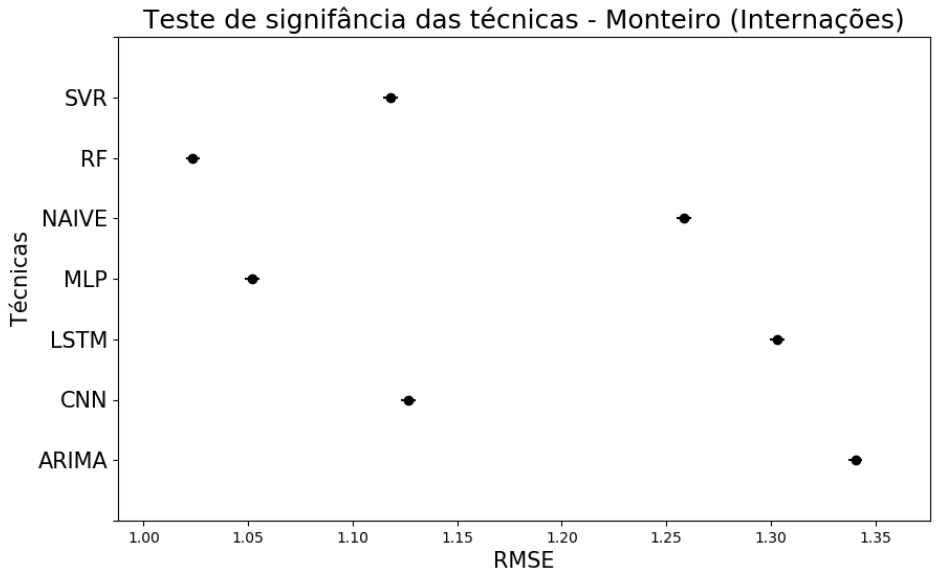
Gráfico 12 - Previsões de internações por técnica para a cidade de João Pessoa



Fonte: Autoria própria

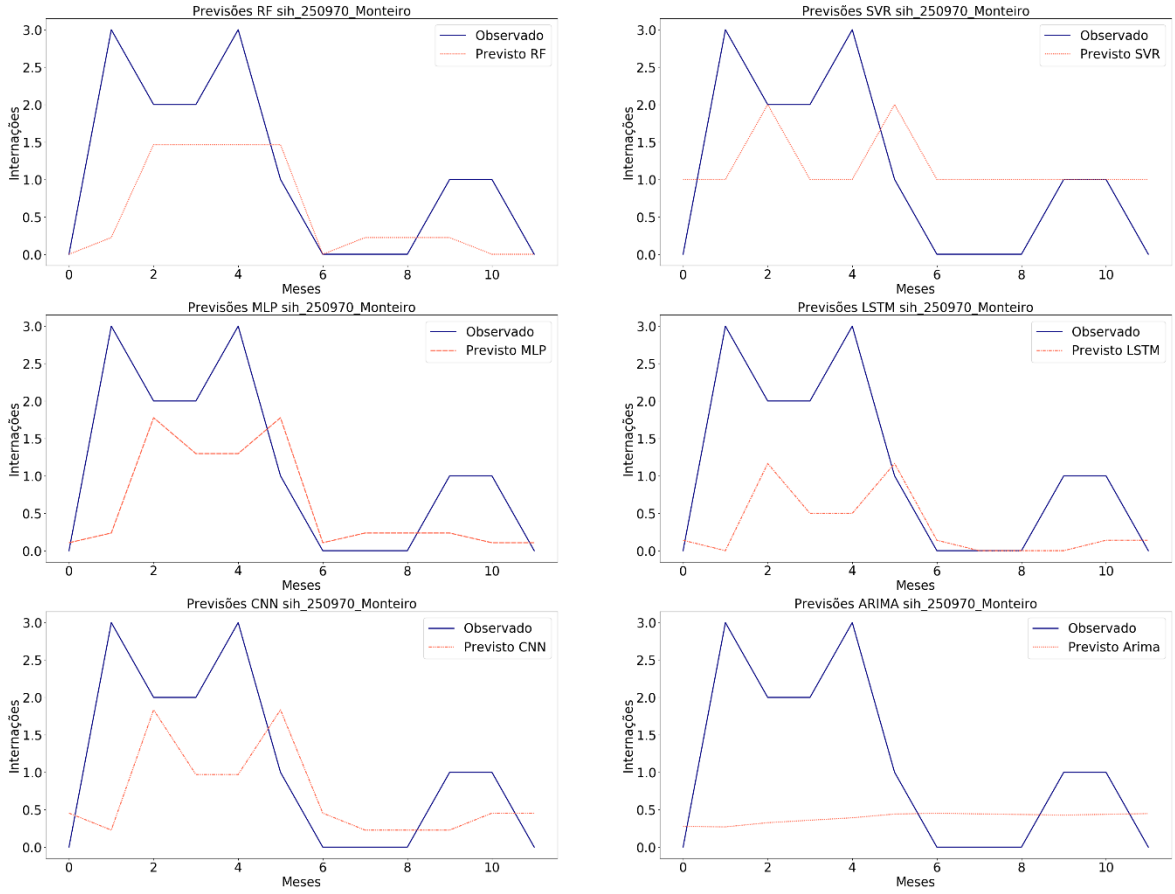
Como evidenciado no Gráfico 13, em relação à cidade de Monteiro, os testes estatísticos relatam a superioridade da técnica *Random Forest* e a diferença estatística em relação às demais técnicas.

Gráfico 13 - Análise de significância estatística para as previsões de internações para a cidade de Monteiro



Fonte: Autoria própria

Gráfico 14 - Previsões de internações por técnica para a cidade de Monteiro

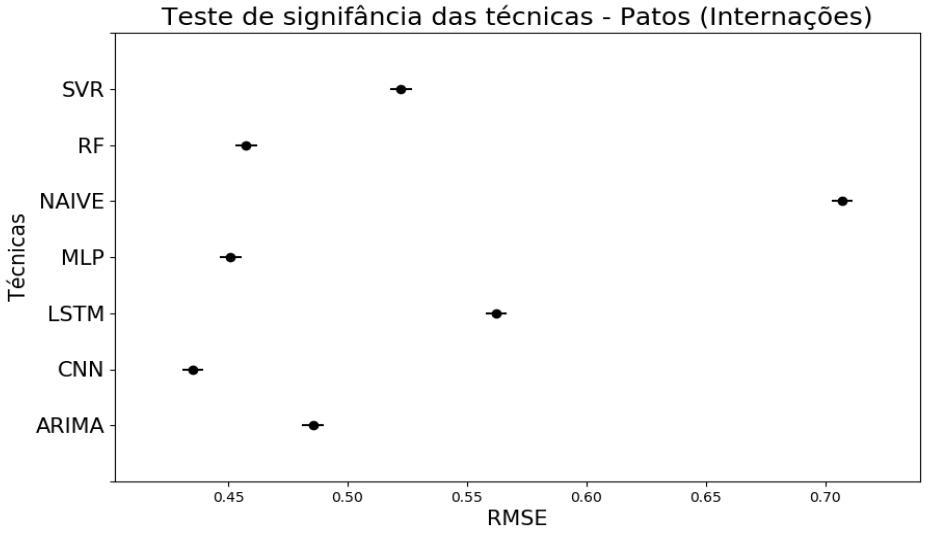


Fonte: Autoria própria

As previsões de internações para Monteiro estão relatadas no Gráfico 14. A técnica RF, seguida por MLP, foram as que melhor seguiram a linha de internações observadas. As técnicas de *Deep Learning* LSTM e CNN, não obtiveram bons resultados durante as previsões de Monteiro. Uma possível justificativa para isso pode ser a quantidade de meses com observações de zero notificações de casos de dengue para essa cidade.

O Gráfico 15 ilustra os resultados dos testes de Tukey para a cidade de Patos. A técnica CNN obteve a menor taxa de erro e ficou evidenciada a sua diferença estatística após as comparações com as outras técnicas.

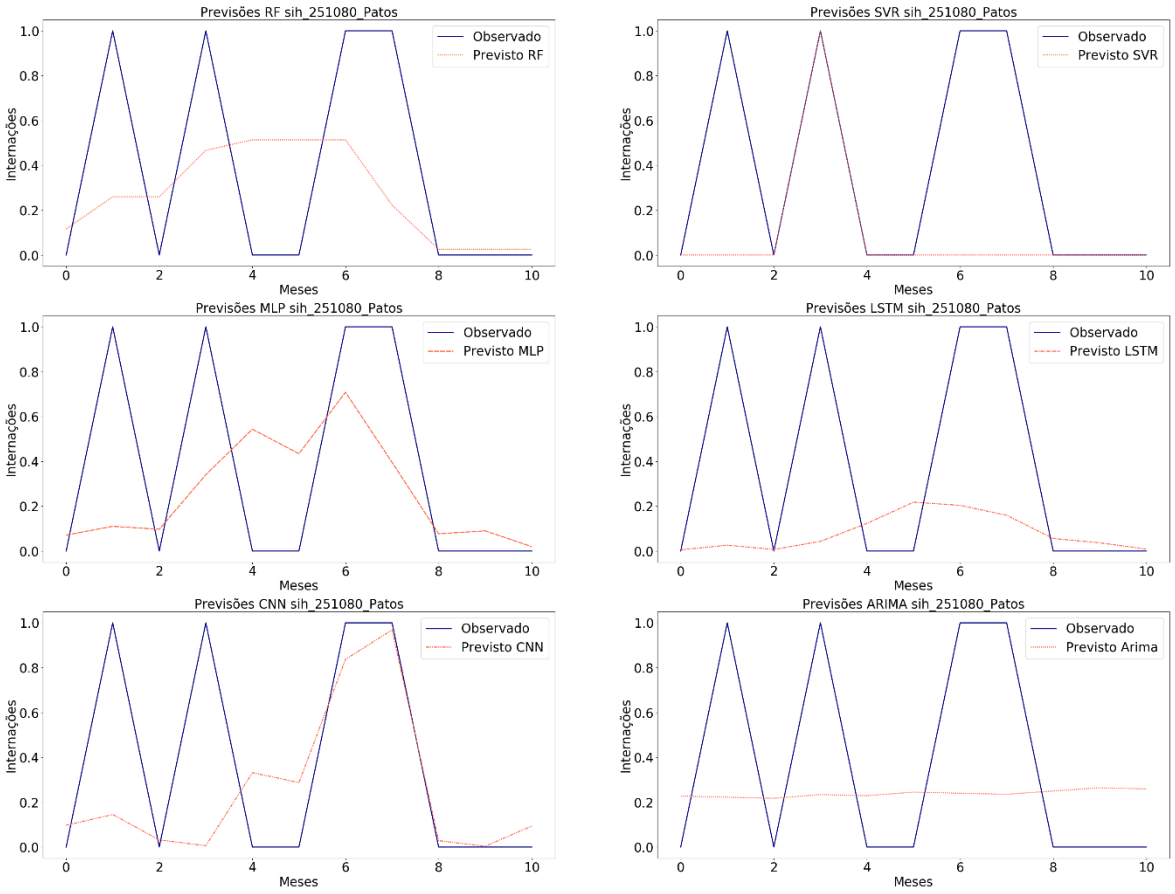
Gráfico 15 - Análise de significância estatística para as previsões de internações para a cidade de Patos



Fonte: Autoria própria

As previsões para Patos estão presentes no Gráfico 16. O principal achado dessas previsões 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 para Partos. Contudo, além de conseguir acompanhar a curva, a técnica CNN obteve a menor taxa de erro para essa cidade.

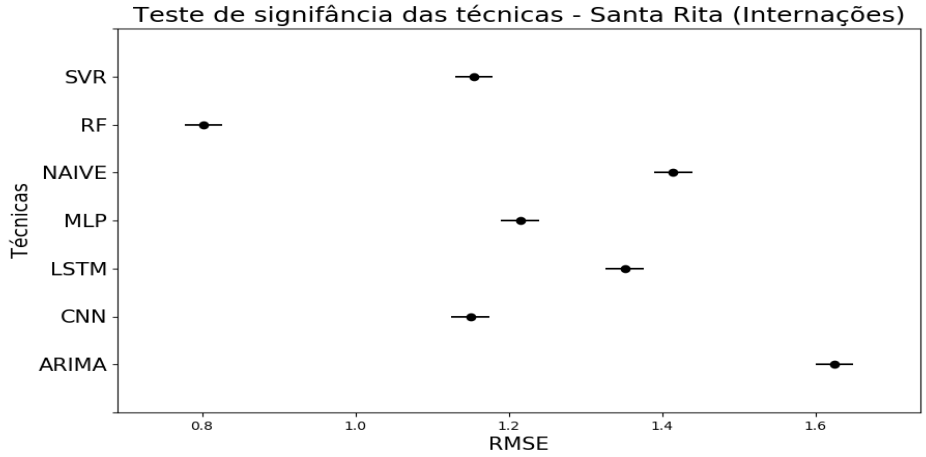
Gráfico 16 - Previsões de internações por técnica para a cidade de Patos



Fonte: Autoria própria

Finalizando as análises das previsões de internações, o Gráfico 17 evidencia os testes de Tukey para Santa Rita.

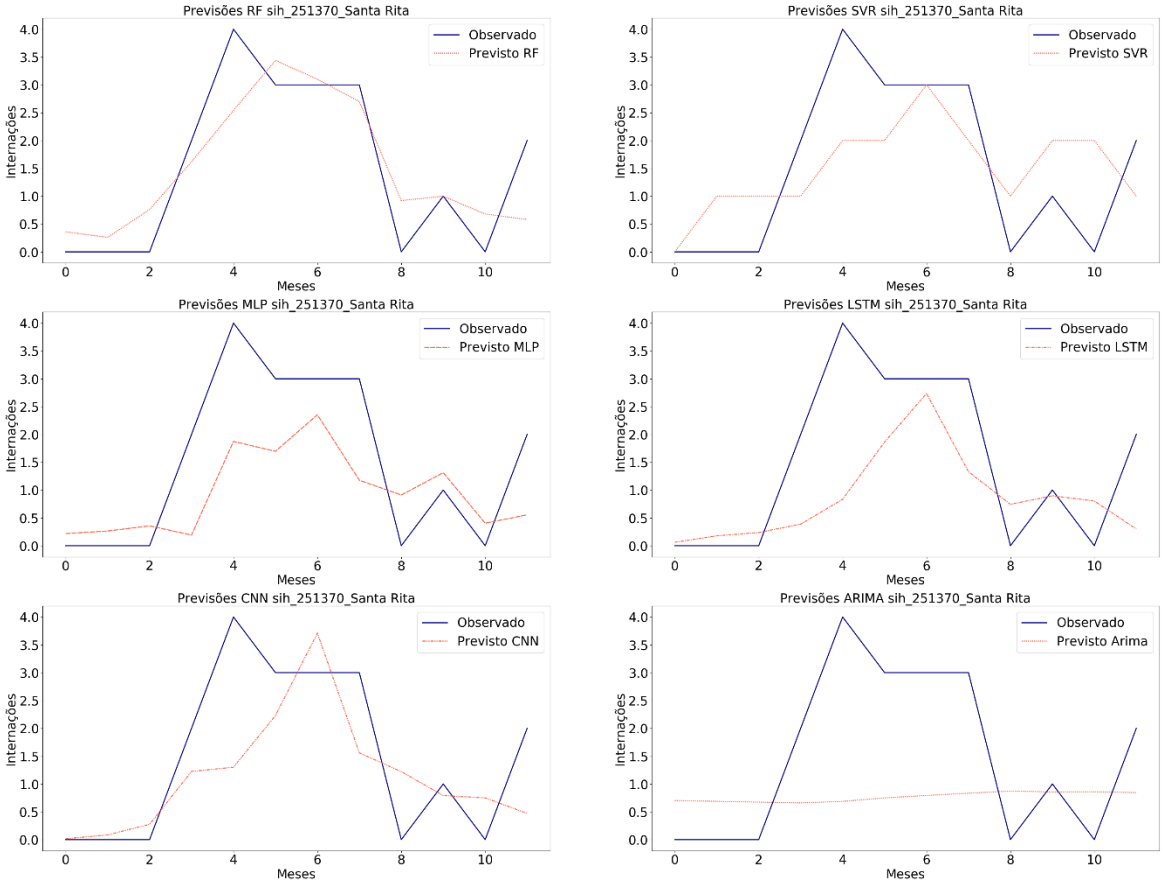
Gráfico 17 - Análise de significância estatística para as previsões de internações para a cidade de Santa Rita



Fonte: Autoria própria

A técnica RF obteve a menor taxa de erro RMSE e ficou comprovada a sua diferença estatística em relação às demais técnicas. As previsões para Santa Rita foram feitas no período de um ano e, conforme observado no Gráfico 18, a técnica vencedora, RF, seguiu a tendência de crescimento e de decréscimo dos casos de internações. Ademais, a segunda melhor técnica foi SVR. Com isso, para Santa Rita, as técnicas de ML performaram melhor que as técnicas de DL.

Gráfico 18 - Previsões de internações por técnica para a cidade de Santa Rita

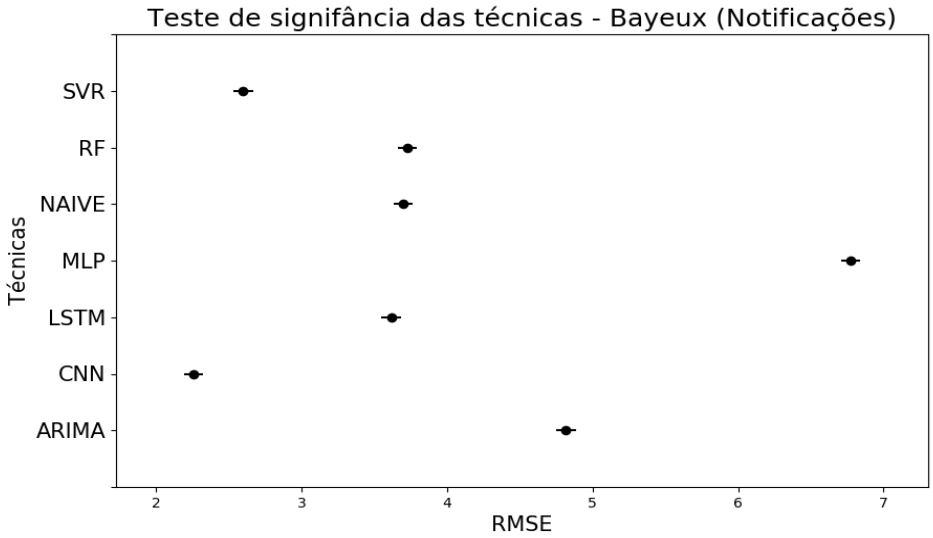


Fonte: Autoria própria

4.4.2 Notificações

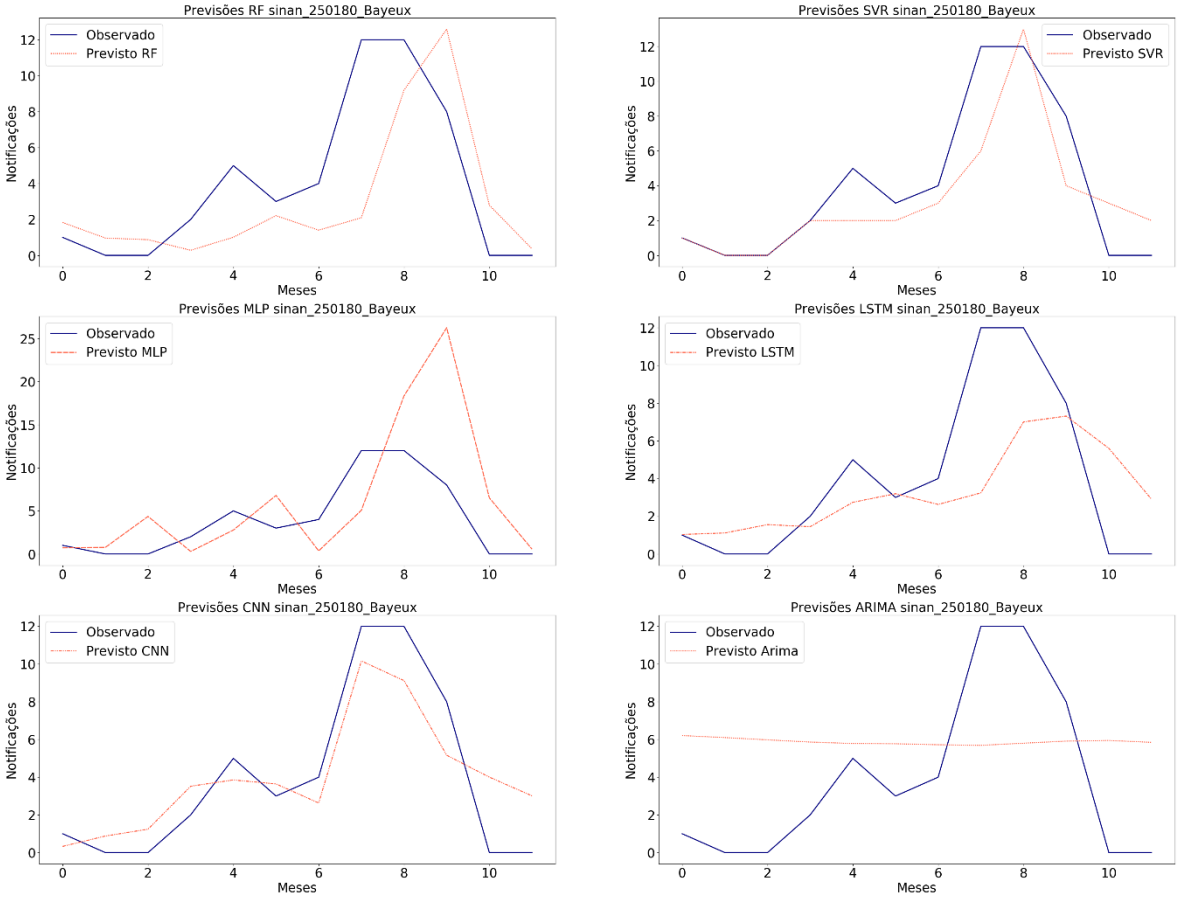
A análise da significância estatística e das previsões de notificações para a cidade de Bayeux estão demonstrados nos Gráfico 19 e Gráfico 20. Examinando os testes de Tukey, fica evidente a diferença estatística entre a técnica com menor taxa de erro (CNN) e as demais. Em relação às curvas de previsões, a CNN demonstra melhor adequação às observações de notificações nesta cidade.

Gráfico 19 - Análise de significância estatística para as previsões de notificações para a cidade de Bayeux



Fonte: Autoria própria

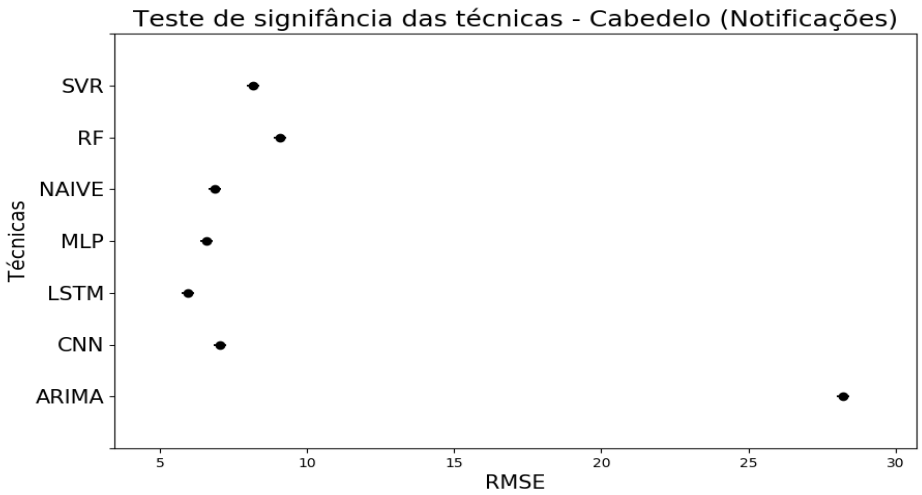
Gráfico 20 - Previsões de notificações por técnica para a cidade de Bayeux



Fonte: Autoria própria

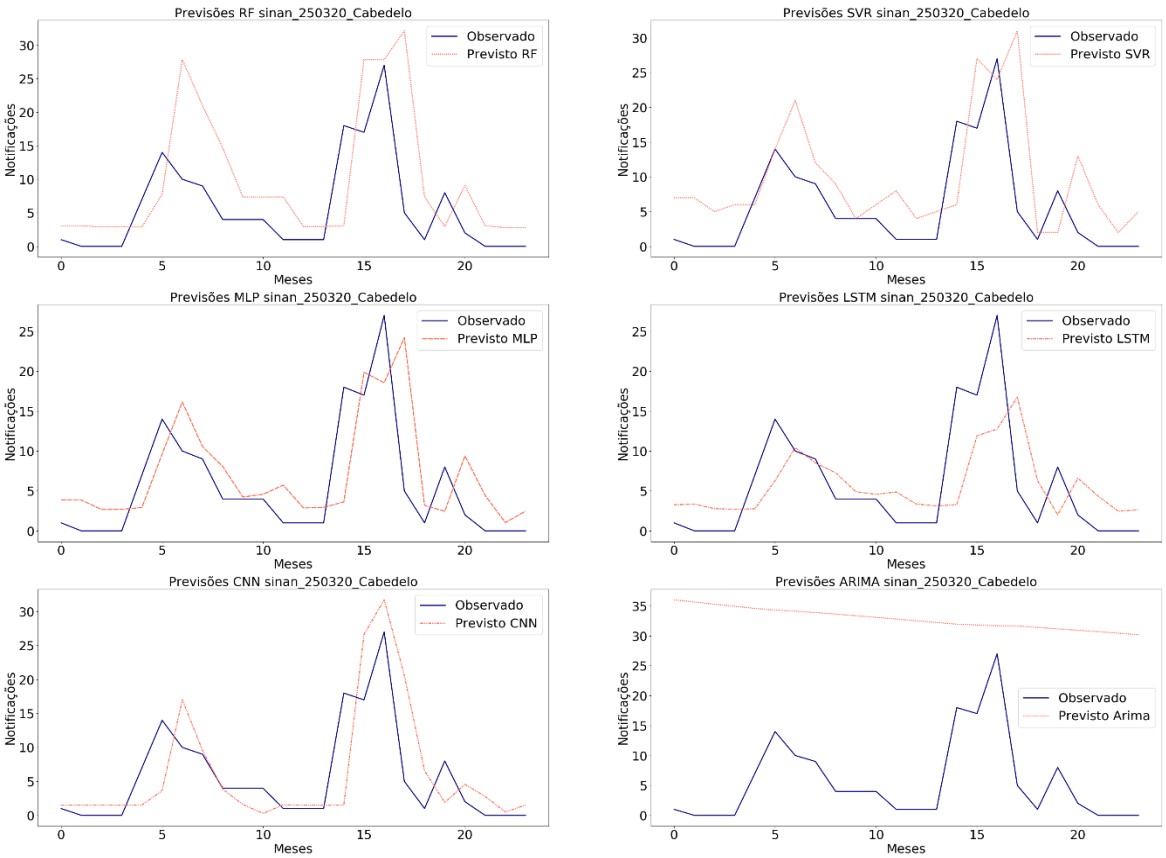
Os gráficos Gráfico 21 e Gráfico 22 ilustram, respectivamente, os testes de Tukey e as curvas de previsões para a cidade de Cabedelo.

Gráfico 21 - Análise de significância estatística para as previsões de notificações para a cidade de Cabedelo



Fonte: Autoria própria

Gráfico 22 - Previsões de notificações por técnica para a cidade de Cabedelo

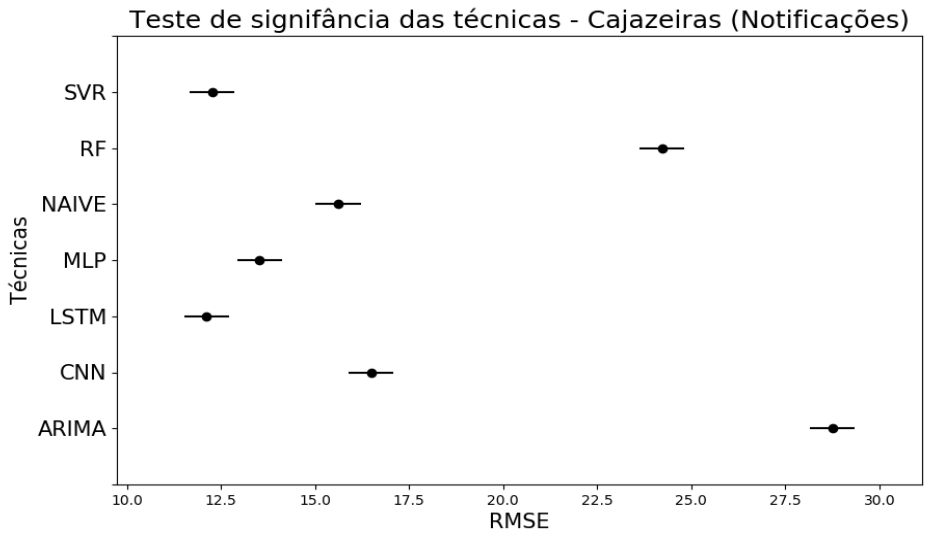


Fonte: Autoria própria

Como notado, existe diferença estatística entre os resultados produzidos pela técnica LSTM e as demais. Sobre as previsões, visualmente, é comprovada a melhor adequação das notificações previstas pela LSTM durante o período de testes realizados (24 meses).

Em relação à Cajazeiras, o teste estatístico produziu um *p-value* de 0,9 ao comparar as melhores técnicas para previsões de notificações para essa cidade: LSTM e SVR. Logo, estatisticamente, não podemos afirmar que existe diferença entre os resultados produzidos por elas. A representação visual dos testes de Tukey está presente no Gráfico 23.

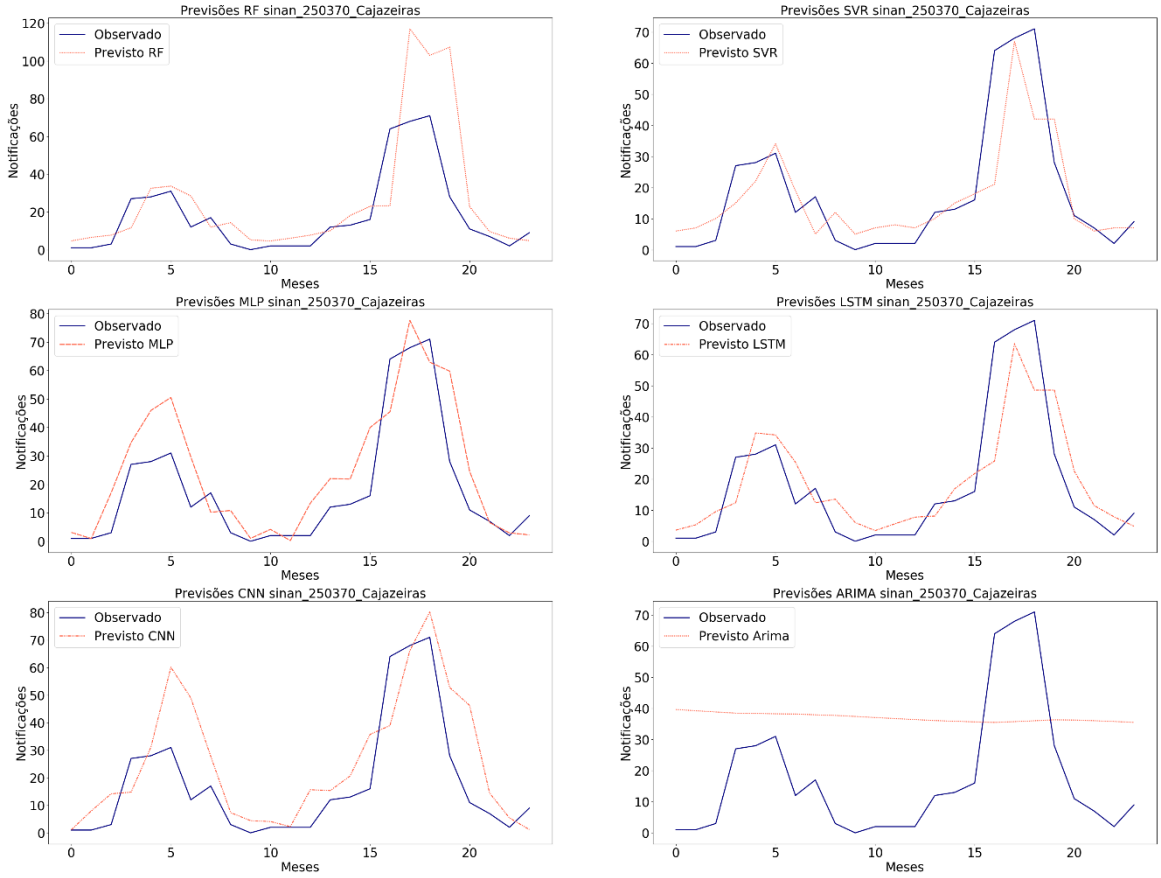
Gráfico 23 - Análise de significância estatística para as previsões de notificações para a cidade de Cajazeiras



Fonte: Autoria própria

As curvas de notificações produzidas para a cidade de Cajazeiras, vão ao encontro dos resultados produzidos pelo teste de Tukey. As previsões da LSTM e SVR, representadas no Gráfico 24, são bem similares. Contudo, LSTM demonstra uma leve melhor adequação.

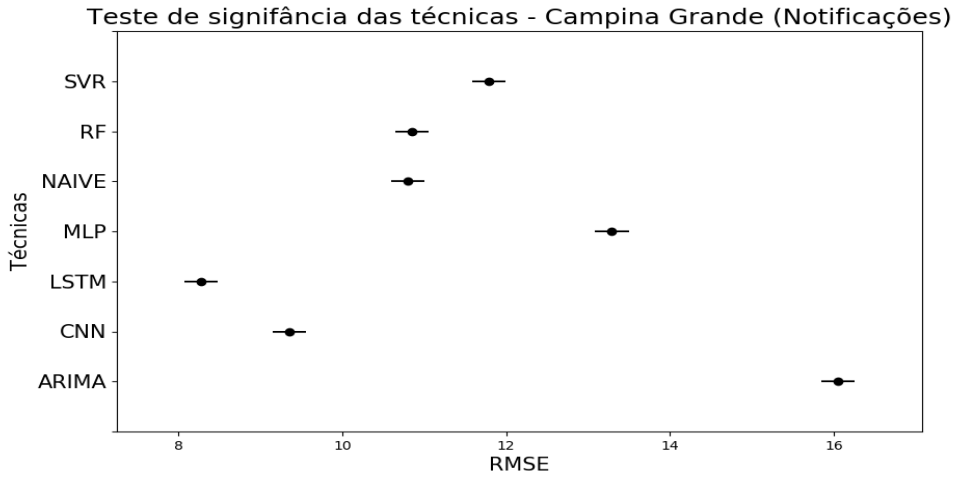
Gráfico 24 - Previsões de notificações por técnica para a cidade de Cajazeiras



Fonte: Autoria própria

O Gráfico 25 relata os resultados dos testes estatísticos para a cidade de Campina Grande.

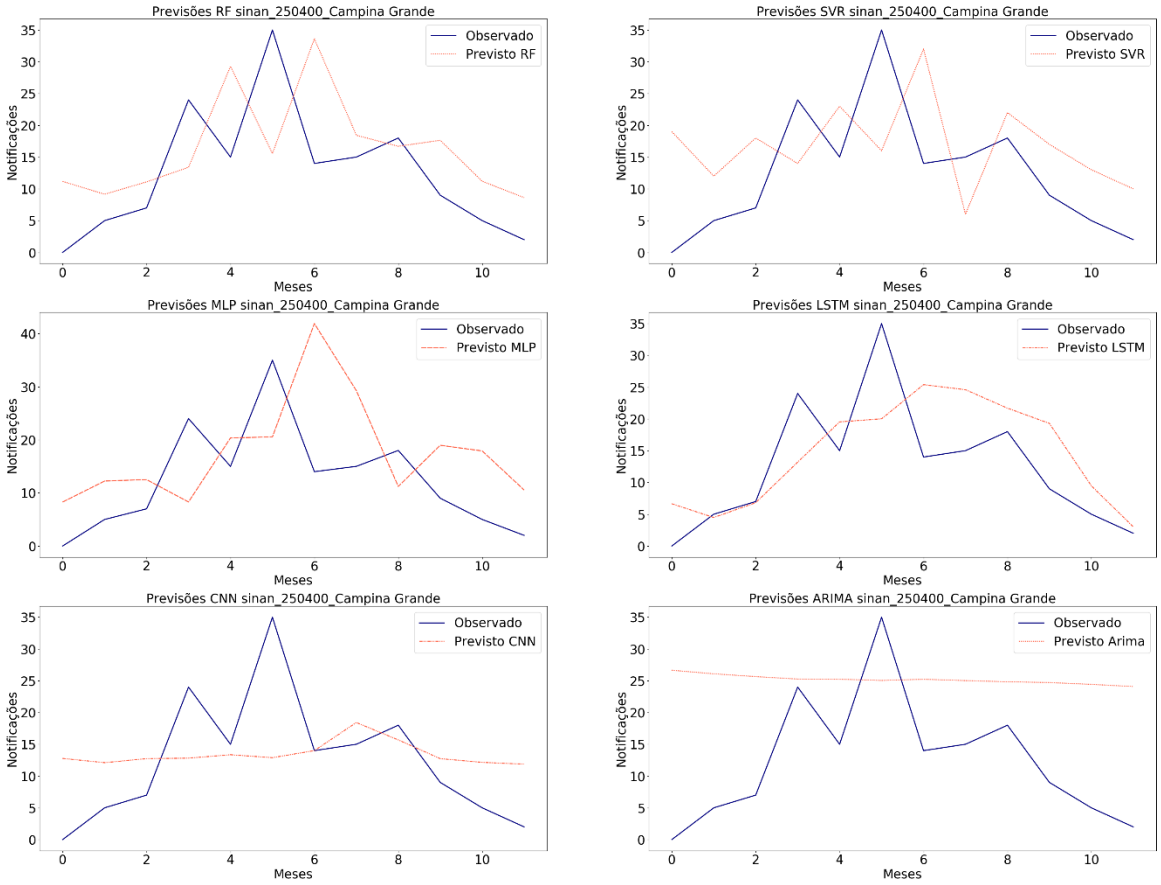
Gráfico 25 - Análise de significância estatística para as previsões de notificações para a cidade de Campina Grande



Fonte: Autoria própria

Conforme verificado, ficou evidenciado a diferença estatística entre as duas técnicas com menores taxas de erro: LSTM e CNN. Adicionalmente, os testes também evidenciam a diferença da menor técnica, LSTM, em relação às demais. Sobre as curvas de previsões, embora as linhas não estejam tão fidedignas como em outras cidades, as previsões feitas por LSTM conseguiram acompanhar a tendência de crescimento e diminuição dos casos. Os resultados podem ser observados no Gráfico 26.

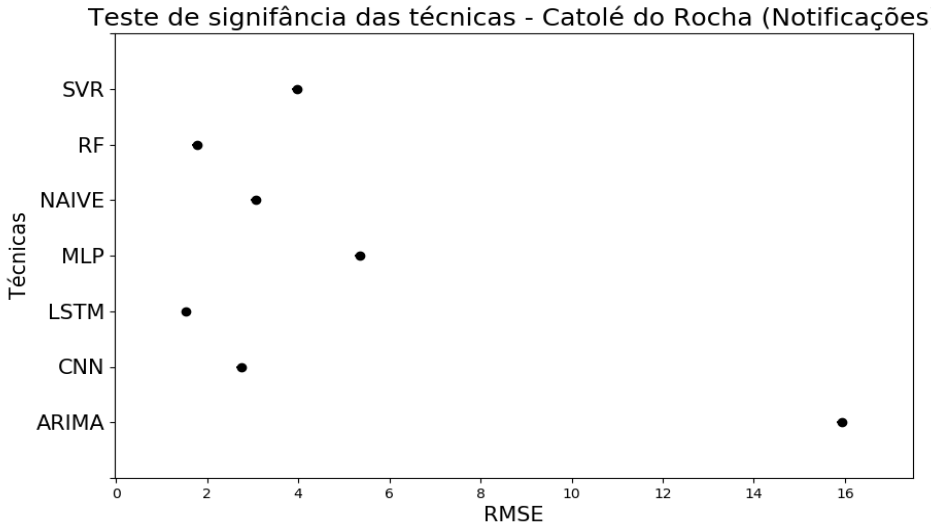
Gráfico 26 - Previsões de notificações por técnica para a cidade de Campina Grande



Fonte: Autoria própria

Os testes de significância estatística relativos à cidade de Catolé do Rocha estão demonstrados no Gráfico 25.

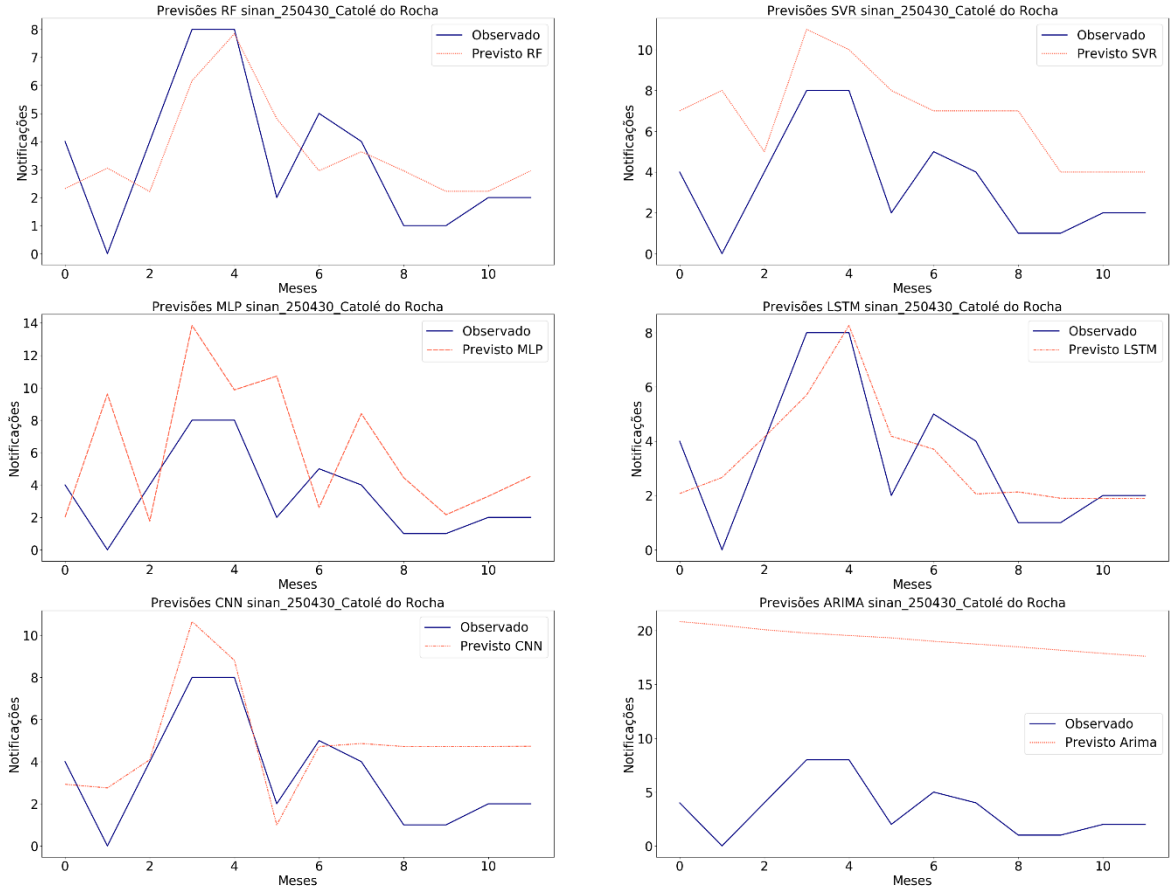
Gráfico 27 - Análise de significância estatística para as previsões de notificações para a cidade de Catolé do Rocha



Fonte: Autoria própria

De acordo com os resultados, há diferença estatística entre a técnica que produziu a menor taxa de erro durante as previsões (LSTM) e as demais técnicas: RF, CNN, NAIVE, SVR, MLP e ARIMA. Ao examinar o Gráfico 26, fica claro a melhor adequação da LSTM às notificações de dengue presentes nessa cidade. Ainda desse gráfico, pode-se notar também a boa adequação da curva das previsões via técnica CNN até o 6º mês de previsão. Contudo, os resultados não seguiram a tendência de observações nos seis meses finais.

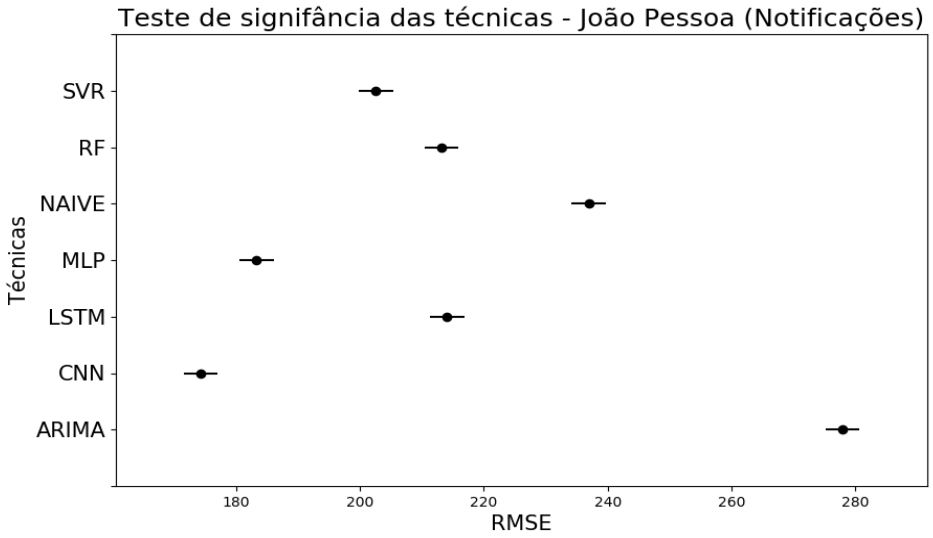
Gráfico 28 - Previsões de notificações por técnica para a cidade de Catolé do Rocha



Fonte: Autoria própria

Os testes de Tukey para a cidade de João Pessoa estão relatados no Gráfico 29.

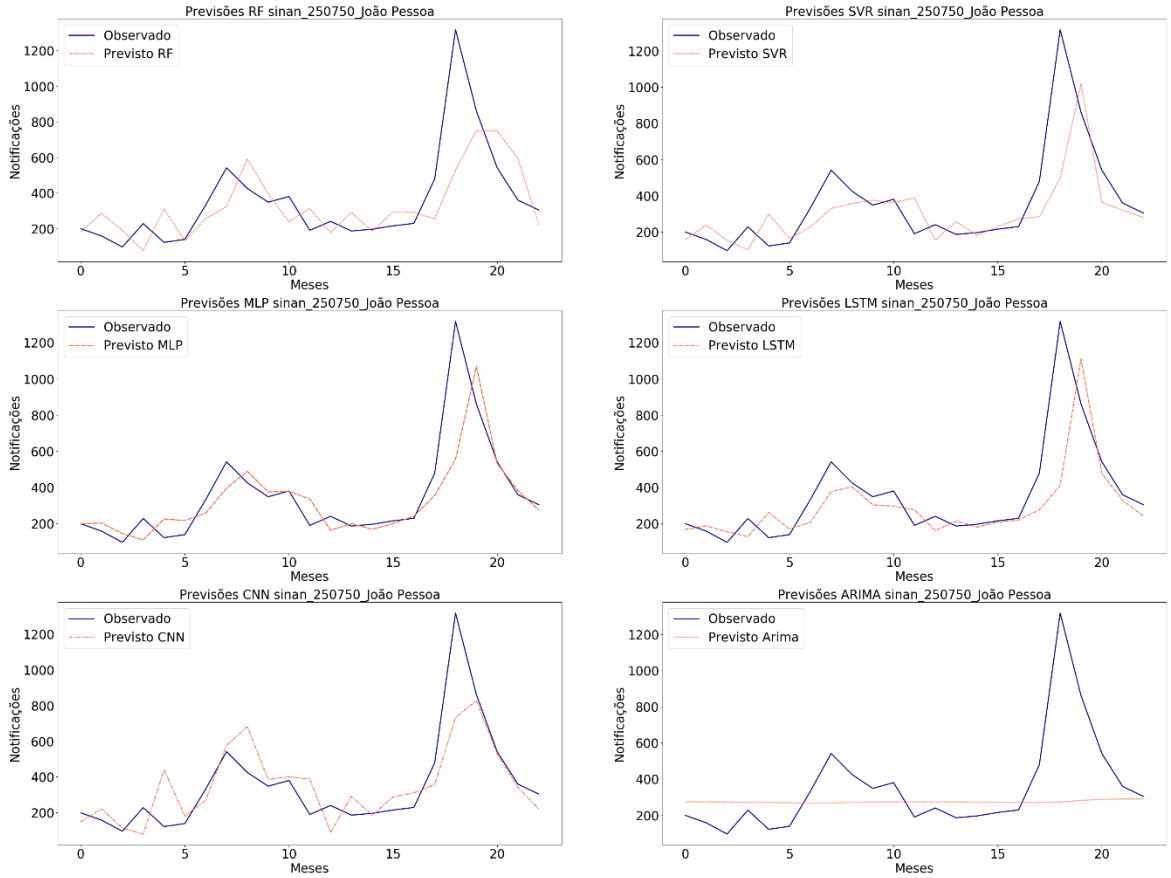
Gráfico 29 - Análise de significância estatística para as previsões de notificações para a cidade de João Pessoa



Fonte: Autoria própria

A técnica CNN obteve o menor valor de RMSE e foi considerada a melhor técnica. Segundo o teste estatístico, ficou comprovada a sua diferença em relação às demais. Ao estudar as previsões calculadas para João Pessoa, presentes no Gráfico 30, fica confirmada que, até o 15º mês, com exceção do ARIMA, as técnicas estavam acompanhando bem a linha de observações de dengue. Contudo, após o 15 mês, a CNN foi a técnica que melhor se adequa a curva de observações. Provavelmente, a camada convolucional conseguiu encontrar padrões adicionais não identificados pelas demais técnicas.

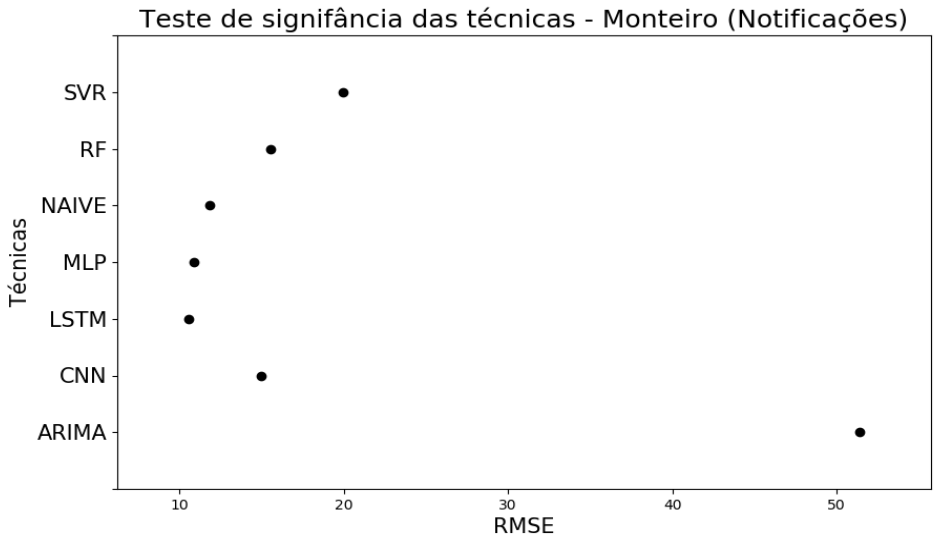
Gráfico 30 - Previsões de notificações por técnica para a cidade de João Pessoa



Fonte: Autoria própria

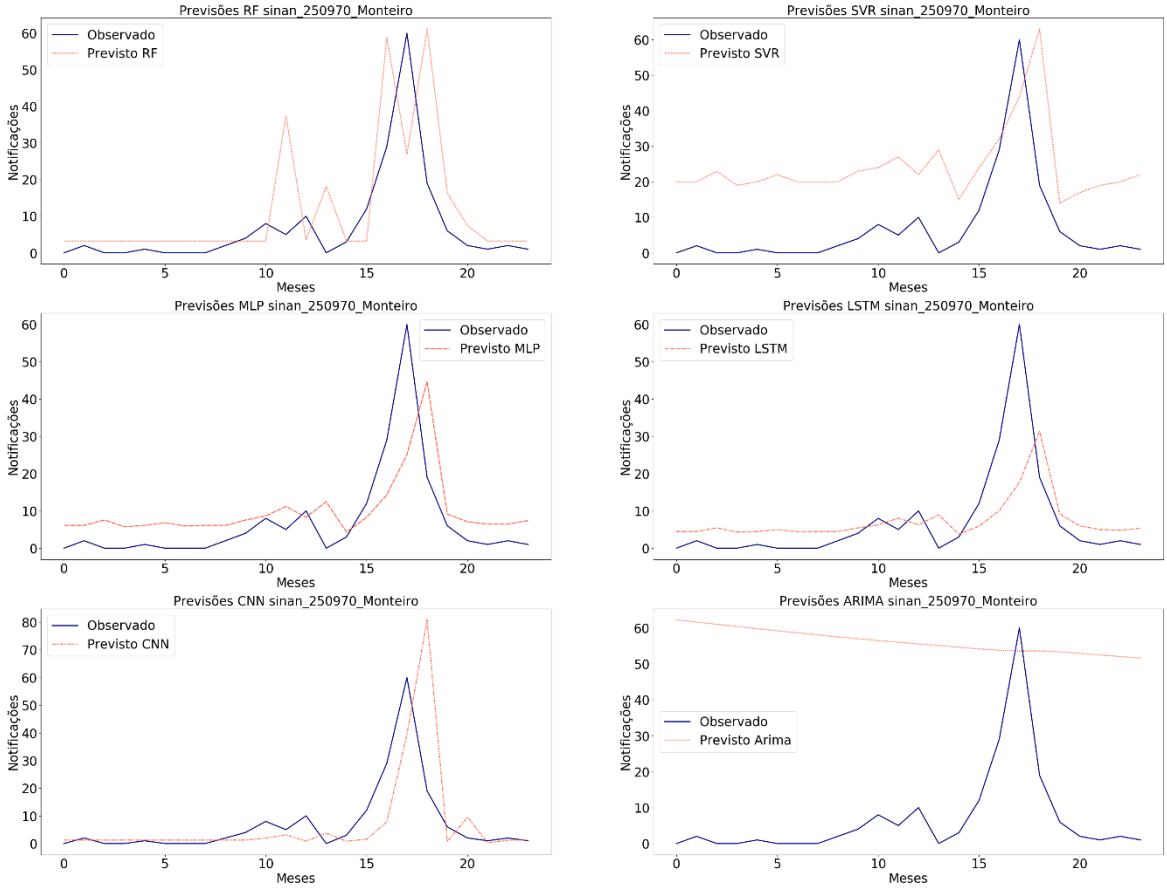
Para a cidade de Monteiro, os resultados apontam LSTM como a técnica com menor taxa de erro. Todavia, ao executar os testes de Tukey da LSTM com MLP, obtivemos um *p-value* de 0,1829. Logo, não podemos rejeitar a hipótese nula do teste e, estatisticamente, não há diferença entre os resultados produzidos por elas. Os demais resultados do teste estão presentes no Gráfico 31. Em relação às previsões, os resultados estão ilustrados no Gráfico 32 e demonstram as similaridades entre as curvas de previsões da LSTM e MLP.

Gráfico 31 - Análise de significância estatística para as previsões de notificações para a cidade de Monteiro



Fonte: Autoria própria

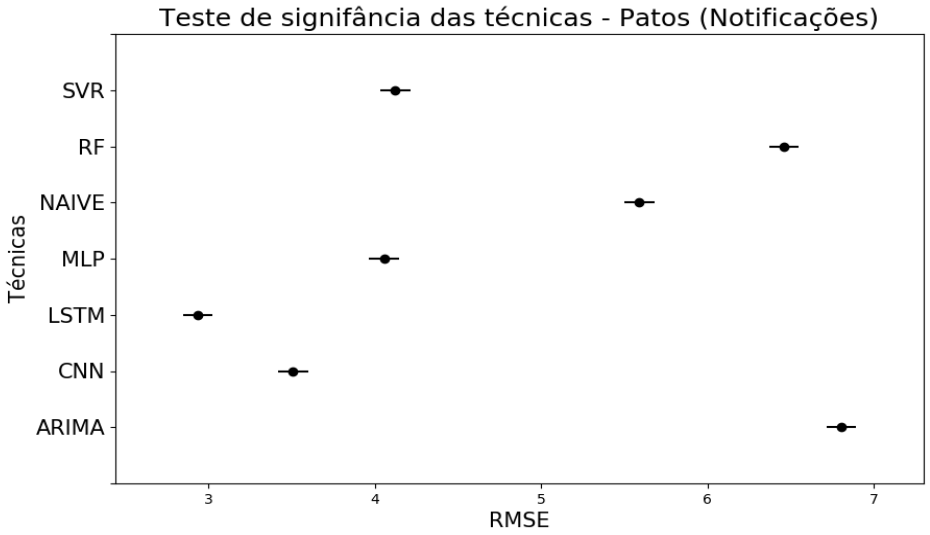
Gráfico 32 - Previsões de notificações por técnica para a cidade de Monteiro



Fonte: Autoria própria

Os resultados dos testes de Tukey para a cidade de Patos estão representados no Gráfico 33.

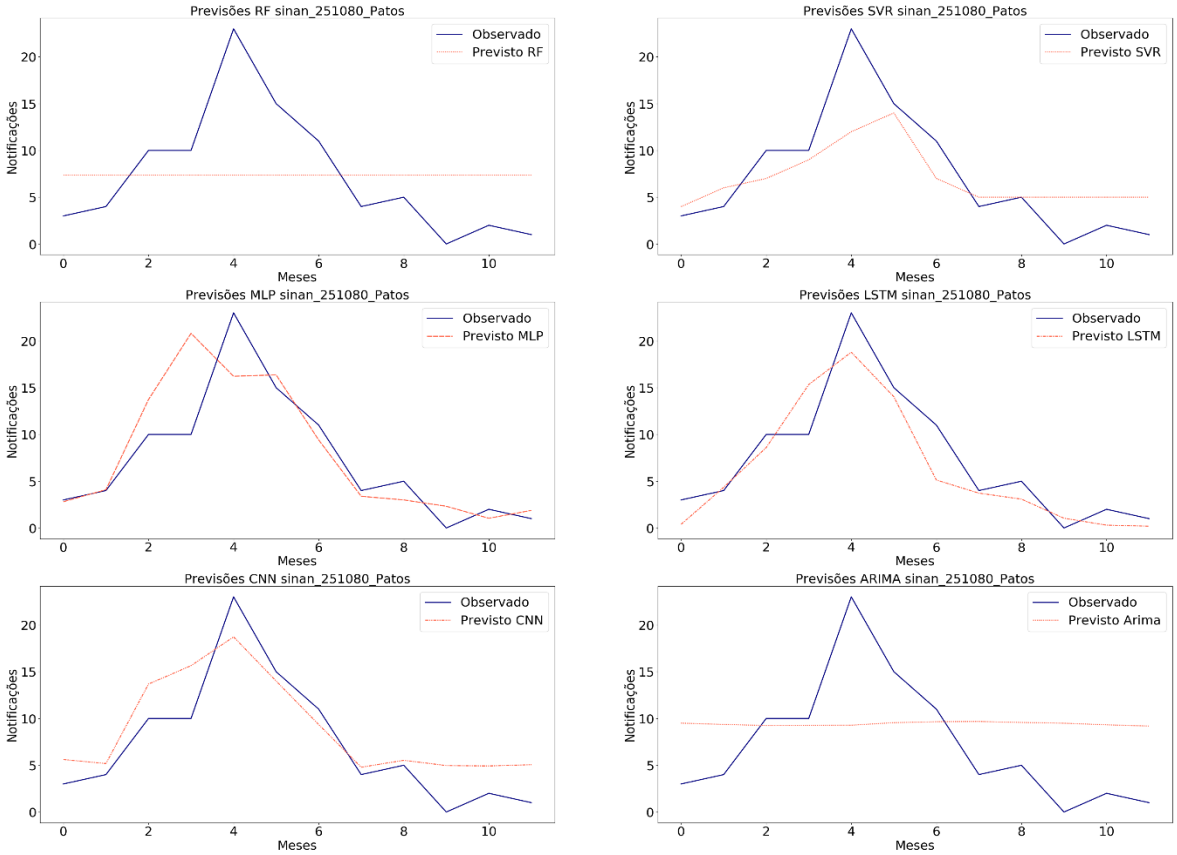
Gráfico 33 - Análise de significância estatística para as previsões de notificações para a cidade de Patos



Fonte: Autoria própria

O gráfico evidencia a diferença estatística entre a LSTM, técnica com melhor previsões para a cidade de Patos, e as outras técnicas abordadas. Ao analisarmos as previsões presentes no Gráfico 34, fica claro o melhor ajuste da curva da técnica LSTM em relação aos valores de observações reais. Ademais, as outras técnicas de *Deep Learning*, CNN e MLP, também conseguiram acompanhar bem a tendencia de aumento e de diminuição dos casos de dengue durante os doze meses do período de testes.

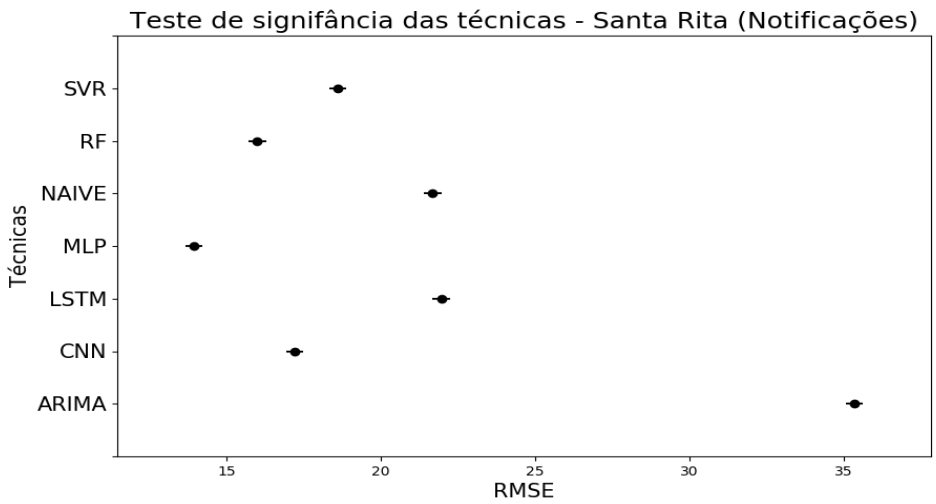
Gráfico 34 - Previsões de notificações por técnica para a cidade de Patos



Fonte: Autoria própria

As comparações estatísticas dos resultados para a cidade de Santa Rita estão ilustradas no Gráfico 35.

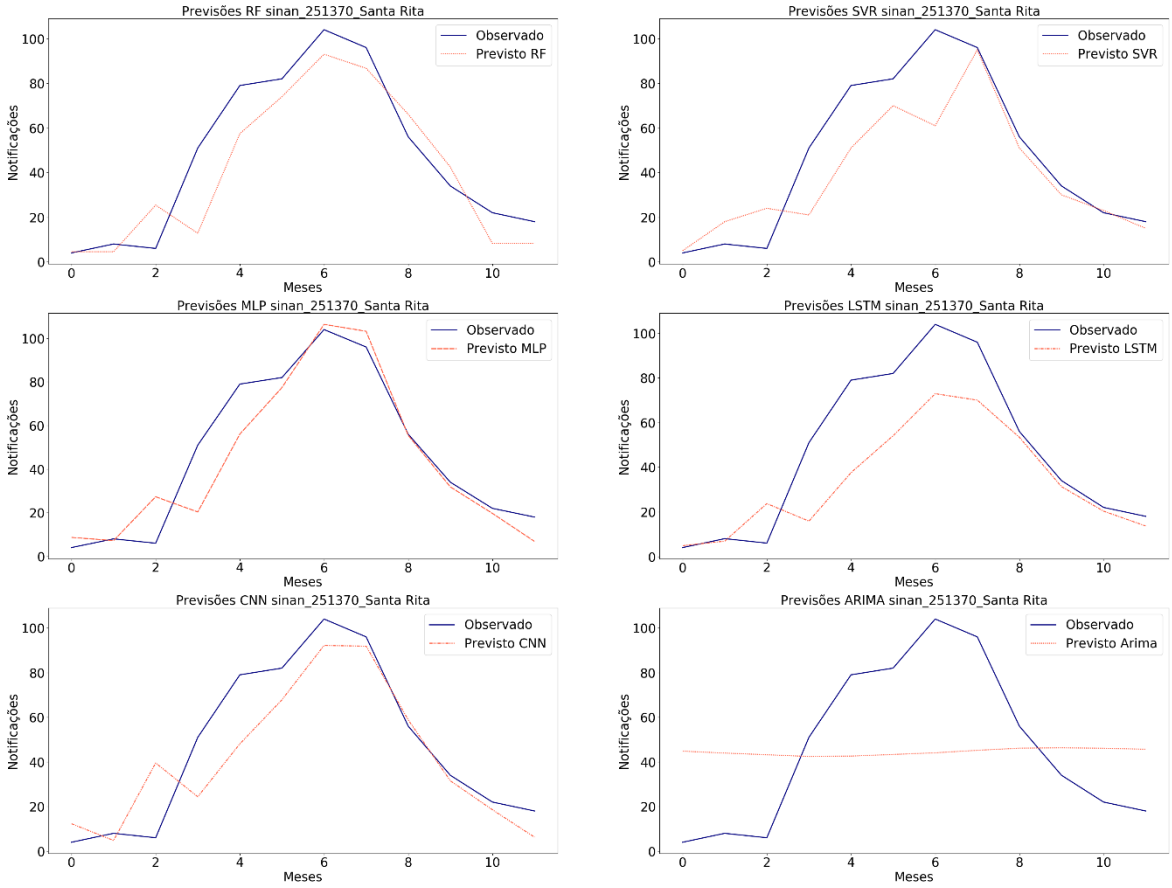
Gráfico 35 - Análise de significância estatística para as previsões de notificações para a cidade de Santa Rita



Fonte: Autoria própria

A técnica MLP conseguiu prever os casos de notificações para Santa Rita com menor taxa de erro e os seus resultados são significantemente diferentes das demais técnicas. A validação visual (Gráfico 36) ilustra a capacidade da técnica de MLP em capturar a tendência de crescimento e, posteriormente, decréscimo de casos de notificações em Santa Rita.

Gráfico 36 - Previsões de notificações por técnica para a cidade de Santa Rita



Fonte: Autoria própria

5 CONSIDERAÇÕES FINAIS

Após coletar dados epidemiológicos, climáticos e sanitários, armazená-los em um banco de dados e empregando técnicas de ML (*Random Forest* e *Support Vector Regression*) e de DL (*Multilayer Perceptron, Long Short-Term Memory e Convolutional Neural Network*), foi possível criar um sistema capaz de encontrar os melhores atributos previsores e realizar previsões de notificações e de internações causadas por dengue para municípios Paraibanos. Além das previsões, foram determinadas e comprovadas, estatisticamente, quais técnicas produziram os melhores resultados por cidade.

Com base nos resultados produzidos, os governantes e os profissionais da saúde poderão utilizar as informações em ações preventivas no combate à dengue, na melhor gestão de recursos financeiros, humanos e hospitalares.

Ao final das análises, ficou demonstrada a superioridade das técnicas de DL em comparação as técnicas de ML. Durante a previsão de casos de notificações, a técnica LSTM obteve melhores resultados em 66,67% das cidades, CNN em 22,22% e MLP em 11,11%. Em relação às internações, LSTM obteve menor taxa de erro em 33,34% dos munícipios, CNN, MLP e RF, cada uma delas, obtiveram melhores resultados em 22,22% das cidades.

As escolhas dos parâmetros previsores mostram significantes achados do trabalho referentes aos aspectos epidemiológicos, climáticos, sanitários e sociais nos munícipios. Para os municípios da Mata Paraibana, o sistema indicou a utilização da coleta de esgoto durante as previsões de internações e de notificações para Santa Rita (cidade com menor taxa de esgotamento da Mata Paraibana). Em contrapartida, não foi utilizado esse atributo para a cidade com maior taxa de coleta (João Pessoa). Esse fato pode servir como base de futuros estudos para a investigação da influência da rede sanitária no número de casos de dengue, bem como nas ações contra à doença.

O padrão na utilização de parâmetros também foi observado nos municípios do Sertão Paraibano (Cajazeiras, Catolé do Rocha e Patos). Para essas cidades, os parâmetros epidemiológicos e a pluviometria mensal mostraram maior significância. Essa descoberta demonstra que, para cada mesorregião, alguns fatores são mais relevantes na proliferação da dengue e, consequentemente, deve haver um combate

personalizado para cada região. Adicionalmente, a escolha dos parâmetros pode embasar outros estudos, evitando, assim, a busca de parâmetros e técnicas para cidades com características similares as aqui trabalhadas.

Percebeu-se também a disparidade na detecção de *outliers* durante a escolha dos cenários de pesquisa. Para os casos de internações, a exclusão de *outliers* ocorreu em 72% das cidades contra 44% nos casos de notificações. Estudos futuros podem averiguar se há uma melhor qualidade nos dados providos pelo SIH em comparação aos relatados pelo SINAN.

Embora a criação do sistema tenha sido feita de forma estruturada e pensando na utilização em ambiente real, a falta de uma interface gráfica contendo os resultados produzidos é uma limitação da pesquisa. A completude dos dados para apenas nove das 223 cidades da Paraíba pode ser destacada como deficiência. Adicionalmente, a ausência dos dados referentes ao ano de 2020 também é uma deficiência. Contudo, para o último levamento há justificava, pois os dados do SNIS relacionados ao ano de 2020 serão disponibilizados apenas em dezembro de 2021. Por fim, a falta do teste de homoscedasticidade, pressuposto do teste ANOVA, pode ser um risco e deficiência desta pesquisa.

Em relação às sugestões para trabalhos futuros, novos atributos previsores como, por exemplo, informações acerca da coleta de lixo, informações demográficas das cidades, dados de temperatura, de umidade, além da combinação de novas informações epidemiológicas podem ser utilizados durante as previsões. Além disso, se possível, a coleta dos dados tendo como a unidade os bairros das cidades e a utilização de um período maior de dados, poderiam elevar o grau de precisão dos resultados e, consequentemente, a maior personalização das ações de combate à dengue.

Devido à constante evolução da tecnologia da informação, também é sugerido para trabalhos futuros a investigação de novas técnicas de ML e de DL juntamente com novas combinações e configurações das técnicas utilizadas por este trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, L. S.; COTA, A. L. S.; RODRIGUES, D. F. Saneamento, Arboviroses e Determinantes Ambientais: impactos na saúde urbana. **Ciência & Saúde Coletiva**, v. 25, n. 10, p. 3857–3868, 2020.

APPICE, A. et al. A Multi-Stage Machine Learning Approach to Predict Dengue Incidence: A Case Study in Mexico. **IEEE Access**, v. 8, p. 52713–52725, 2020.

AWAD, M.; KHANNA, R. **Efficient learning machines: Theories, concepts, and applications for engineers and system designers**. Springer nature, 2015.

BARBOSA, R. M. R. et al. Infestation of an endemic arbovirus area by sympatric populations of Aedes aegypti and Aedes albopictus in Brazil. **Memórias do Instituto Oswaldo Cruz**, v. 115, n. 4, 2020.

BATISTA, E. D. DE A. et al. Previsão de casos de dengue através de Machine Learning e Deep Learning: uma revisão sistemática. **Research, Society and Development**, v. 10, n. 11, p. e33101119347, 22 ago. 2021.

BESERRA, E. B. et al. Ciclo de vida de Aedes (Stegomyia) aegypti (Diptera, Culicidae) em águas com diferentes características. **Iheringia. Série Zoologia**, v. 99, n. 3, p. 281–285, 2009.

BONACCORSO, G. **Machine Learning Algorithms: Popular algorithms for data science and machine learning**. Packt Publishing Ltd, 2018.

BRASIL. Ministério da Saúde alerta para aumento de 149% dos casos de dengue no país. **Ministério da Saúde, Brasil**, p. 2020, 2019.

CÂMARA, F. P. et al. Estudo retrospectivo (histórico) da dengue no Brasil: características regionais e dinâmicas. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 40, n. 2, p. 192–196, 2007.

CARVAJAL, T. M. et al. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. **BMC Infectious Diseases**, v. 18, n. 1, 2018.

CARVALHO, F. D.; MOREIRA, L. A. Why is Aedes aegypti Linnaeus so Successful as a Species?. **Neotropical Entomology**, v.46, n3, p. 243-255, 2017.

CHEN, Y. et al. Neighbourhoodlevel real-time forecasting of dengue cases in tropical urban Singapore. **BMC Medicine**, v. 16, n. 1, p. 129, 2018.

CHOLLET, F. **Deep Learning with Python**. Simon and Schuster, 2017.

DE JESUS, J. G. et al. Genomic detection of a virus lineage replacement event of dengue virus serotype 2 in Brazil, 2019. **Memorias do Instituto Oswaldo Cruz**, v. 115, n. 4, 2020.

DEITEL, PAUL J.; DEITEL, H. **Intro to Python for Computer Science and Data**

Science: Learning to Program with AI, Big Data and the Cloud. Pearson Education, 2019.

DONI, A. R.; SASIPRABA, T. Lstm-Rnn Based Approach for Prediction of Dengue Cases in India. **Ingenierie des Systemes d'Information**, v. 25, n. 3, p. 327–3355, 2020.

DUARTE, D.; FAERMAN, J. Comparison of Time Series Prediction of Healthcare Emergency Department Indicators with ARIMA and Prophet. In: In: **9th International Conference on Computer Science, Engineering and Applications (ICCSEA 2019)**. p. 123-133, 2019.

FARES, R. C. G. et al. Epidemiological Scenario of Dengue in Brazil. **BioMed Research International**, v. 2015, 2015.

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Rio de Janeiro: Alta Books, 2019.

GOODFELLOW, IAN; BENGIO, YOSHUA; COURVILLER, A. **Deep learning**. Mit press, 2016.

GRACIANO, A. R. et al. Morbimortalidade da dengue em idosos no Brasil - Dengue morbidity and mortality in elderly in Brazil. **Revista Educação em Saúde**, v. 5, n. 1, p. 56-65, 2017.

GUERRIERO, I. C. Z. Resolução nº 510 de 7 de abril de 2016 que trata das especificidades éticas das pesquisas nas ciências humanas e sociais e de outras que utilizam metodologias próprias dessas áreas. **Ciência & Saúde Coletiva**, v. 21, n. 8, p. 2619–2629, 2016.

GUO, P. et al. Developing a dengue forecast model using machine learning: A case study in China. **PLoS Neglected Tropical Diseases**, v. 11, n. 10, p. e0005973, 2017.

HARISSON, M. **Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python**. Novatec, 2020.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. OTexts, 2018.

IOC/FIOCRUZ. **Dengue**. Disponível em: <http://www.ioc.fiocruz.br/dengue/textos/longatraje.html>. Acesso em: 23 ago. 2021.

IOC/FIOCRUZ. **Como é o ciclo de vida do mosquito "Aedes aegypti"? - Perguntas e respostas Fiocruz**. Disponível em: <https://portal.fiocruz.br/pergunta/como-e-o-ciclo-de-vida-do-mosquito-aedes-aegypti>. Acesso em: 24 ago. 2021.

IZBICKI, RAFAEL; SANTOS, T. M. DOS. **Aprendizado de máquina: uma abordagem estatística**. Rafael Izbicki, 2020.

KANJI, G. **100 Statistical Tests**. Sage, 2006.

KIRBY, S.; PARAMAGURU, K.; WARREN, J. The Accuracy of NIESR's GDP Growth Forecasts. **National Institute Economic Review**, v. 232, n. 1, p. 41–47, 2015.

KUHN, MAX; JOHNSON, K. **Feature engineering and selection: A practical approach for predictive models**. CRC Press, 2019.

LEANDRO, C. DOS S. et al. Redução da incidência de dengue no Brasil em 2020: controle ou subnotificação de casos por COVID-19? **Research, Society and Development**, v. 9, n. 11, p. e76891110442, 2020.

LEITE, P. L. E. Impacto da dengue no Brasil em período epidêmico e não epidêmico: incidência, mortalidade, custo hospitalar e disability adjusted life years (DALY). 2015.

LONGARAY, André Andrade; CASTELLI, Tiago Machado. Avaliação do desempenho do uso da tecnologia da informação na saúde: revisão sistemática da literatura sobre o tema. **Ciência & Saúde Coletiva**, v. 25, p. 4327-4338, 2020.

MAIA, C. V. DE A. et al. DISTRIBUIÇÃO ESPACIAL DE CRIADOUROS DE AEDES AEGYPTI EM JAGUARUANA – CE – BRASIL E SUAS CORRELAÇÕES COM INDICADORES SOCIODEMOGRÁFICOS. **Hygeia - Revista Brasileira de Geografia Médica e da Saúde**, v. 15, n. 31, p. 71–81, 2019.

MARQUES-TOLEDO, C. DE A. et al. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. **PLoS Neglected Tropical Diseases**, v. 11, n. 7, p e0005729, 2017.

MENDONÇA, F. DE A.; SOUZA, A. V. E; DUTRA, D. DE A. Saúde pública, urbanização e dengue no Brasil. **Sociedade & Natureza**, v. 21, n. 3, p. 257–269, 2009.

MOHER, D. et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. **PLoS Medicine**, v.6, n.7, p e 1000097, 2009.

MUSSUMECI, E.; CODEÇO COELHO, F. Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression. **Spatial and Spatio-temporal Epidemiology**, v. 35, p. 100372, 2020.

NASCIMENTO, C. S. et al. Impactos no perfil epidemiológico da Dengue em meio a Pandemia da COVID-19 em Sergipe. **Research, Society and Development**, v. 10, n. 5, p. e3610514544, 2021.

NORRBY, R. **Outlook for a dengue vaccineClinical Microbiology and Infection**, 2014. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>. Acesso em: 21 set. 2020.

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

PARAÍBA. **Boletim Epidemiológico de Arbovirores - Julho - 2021**. Disponível em: <https://paraiba.pb.gov.br/diretas/saude/arquivos-1/vigilancia-em-saude/be_arbo_07_2021-2-6-2.pdf>. Acesso em: 24 ago. 2021.

PARREIRA, R. M. S.; ATOUGUIA, J. L. M. DA S. DE; SOUSA, C. A. Uma nova ameaça à nossa porta: consequências da dispersão do vírus da dengue num mundo em constante mudança. **Anais do Instituto de Higiene e Medicina Tropical**, v. 12, p. 46–51, 2013.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. 85, p. 2825–2830, 2011.

PHAM, D. N. et al. How to Efficiently Predict Dengue Incidence in Kuala Lumpur.In: In: **2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)**. IEEE, 2018.

PINOCHET, Luis Hernan Contreras. Tendências de tecnologia de informação na gestão da saúde. **Mundo saúde**, v. 35, n. 4, p. 382-94, 2011.

REN, J. et al. On software defect prediction using machine learning. **Journal of Applied Mathematics**, v. 2014, 2014.

RIBEIRO, M. S. et al. Índices larvais de *Aedes aegypti* e incidência de dengue: um estudo ecológico no Estado do Rio de Janeiro, Brasil. **Cadernos de Saúde Pública**, v. 37, n. 7, 2021.

ROSA, R. S. DA; BRAIDO, J. A.; CAPORLINGUA, V. H. CONCEPÇÕES E PRÁTICAS EM EDUCAÇÃO AMBIENTAL DOS AGENTES DE COMBATE A ENDEMIAS NO MUNICÍPIO DE NOVO HAMBURGO/RS. **Revista Prâksis**, v. 1, n. 1, p. 118–136, 2020.

RUSSELL, STUART; NORVIG, P. **Inteligência Artificial**. Elsevier, 2013.

SHMUELI, G.; LICHTENDAHL JR, K. C. **Practical time series forecasting with r: A hands-on guide**. Axelrod Schnall Publishers, 2016.

SILVA, F. J. F. DA. Previsão de internações hospitalares de dengue por meio de séries temporais, 2019.

SILVEIRA, L. T. C. DA; TURA, B.; SANTOS, M. Systematic review of dengue vaccine efficacy. **BMC Infectious Diseases**, v. 19, n. 1, p. 1-8, 2019.

SIPPYID, R. et al. Severity index for suspected arbovirus (SISA): Machine learning for accurate prediction of hospitalization in subjects suspected of arboviral infection. **PLoS Neglected Tropical Diseases**, v. 14, n. 2, p. e0007969, 2020.

SOBRAL, M. F. F.; SOBRAL, A. I. G. DA P. Casos de dengue e coleta de lixo urbano: um estudo na Cidade do Recife, Brasil. **Ciência & Saúde Coletiva**, v. 24, n. 3, p. 1075–1082, 2019.

SOUZA, R. F. DE; ALBUQUERQUE, A. R. DA C. GEOGRAFIA DA DENGUE: UMA ANÁLISE DAS POLÍTICAS DE CONTROLE E MONITORAMENTO DO AEDES AEGYPTI EM MANAUS/ Geography of Dengue: an analysis of the control and monitoring policies of Aedes aegypti in Manaus. **REVISTA GEONORTE**, v. 9, n. 31, p. 68–76, 2018.

STOLERMAN, L. M.; MAIA, P. D.; NATHAN KUTZ, J. Forecasting dengue fever in Brazil: An assessment of climate conditions. **PLoS ONE**, v. 14, n. 8, p. e0220106, 2019.

SWAMINATHAN, S.; KHANNA, N. Dengue vaccine development: Global and Indian scenarios. **International Journal of Infectious Diseases**, v. 84, p. S80–S86, 2019.

TEICH, V.; ARINELLI, R.; FAHHAM, L. Aedes aegypti e sociedade: o impacto econômico das arboviroses no Brasil. **Jornal Brasileiro de Economia da Saúde**, v. 9, n. 3, p. 267–276, 2017.

VIANA, D. V.; IGNOTTI, E. A ocorrência da dengue e variações meteorológicas no brasil: Revisão sistemática. **Revista Brasileira de Epidemiologia**, v. 16, n. 2, p. 240–256, 2013.

XU, J. et al. Forecast of dengue cases in 20 chinese cities based on the deep learning method. **International Journal of Environmental Research and Public Health**, v. 17, n. 2, p. 453, 2020.

ZARA, A. L. DE S. A. et al. Estratégias de controle do Aedes aegypti: uma revisão. **Epidemiologia e Serviços de Saúde**, v. 25, p. 391-404, 2016.

APÊNDICE A – REVISÃO SISTEMÁTICA

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Previsão de casos de dengue através de *Machine Learning* e *Deep Learning*: uma revisão sistemática

Predicting dengue cases through Machine Learning and Deep Learning: a systematic review

Predicción de casos de dengue a través del aprendizaje automático y el aprendizaje profundo: una revisión sistemática

Recebido: 12/08/2021 | Revisado: 16/08/2021 | Aceito: 18/08/2021 | Publicado: 22/08/2021

Ewerthon Dyego de Araújo Batista
ORCID: https://orcid.org/0000-0003-4993-9900
Universidade Estadual da Paraíba, Brasil
E-mail: edabew@gmail.com
Wellington Candeia de Araújo
ORCID: https://orcid.org/0000-0003-2102-7993
Universidade Estadual da Paraíba, Brasil
E-mail: wcandeia@uepb.edu.br
Romeryto Vieira Lira
ORCID: https://orcid.org/0000-0003-2567-0839
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Brasil
E-mail: romeryto.lira@academico.ifpb.edu.br
Laryssa Izabel de Araújo Batista
ORCID: https://orcid.org/0000-0002-0188-9425
Universidade Federal da Paraíba, Brasil
E-mail: laryssa.izabel@gmail.com

Resumo

Introdução: a dengue é uma arbovirose causada pelo vírus DENV e transmitida para o homem através do mosquito *Aedes aegypti*. Atualmente, não existe uma vacina eficaz para combater todas as sorologias do vírus. Diante disso, o combate à doença se volta para medidas preventivas contra a proliferação do mosquito. Os pesquisadores estão utilizando *Machine Learning (ML)* e *Deep Learning (DL)* como ferramentas para prever casos de dengue e ajudar os governantes nesse combate. Objetivo: identificar quais técnicas e abordagens de ML e de DL estão sendo utilizadas na previsão de dengue. Métodos: revisão sistemática realizada nas bases das áreas de Medicina e de Computação com intuito de responder as perguntas de pesquisa: é possível realizar previsões de casos de dengue através de técnicas de ML e de DL, quais técnicas são utilizadas, onde os estudos estão sendo realizados, como e quais dados estão sendo utilizados? Resultados: após realizar as buscas, aplicar os critérios de inclusão, exclusão e leitura aprofundada, 14 artigos foram aprovados. As técnicas *Random Forest* (RF), *Support Vector Regression* (SVR), e *Long Short-Term Memory* (LSTM) estão presentes em 85% dos trabalhos. Em relação aos dados, na maioria, foram utilizados 10 anos de dados históricos da doença e informações climáticas. Por fim, a técnica *Root Mean Absolute Error* (RMSE) foi a preferida para mensurar o erro. Conclusão: a revisão evidenciou a viabilidade da utilização de técnicas de ML e de DL para a previsão de casos de dengue, com baixa taxa de erro e validada através de técnicas estatísticas.
Palavras-chave: Dengue; Previsão; Machine learning; Deep learning.

Abstract

Introduction: dengue is an arbovirus caused by the DENV virus and transmitted to humans through the Aedes aegypti mosquito. Currently, there is no effective vaccine to combat all serology of the virus. Therefore, the fight against the disease turns to preventive measures against the proliferation of the mosquito. Researchers are using Machine Learning (ML) and Deep Learning (DL) as tools to predict cases of dengue and help governments in this fight. Objective: to identify which ML and DL techniques and approaches are being used to predict dengue. Methods: systematic review carried out on the bases of the areas of Medicine and Computing in order to answer the research questions: it is possible to make predictions of dengue cases using ML and DL techniques, which techniques are used, where the studies are being performed, how and what data is being used? Results: after performing the searches, applying the inclusion, exclusion and in-depth reading criteria, 14 articles were approved. The Random Forest (RF), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) techniques are present in 85% of the works. Regarding the data, most were used 10 years of historical data on the disease and climate information. Finally, the Root Mean Absolute Error (RMSE) technique was preferred to measure the error. Conclusion: the review showed the feasibility of using ML and DL techniques to predict dengue cases, with a low error rate and validated through statistical techniques.
Keywords: Forecast; Machine learning; Deep learning.

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Resumen
Introducción: el dengue es un arbovirus causado por el virus DENV y transmitido al ser humano a través del mosquito Aedes aegypti. Actualmente, no existe una vacuna eficaz para combatir todas las serologías del virus. Por tanto, la lucha contra la enfermedad se convierte en medidas preventivas contra la proliferación del mosquito. Los investigadores están utilizando Machine Learning (ML) y Deep Learning (DL) como herramientas para predecir casos de dengue y ayudar a los gobiernos en esta lucha. Objetivo: identificar qué técnicas y enfoques de LD y LD se están utilizando para predecir el dengue. Métodos: revisión sistemática realizada sobre las bases de las áreas de Medicina y Computación para dar respuesta a las preguntas de investigación: es posible realizar predicciones de casos de dengue utilizando técnicas de ML y DL, qué técnicas se utilizan, dónde se están realizando los estudios, ¿cómo y qué datos se utilizan? Resultados: luego de realizar las búsquedas, aplicando los criterios de inclusión, exclusión y lectura en profundidad, se aprobaron 14 artículos. Las técnicas Random Forest (RF), Support Vector Regression (SVR) y Long Short-Term Memory (LSTM) están presentes en el 85% de los trabajos. En cuanto a los datos, la mayoría se utilizaron 10 años de datos históricos sobre la enfermedad y la información climática. Finalmente, se prefirió la técnica de Root Mean Absolute Error (RMSE) para medir el error. Conclusión: la revisión mostró la viabilidad de utilizar técnicas de LD y LD para predecir casos de dengue, con una baja tasa de error y validadas mediante técnicas estadísticas.
Palabras clave: Dengue; Pronóstico; Aprendizaje automático; Aprendizaje profundo.

1. Introdução

Em circulação no Brasil desde 1981, a dengue foi considerada uma doença extinta, porém reapareceu e passou a ser classificada como endêmica (Câmara et al., 2007). Os principais sintomas da dengue são: febre alta, dores musculares, mal-estar, falta de apetite e dores de cabeça. Os casos mais graves da dengue podem causar hemorragias e levar o paciente a óbito (Graciano et al., 2017).

A arbovirose dengue pode ser originada por cinco variantes do vírus DENV (Mustafa et al., 2015) e, no Brasil, estão em circulação os sorotipos DENV-1, DENV-2, DENV-3 e DENV-4. Uma vez infectado por algum dos quatro tipos do vírus, o paciente continua vulnerável aos demais (Neto, Nascimento, Sousa, 2016). O mosquito *Aedes aegypti* encontra em países de clima tropical, como o Brasil, ambientes propícios para sua reprodução. Adicionalmente, problemas sociais e sanitários como, por exemplo, o descarte incorreto de lixo e o indevido lançamento de esgotos, aumenta o número de locais para as fêmeas do mosquito depositarem seus ovos (de Souza & Albuquerque, 2018).

No Brasil, existe a regulamentação da vacina Dengvaxia para combater a dengue. Contudo, a vacina está disponível apenas na rede particular e o seu uso é indicado, exclusivamente, para pessoas que já tiveram a doença (da Silveira, 2019). Logo, para combater a doença, os sistemas governamentais utilizam campanhas de conscientização e ações contra a proliferação do vetor (Ferreira et al., 2019). No ano de 2019, a *World Health Organization* (WHO) contabilizou cerca de 4.2 milhões de manifestações de dengue em todo o planeta. Anteriormente, esse mesmo órgão, emitiu um alerta classificando a dengue como uma das principais doenças para o ano de 2019. No Brasil, em 2019, devido ao aumento da circulação de um nova variante do vírus DENV-2, houve um novo surto de dengue com crescimento de 149% dos casos em alguns estados (Brasil, 2019; De Jesus et al., 2020).

A criação de ferramentas para prever dengue é uma tarefa complexa, pois vários fatores contribuem para o aparecimento e proliferação da doença. Entretanto, técnicas de *Machine Learning* e de *Deep Learning* vêm ajudando as pesquisas nessa área. Doni e Sasipraba (2020), conduziram um estudo na Índia, utilizando técnicas de *Deep Learning*, para analisar dados climáticos como temperatura, dados de precipitação e umidade. O trabalho conseguiu prever casos de dengue com 89% de acurácia. Outro exemplo da utilização de predição através de ML e de DL ocorreu na Tailândia. Em seu trabalho, os pesquisadores (Puengpreeda, Yhusumram, Sirikulvadhana, 2020) utilizaram dados climáticos e informações de pesquisas realizadas no Google a respeito de dengue.

Ao realizar pesquisa sobre *Machine Learning* e *Deep Learning* para a doença em estudo, na base *Scopus*, mais de 250 artigos foram retornados. Esse fato demonstra a atenção dada pela ciência ao tema. Portanto, combinando a importância da ciência e os problemas de saúde aqui listados, é justificável realizar uma revisão sistemática e verificar quais temas,

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

abordagens e técnicas estão sendo utilizadas nessa área.

O objetivo da pesquisa é, através de uma revisão sistemática, verificar a viabilidade de utilizar técnicas de *Machine Learning* e de *Deep Learning* durante a previsão de dengue, quais técnicas são utilizadas, onde os estudos estão sendo realizados, como e quais dados estão sendo utilizados na previsão e, por fim, quais técnicas estão demonstrando melhores resultados.

2. Metodologia

Esta revisão foi estruturada com base no modelo definido por Levac (Levac, Colquhoun, O'brien, 2010). Em sua metodologia, a autora elencou cinco fases obrigatórias e uma fase opcional para a confecção de uma revisão sistemática. Para este trabalho, adotamos apenas as fases obrigatórias, são elas: 1 – Identificar a questão da pesquisa, 2 – Identificar os estudos relevantes, 3 – Selecionar os estudos, 4 – Mapear os dados e 5 – Coleta, sumarização e relato dos dados. Além da metodologia citada anteriormente, os pesquisadores deste trabalho utilizaram o *software* StArt (Fabbri et al., 2016) como forma de auxílio na definição e execução do protocolo de pesquisa.

2.1 Identificação da questão de pesquisa

O objetivo principal desta revisão é verificar a viabilidade da utilização de técnicas de *Machine Learning* e de *Deep Learning* na previsão de casos da doença dengue. Objetivando um melhor direcionamento do assunto, a equipe subdividiu o problema de pesquisa em quatro frentes: análise de técnicas, quais dados foram utilizados na previsão, quais abordagens foram utilizadas, como foram feitas as validações dos resultados e, por fim, quais foram os melhores resultados. Feito isso, foram criadas as seguintes perguntas derivadas:

1. Quais técnicas de *Machine Learning* e de *Deep Learning* são utilizadas nas previsões?
2. Em qual país foi realizado o estudo? Como foram coletados os dados?
3. Quantos anos de dados foram utilizados nos modelos e quais itens foram considerados na criação dos modelos? Exemplo: Fatores climáticos, econômicos, dados de redes sociais, entre outros.
4. Como foi feita a validação dos modelos? Quais técnicas estatísticas o estudo utilizou na avaliação?
5. Qual técnica ou combinação de técnicas obtiveram os melhores resultados?

2.2 Identificação de estudos relevantes

Após a definição da questão primária da pesquisa e suas derivações, o próximo passo foi definir quais bases de dados seriam relevantes para a revisão. Durante o levantamento dos estudos, foram utilizadas as principais bases eletrônicas nas áreas da Saúde e da Ciência da Computação: *Scopus*, *IEEE Xplore*, *PubMed*, *ACM Digital Library* e *Web of Science*. Outras bases, como *Cochrane*, foram testadas, porém não tiveram adequada indexação de artigos para o tema desta revisão ou o acesso aos artigos estava limitado.

Definidas as bases, a equipe iniciou o estudo sobre os termos para a formação das *strings* de busca. Os termos utilizados nas *strings* de busca fazem referência à previsão, à *Machine Learning* (aqui entende-se *Deep Learning* como sendo um tipo de ML) e à dengue. Segue a listagem das palavras-chave, o conjunto e o escopo abordado por cada uma delas: *predict** (referenciam termos como *predict, prediction, predicted*), *forecast** (contemplando as palavras *forecast, forecasting, forecasted*), *Machine Learning* (não há variação para esse termo), *Deep Learning* (não há variação para esse termo) e dengue (referência à dengue *fever, fever hemorrhagic* dengue e, em alguns países, apenas dengue).

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Utilizando a estratégia de *rounds* e a variação nos operadores lógicos, as combinações das *strings* foram testadas e, a cada *round*, pequenos ajustes foram feitos. Para a avaliar a qualidade da *string*, a equipe elencou dezessete artigos como indispensáveis no retorno das bases. Caso algum desses artigos não fosse retornado, a alteração na *string* era descartada. Por fim, após quatro *rounds*, a equipe chegou no consenso e definiu as seguintes *strings*:

- Base *Scopus*: *TITLE-ABS-KEY ((predict* or forecast*) AND ("machine learning" or "deep learning") AND (dengue))*;
- Base *IEEE*: *((("Full Text & Metadata":predict* or forecast*) AND "Full Text & Metadata":machine learning or deep learning) AND "Full Text & Metadata":dengue);*
- Base *PubMed*: *All Fields ((predict* or forecast*) AND ("machine learning" or "deep learning")) AND (dengue)*;
- Base *ACM*: *[All: predict or forecast] AND [All: machine learning or deep learning] AND [All: dengue] AND [All: predict* or forecast*] AND [All: machine learning or deep learning] AND [All: dengue]*;
- Base *Web of Science*: *TOPIC: (predict* or forecast*) AND TOPIC: (machine learning or deep learning) AND TOPIC: (dengue)*.

2.3 Seleção dos estudos

Machine Learning e de *Deep Learning* são temas atuais, estão em constante evolução, e são cada vez mais empregados na resolução de problemas complexos. Adicionalmente, são amplamente pesquisados nas áreas de Medicina e de Ciência da Computação. Primordialmente, foram selecionados artigos contendo previsões de casos de dengue realizados através de técnicas de ML e de DL. A equipe definiu os critérios de inclusão (INC) e exclusão (EXC) para ter um melhor direcionamento junto à pergunta desta pesquisa. São eles:

- INC01 – O estudo utiliza técnicas de *Machine Learning*;
- INC02 – O estudo utiliza técnicas de *Deep Learning*;
- INC03 – O estudo foi validado estatisticamente;
- INC04 – Existe no estudo a comparação e a utilização de mais de uma técnica ou modelo de ML ou DL;
- INC05 – O estudo contém previsões de casos de dengue;
- INC06 – O artigo deve ser escrito no idioma inglês ou português;
- EXC01 – Publicações além de 5 anos;
- EXC02 – Publicações que utilizam apenas um método de *Machine Learning* ou *Deep Learning*;
- EXC03 – Publicações que utilizam apenas técnicas clássicas da estatística;
- EXC04 – Estudos com baixo grau de validação estatística;
- EXC05 – Trabalhos sem resultados de previsão ou internação de casos de dengue;
- EXC06 – Estudos que não utilizaram bases oficiais ou formais;
- EXC07 – Estudos utilizados para classificar a doença de acordo com os sintomas;
- EXC08 – Estudos não primários;
- EXC09 – Problema de acesso ou não acesso total aos dados;
- EXC10 – Linguagem diferente de inglês ou português.

Após a definição dos critérios, dois pesquisadores iniciaram a triagem dos artigos utilizando os conceitos e as técnicas descritos pelo PRISMA (Moher et al., 2009). Com intuito de evitar influência nos resultados, cada um deles fez a classificação

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

individualmente e sem saber o resultado do outro pesquisador. Durante a seleção dos estudos, as seguintes atividades foram realizadas: 1 – eliminar os artigos duplicados, 2 – leitura rápida do título, *abstract* e resultados dos artigos e 3 – aplicar os critérios de inclusão e exclusão definidos. As exclusões por não obediência aos critérios tiveram o seu registro realizado. Finalizado esse processo, houve comparação dos resultados obtidos. Os artigos com opiniões conflitantes foram submetidos ao crivo de um terceiro pesquisador. O terceiro pesquisador discutiu os pontos levantados preliminarmente, e, por fim, deferiu o parecer final.

2.4 Mapeamento dos dados

Na etapa de mapear os dados, a equipe definiu as informações a serem extraídas dos artigos. Inicialmente, ocorreu uma reunião para a equalização das informações a serem extraídas. Como resultado da reunião, os pesquisadores entraram em consenso para extrair: dados bibliográficos, quais técnicas de ML ou de DL foram utilizadas, qual país o estudo foi feito, como os dados foram coletados e se são oriundos de bases oficiais, quantos anos de dados foram utilizados nas amostras, quais técnicas estatísticas foram utilizadas na validação das predições e, enfim, qual técnica obteve melhor performance. Ademais, os revisores acresceram as suas considerações em relação aos artigos.

À medida que os artigos foram lidos, os revisores, individualmente, incluíram as informações na ferramenta StART. A etapa final do mapeamento foi sumarizar em um arquivo de planilha eletrônica as extrações. Aqui, mais uma vez, surgiram conflitos. Novamente, as divergências foram resolvidas por um terceiro avaliador.

2.5 Coleta, sumarização e relato dos dados

Durante a coleta, sumarização e relato dos resultados, (Levac, Colquhoun, O'brien, 2010) sugere a quebra nos tópicos *analysis*, *reporting* e *implications*. No primeiro, a equipe utilizou a análise quantitativa. Durante a análise quantitativa, os artigos foram classificados e verificada a forma de resposta às subquestões da pesquisa. O relato da revisão foi feito através de quadros contendo as respostas e os dados coletados. Como último passo desta revisão, a equipe emitiu recomendações em relação ao tema estudado e aos achados da revisão.

2.6 Registro de protocolo

Esta revisão sistemática foi guiada através da recomendação *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) (Moher et al., 2009). O protocolo de estudo foi registrado no *Open Science Framework* (OSF) e está disponível através do *link*: https://osf.io/fqa57.

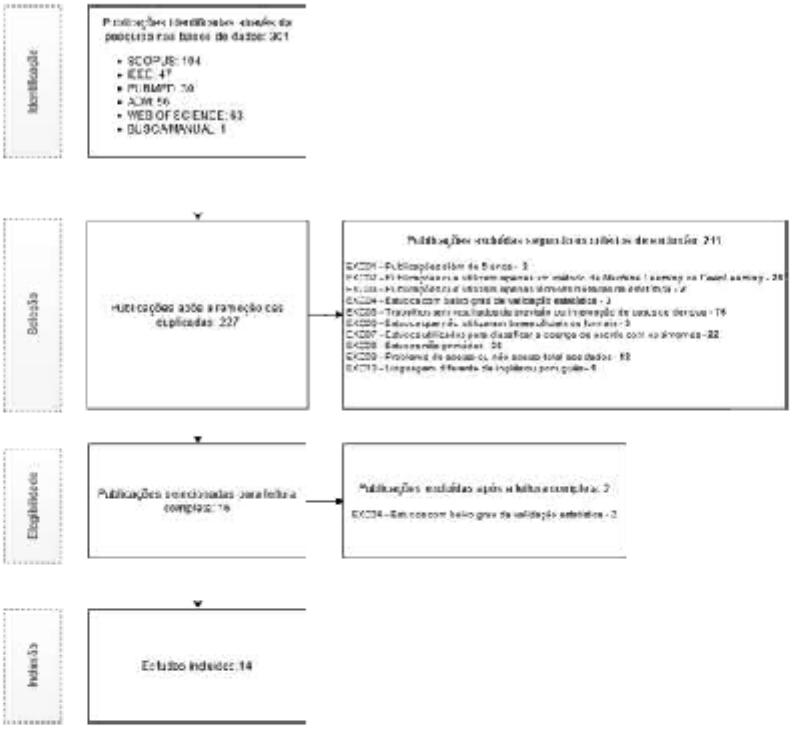
3. Resultados e Discussão

Ao executar o protocolo de revisão, os pesquisadores identificaram 300 artigos. Adicionalmente, 1 artigo foi adicionado manualmente. Após a exclusão dos duplicados, esse número reduziu a 227. Durante as leituras de título, resumo, resultados e aplicando os critérios de inclusão e exclusão, foram selecionados 16 artigos para a leitura aprofundada. Nessa etapa, dois artigos foram excluídos e, enfim, o número de trabalhos aprovados ficou em 14 artigos.

A Figura 1 detalha o processo de seleção, listando o quantitativo de artigos por base e a minúcia das exclusões.

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Figura 1 - Resultado da seleção dos estudos após a execução do protocolo da revisão sistemática.



Fonte: Autores (2021)

As repostas para o questionamento "Quais técnicas de *Machine Learning* e *Deep Learning* são utilizadas nas previsões?" estão representadas no Quadro 1. Como se pode observar, há uma grande variação no emprego das técnicas de *Machine Learning* e de *Deep Learning*. Contudo, técnicas como *Support Vector Machine for regression* (SVR), *Random forest* (RF), *Long short-term memory* (LSTM) estão presentes em 12 dos 14 trabalhos.

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Quadro 1 - Listagem das técnicas de *Machine Learning* e *Deep Learning* utilizada para os trabalhos aprovados.

Estudo	Técnicas ML e DL
(Manogaran & Lopez, 2018)	*Gaussian process regression* (GPR), RF, SVR, *Multiple regression* (MR).
(Appice et al., 2020)	*AUTOencoding based TIme series Clustering with Nearest. Neighbour (*AutoTiC-NN), *K-Nearest Neighbourhood* (KNN), SVR, *Autoregressive integrated moving average* (ARIMA), *M5 regression* (M5).
(Dhaka & Singh, 2020)	RF, *Decision Tree* (DT), *Multiple linear regression* (MLR) e SVR.
(Mishra, Tiwari, Ajaymon, 2019)	*Neural Network* (NN), RF, *Boosted Trees, Least absolute shrinkage and selection operator* (LASSO), *Ridge*, *Extreme Gradient Boosting* (XGBoost) e SVR.
(Guo et al., 2017)	SVR, *Gradient boosted* (GBM), LASSO e *Generalized additive model* (GAM).
(Raju et al., 2019)	SVR, *Ridge* e *Linear Regression* (LR).
(Xu et al., 2020)	LSTM, *Back propagation neural network* (BPNN), *Generalized additive model* (GAM), SVR e *Gradient boosted* (GBM).
(Kerdprasop, Kerdprasop, Chuaybamroong, 2019)	*Chi-squared automatic interaction detection* (CHAID), LR, GLM, *Artificial Neural Network* (ANN) e SVR.
(Pham et al., 2018)	*Genetic Algorithm Enhanced Recurrent Neural Network* (GA_RNN), LR e DT.
(Mussumeci & Codeço Coelho, 2020)	LSTM, RF e LASSO.
(Doni & Sasipraba, 2020)	LSTM, SVR, XGboost, RF, GAM, BPNN
(Carvajal et al., 2018)	RF, GAM e GB.
(Dharmawardana et al., 2018)	ANN E XGBoost.
(Puengpreeda, Yhusumran, Sirikulvadhana, 2020)	RF e *Ridge*.

Fonte: Autores (2021).

O Quadro 2 contém os esclarecimentos para os questionamentos 2 e 3, respectivamente, "Em qual país foi realizado o estudo? Como foram coletados os dados?" e "Quantos anos de dados foram utilizados nos modelos e quais itens foram considerados na criação dos modelos". Para cada pesquisa, foram elencados o país alvo, quantos anos de dados foram utilizados, quais informações foram utilizadas e, por fim, as suas origens.

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Quadro 2 - Levantamento dos países, período, dados utilizados e suas origens para os trabalhos selecionados.

Estudo	País	Período	Dados utilizados	Origem dos dados
(Manogaran & Lopez, 2018)	India	1998 a 2006	Meteorológicos e epidemiológicos	Ministério da saúde
(Appice et al., 2020)	México	1985 a 2010	Epidemiológicos e temperatura	Governo do México
(Dhaka & Singh, 2020)	India	2013 a 2017	Epidemiológicos e climáticos	Ministério da saúde da India
(Mishra, Tiwari Ajaymon, 2019)	Peru	Não informado	Epidemiológicos, sociais e climáticos	*Dataset* dos USA
(Guo et al., 2017)	China	2011 a 2014	Epidemiológicos	Sistema nacional de vigilância da China
(Raju et al., 2019)	India	2001 a 2018	Epidemiológicos e climáticos	Dados do governo de Kerala
(Xu et al., 2020)	China	2005 a 2018	Meteorológicos	Dados do Centro nacional da China
(Kerdprasop, Kerdprasop, Chuaybamroong, 2019)	Tailândia	2003 a 2017	Epidemiológicos e climáticos	Ministério da saúde
(Pham et al., 2018)	Malasia	2002 a 2012	Epidemiológicos e climáticos	Câmara municipal de Kuala Lumpur
(Mussumeci & Codeço Coelho, 2020)	Brasil	2010 a 2018	Epidemiológicos e climáticos	Base InfoDengue
(Doni & Sasipraba, 2020)	India	2015 a 2018	Epidemiológicos e climáticos	Governo da India
(Carvajal et al., 2018)	Filipinas	2009 a 2013	Climaticos	Ministério da saúde
(Dharmawardana et al., 2018)	Sri Lanka	2012 a 2017	Epidemiológicos e de telefonia móvel	Centro nacional de controle a dengue
(Puengpreeda, Yhusumrarn, Sirikulvadhana, 2020)	Tailândia	2014 a 2018	Epidemiológicos e Google *Trend topics*	Google e Departamento de meteorologia

Fonte: Autores (2021).

Na sua grande maioria (79%), as pesquisas foram realizadas nos países da Ásia. O restante (11%) correspondeu a estudos realizados nas Américas. Com exceção do trabalho conduzido por (Mishra, Tiwari, Ajaymon, 2019), os artigos informaram o período de anos utilizado em seus modelos de predição. Ainda sobre os dados, a grande parte deles utilizou dados epidemiológicos ou climáticos fornecidos por órgãos oficiais do país onde o estudo foi conduzido. Os trabalhos de (Dharmawardana et al., 2018) e de (Puengpreeda, Yhusumrarn, Sirikulvadhana, 2020) chamam atenção por inovar e utilizar, respectivamente, dados de telefonia móvel e de pesquisas realizadas no Google.

Os resultados obtidos pelos estudos, qual abordagem foi utilizada para mensurar a taxa de erro, os resultados para cada técnica e, enfim, a técnica vencedora, estão listados no Quadro 3. Este quadro apresenta as repostas aos questionamentos 4 ("Como foi feita a validação dos modelos? Quais técnicas estatísticas o estudo utilizou na avaliação?") e 5 ("Qual técnica ou combinação de técnicas obtiveram os melhores resultados?").

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Quadro 3 - Quadro contendo a forma de validação, resultados e técnica vencedora de cada estudo.

Estudo	Forma de Validação	Resultados	Vencedor
(Manogaran & Lopez, 2018)	RMSE	MR – RMSE 0,525 GPR – RMSE 0,281 SVR – RMSE 0,352 RF – RMSE 0,323	GPR
(Appice et al., 2020)	RMSE	AutoTiC-NN – RMSE 5,18 KNN – RMSE 8,20 SVR – RMSE 19,62 ARIMA – RMSE 12,23 M5- RMSE 115,34	AutoTiC-NN
(Dhaka & Singh, 2020)	*Sum of Absolute Difference* (SAD)	RF – SAD 91560,56 DT – SAD 101911,00 MLR – SAD 80901,42 SVR – SAD 11376,1	MLR
(Mishra, Tiwari, Ajaymon, 2019)	*Mean absolute error* (MAE)	NN – MAE 25,621 RF – MAE 25,012 *Boosted Trees* – MAE 24,985 LASSO – MAE 27,045 Ridge – MAE 28,052 XGBoost – MAE 24,802 SVR – MAE 25,011	XGBoost
(Guo et al., 2017)	RMSE	SVR – RMSE 0,2681 LASSO – RMSE 2,0621 GAM – RMSE 4,4973 GBM – RMSE 3,4527	SVR
(Raju et al., 2019)	MAE	SVR – MAE 180,61 Ridge – MAE 366,570 LR – MAE 190,04	SVR
(Xu et al., 2020)	RMSE	LSTM – RMSE 36,50 BPNN – RMSE 48,61 GAM – RMSE 41,95 SVR – RMSE 44,37 GBM – RMSE 42,33	LSTM
(Kerdprasop, Kerdprasop, Chuaybamroong, 2019)	Erro preditivo	CHAID – Erro preditivo 0,275 LR – Erro preditivo 0,598 GLM – Erro preditivo 0,598 ANN – Erro preditivo 0,901 SVR – Erro preditivo 1,034	CHAID
(Pham et al., 2018)	RMSE	GA-RNN – RMSE 13,06 LR – RMSE 22,99 DT – RMSE 34,89	GA-RNN
(Mussumeci & Codeço Coelho, 2020)	RMSE	LSTM – RMSE 0,45 RF – RMSE 0,47 LASSO – RMSE 0,50	LSTM
(Doni & Sasipraba, 2020)	RMSE	LSTM – RMSE 42,00 SVR – RMSE 49,00 XGboost – RMSE 48,00 RF – RMSE 51,00 GAM – RMSE 53,00 BPNN – RMSE 48,00	LSTM
(Carvajal et al., 2018)	RMSE	RF– RMSE 0,29 GAM – RMSE 0,33 GB – RMSE 0,30	RF
(Dharmawardana et al., 2018)	RMSE	ANN – RMSE 0,67 XGBoost – RMSE 0,54	XGBoost
(Puengpreeda, Yhusumram, Sirikulvadhana, 2020)	MAE	RF – MAE 10,98 Ridge – MAE 16,44	RF

Fonte: Autores (2021).

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Com exceção de (Raju et al., 2019), as melhores técnicas produzem resultados de previsões bem similares aos valores reais. As técnicas SVR, RF, LSTM obtiveram os melhores resultados e venceram em 50% dos artigos estudados. Vale ressaltar que, para todas as técnicas de medição de erro aqui abordadas, quanto menor for o seu valor melhor é o resultado da previsão.

Não foi encontrado na literatura um consenso de quais técnicas ou algoritmos devem ser utilizados durante as atividades de previsão. Inúmeros fatores influenciam positivamente ou negativamente nos resultados de um modelo de previsão. Entre eles, pode-se citar a quantidade de dados, como eles são tratados e quais foram utilizados para a previsão. Contudo, os resultados desta revisão apontam, que, ao menos, uma das técnicas RF, SVR e LSTM são utilizadas em 85% das pesquisas.

As técnicas de *Deep Learning*, como é o caso da LSTM, vêm ganhando espaço nos problemas de previsão frente às demais técnicas de *Machine Learning*. Cientistas atribuem o sucesso da LSTM devido à capacidade de possuir uma memória sobre o que já foi calculado, bem como capacidade de decidir quais dados precisam ser utilizados no futuro ou esquecidos (Doni & Sasipraba, 2020). Essa capacidade credencia as LSTMs na utilização de modelagem de problemas complexos como a previsão de dengue, visto que, vários fatores contribuem para a proliferação da doença: climáticos, sanitários, econômicos e sociais.

Em relação aos locais dos estudos desta revisão, eles foram realizados em países da Ásia e Américas. Embora exista uma disparidade espacial entre os continentes, as características do clima e chuva se assemelham: altos níveis de precipitações, temperaturas elevadas e elevados níveis de umidade. Combinando com registros epidemiológicos, os dados climáticos mostraram a sua capacidade em participar da criação de modelos de previsão de dengue.

Ademais, o trabalho conduzido por (Dharmawardana et al., 2018) mostra uma abordagem interessante sobre a migração de pessoas entre os países e a sua capacidade de proliferação de doenças. O estudo foi conduzido em 2018 e, a partir de 2020, o fenômeno foi bastante observado durante a epidemia do coronavírus. Sendo assim, fica a pergunta: quantas vidas poderiam ter sido salvas na epidemia do COVID19 caso medidas sanitárias e efetivos rastreios da doença fossem feitos antes da sua infestação pandêmica?

Em média, as previsões foram feitas com 10 anos de informação base e a coleta dos dados foi realizada em fontes oficiais dos governos daqueles países. Logo, problemas relacionados à qualidade ou viés dos dados, podem ser minimizados, pois os dados foram oriundos de fontes oficiais e em uma quantidade significativa. Sobre a significância estatísticas dos resultados obtidos, todos os trabalhos aprovados possuem técnicas de validação estatística atreladas à análise dos seus resultados. Artigos sem validação estatística ou com baixa qualidade no reporte foram excluídos desta revisão.

Outro achado da revisão é a grande aceitação da técnica RMSE para a validação dos resultados previstos pelos modelos, sendo utilizado em 64% dos artigos. Adicionalmente, a utilização do RMSE demonstra com mais fidelidade as discrepâncias entre o resultado previsto versus o resultado esperado (Carvajal et al., 2018).

4. Considerações Finais

A partir desta revisão, pode-se inferir que é possível prever, com baixa taxa de erro, casos de dengue através de técnicas de *Machine Learning* e de *Deep Learning*. A grande maioria dos estudos envolvendo ML e DL na previsão de dengue ocorreu em países asiáticos, embora também tenhamos trabalhos nas américas.

Apesar da existência de uma gama de técnicas de ML e de DL, podemos destacar as técnicas RF, SVR e LSTM como recorrentes nos estudos. Embora cada estudo tenha a sua particularidade, vale destacar os ótimos resultados da LSTM. Sempre que usada, essa técnica saiu vitoriosa.

É perceptível também o padrão de utilização de dados com período de, em média, 10 anos. Outro destaque desta revisão é a padronização da utilização dos dados para a confecção dos modelos. Na maioria deles, foi usado dados históricos e

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

climáticos. Finalmente, em relação à validação estatística, os pesquisadores têm preferência por medir através do RMSE.

Ainda que esta pesquisa tenha sido ampla, realizada em bases de referências e conduzida por pesquisadores experientes, vale ressaltar que, ao realizar revisão com base em artigos publicados, os resultados produzidos pela revisão são direcionados por eles. Por fim, devido às similaridades entre as variantes dos vírus em circulação e semelhança entre os climas, boa parte dos trabalhos aqui listados podem ser reproduzidos no Brasil.

Como sugestões para trabalhos futuros, indicamos a realização da pesquisa em outras bases, como, por exemplo, a *Springer* e realizar variações na *string* de busca. Por fim, sugerimos uma nova execução do protocolo definido por este artigo com intuito de verificar as novas soluções utilizadas para a predição de casos de dengue.

Referências

Appice, A., Gel, Y. R., Iliev, I., Lyubchich, V. & Malerba, D. (2020). A multi-stage machine learning approach to predict dengue incidence: a case study in mexico. *Ieee access*, 8, 52713–52725.

Brasil. (2019). Ministério da saúde alerta para aumento de 149% dos casos de dengue no país. Ministério da saúde, brasil, p. 2020.

Câmara, F. P., Theophilo, R. L. G., Santos, G. T. D., Pereira, S. R. F. G., Câmara, D. C. P., & Matos, R. R. C. D. (2007). Estudo retrospectivo (histórico) da dengue no Brasil: características regionais e dinâmicas. *Revista da Sociedade Brasileira de Medicina Tropical*, *40*, 192-196.

Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D. M., & Watanabe, K. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC infectious diseases*, *18*(1), 1-15.

da Silveira, L. T. C., Tura, B., & Santos, M. (2019). Systematic review of dengue vaccine efficacy. *BMC infectious diseases*, *19*(1), 1-8.

de Jesus, J. G., Dutra, K. R., Sales, F. C. D. S., Claro, I. M., Terzian, A. C., Candido, D. D. S., & Faria, N. R. (2020). Genomic detection of a virus lineage replacement event of dengue virus serotype 2 in Brazil, 2019. *Memórias do Instituto Oswaldo Cruz*, 115.

de Souza, R. F., & da Cunha Albuquerque, A. R. (2018). Geografia Da Dengue: Uma Análise Das Políticas De Controle E Monitoramento Do Aedes Aegypti Em Manaus/Geography of Dengue: an analysis of the control and monitoring policies of Aedes aegypti in Manaus. *Revista Geonorte*, *9*(31), 68-76.

Dhaka, A., & Singh, P. (2020, January). Comparative Analysis of Epidemic Alert System using Machine Learning for dengue and Chikungunya. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 798-804). IEEE.

Dharmawardana, K. G. S., Lokuge, J. N., Dassanayake, P. S. B., Sirisena, M. L., Fernando, M. L., Perera, A. S., & Lokanathan, S. (2017, December). Predictive model for the dengue incidences in Sri Lanka using mobile network big data. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)* (pp. 1-6). IEEE.

Doni, A. R., & Sasipraba, T. (2020). LSTM-RNN Based Approach for Prediction of Dengue Cases in India. *Ingénierie des Systèmes d'Information*, *25*(3).

Fabbri, S., Silva, C., Hernandes, E., Octaviano, F., Di Thommazo, A., & Belgamo, A. (2016, June). Improvements in the StArt tool to better support the systematic review process. In *Proceedings of the 20th international conference on evaluation and assessment in software engineering* (pp. 1-5).

Ferreira, V. M., Nunes, R. C., Ferreira, J. M. S., & Herrera, K. M. S. (2019). Um mosquito e três doenças: ação de combate ao Aedes aegypti e conscientização sobre Dengue, Chikungunya e Zika em Divinópolis/MG, BRASIL. *Revista Brasileira de Extensão Universitária*, *10*(2), 49-54.

Graciano, A. R., de Assis, L. P. F., Cezar, A. M., Amâncio, V. C., & de Oliveira, J. M. R. (2017). Morbimortalidade da dengue em idosos no Brasil-Dengue morbidity and mortality in elderly in Brazil. *Revista Educação em Saúde*, *5*(1), 56-65.

Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., & Ma, W. (2017). Developing a dengue forecast model using machine learning: A case study in China. *PLoS neglected tropical diseases*, *11*(10), e0005973.

Kardprasop, K., Kardprasop, N., & Chuaybamroong, P. (2019, December). Forecasting Dengue Incidence with the Chi-squared Automatic Interaction Detection Technique. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference* (pp. 37-42).

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implementation science*, *5*(1), 1-9.

Manogaran, G., & Lopez, D. (2018). A Gaussian process based big data processing framework in cluster computing environment. *Cluster Computing*, *21*(1), 189-204.

Mishra, V. K., Tiwari, N., & Ajaymon, S. L. (2019, December). Dengue disease spread prediction using twofold linear regression. In *2019 IEEE 9th International Conference on Advanced Computing (IACC)* (pp. 182-187). IEEE.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, *6*(7), e1000097.

Mussumeci, E., & Coelho, F. C. (2020). Large-scale multivariate forecasting models for Dengue-LSTM versus random forest regression. *Spatial and Spatio-temporal Epidemiology*, *35*, 100372.

Research, Society and Development, v. 10, n. 10, e33101119347, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19347

Mustafa, M. S., Rasotgi, V., Jain, S., & Gupta, V. J. M. J. A. F. I. (2015). Discovery of fifth serotype of dengue virus (DENV-5): A new public health dilemma in dengue control. *Medical journal armed forces India*, *71*(1), 67-70.

Neto, A. S. L., do Nascimento, O. J., & de Sousa, G. D. S. (2016). Dengue, zika e chikungunya-desafios do controle vetorial frente à ocorrência das três arboviroses-parte I. *Revista Brasileira em Promoção da Saúde*, *29*(3), 305-312.

Pham, D. N., Aziz, T., Kohan, A., Nellis, S., Khoo, J. J., Lukose, D., & Ong, H. H. (2018, October). How to efficiently predict dengue incidence in kuala lumpur. In *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)* (pp. 1-6). IEEE.

Puengpreeda, A., Yhusumran, S., & Sirikulvadhana, S. (2020). Weekly Forecasting Model for Dengue Hemorrhagic Fever Outbreak in Thailand. *Engineering Journal*, *24*(3), 71-87.

Raju, N. G., Krishna, P. G., Manognya, K., Kiran, G. R., Rohit, P., & Likhith, K. (2019, July). Evolution of predictive model for Dengue incidence by using machine learning algorithms. In *2019 International Conference on Communication and Electronics Systems (ICCES)* (pp. 51-59). IEEE.

Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L., & Liu, Q. (2020). Forecast of dengue cases in 20 Chinese cities based on the deep learning method. *International journal of environmental research and public health*, *17*(2), 453.