Comparison of Machine Learning Models for Total Dengue Cases Prediction

Thiago M. Carvalho[1], **Gabriel L. Tenório**[1], **Karla Figueiredo**[2], **Marley Vellasco**[1], **Wouter Caarls**[1]

[1]Departamento de Engenharia Elétrica – Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) - Rio de Janeiro - RJ - Brazil

[2]Departamento de Ciências da Computação - UERJ – Universidade do Estado do Rio de Janeiro (UERJ) – Rio de Janeiro– RJ –Brazil

{tmedeiros,gbrtenorio,marley,wouter}@ele.puc-rio.br

karlafigueiredo@ime.uerj.br

Abstract. *Dengue is an endemic disease with high prevalence in tropical areas, due to transmission by mosquitoes. Through preprocessing methods and machine learning algorithms, this work aims to develop predictive models for total dengue cases using climatic variables, as part of the 'DengAI-predicting disease spread' competition, hosted by DrivenData. Among all algorithms implemented, the Ensemble method, using Random Forest and Neural network, outperformed the proposed Benchmark, improving the results by 4.5%.*

Resumo. *A dengue é uma doença endêmica que ocorre principalmente em áreas tropicais, devido à sua transmissão através de mosquitos. Usando mecanismos de pré-processamento e de aprendizado de máquina, esse trabalho objetiva desenvolver um modelo de previsão que estabeleça uma relação existente entre as condições de uma cidade e a proliferação de epidemia de dengue, como parte da competição 'DengAI - predicting disease spread', fornecida pela plataforma DrivenData. Dentre os modelos implementados, o método Ensemble entre o Random Forest e Redes Neurais obtiveram a melhor performance, com melhora de 4,5% em relação ao Benchmark.*

1. Introdução

Atualmente, a dengue é considerada uma doença de caráter endêmico, sendo assunto de saúde pública em diversos países [Focks et al. 1995]. Essa doença é transmitida através de mosquito, sendo o *Aedes Aegypt* a espécie mais conhecida. Além da dengue, o *Aedes Aegypt* também é responsável pela proliferação de outras doenças associadas à dengue, como Chikungunya, Zika e Febre Amarela.

Em geral, existem condições climáticas que favorecem a presença e proliferação de mosquitos. Áreas úmidas e quentes, por exemplo, são conhecidas por uma presença maior de mosquitos [Lambrechts et al. 2011]. Consequentemente, países tropicais são mais propensos a epidemias transmitidas por esses vetores, como é o caso da dengue.

Além disso, para que o mosquito seja infectado pelo vírus da dengue e realize a proliferação da doença, são necessários alguns dias de maturação. Segundo estudos, entre

8 e 12 dias após o contato com o vírus, o mosquito já é capaz de infectar um humano através da picada [Rodhain 1997]. Além disso, após a instalação do vírus no mosquito, os ovos gerados podem já ser infectados pela dengue, chegando à forma adulta apto para transmitir a doença.

Por mais que se conheça o ciclo de vida do mosquito e suas preferências climáticas, a modelagem populacional do mosquito ainda é muito dependente da região em que se encontra. Condições climáticas da região e o índice de urbanização, por exemplo, são fatores que podem influenciar na dinâmica da população nos mosquitos e a proliferação de doenças como a dengue [Kuno 1997].

O presente trabalho visa a implementação de modelos de regressão usando *Machine Learning* para a participação da competição '*DengAI - Predicting Disease Spread*'[1]. Esta competição tem como objetivo prever, semanalmente, o número de casos de dengue nas cidades de Iquitos (Peru) e San Juan (Porto Rico). Através de uma base de dados contendo múltiplos dados coletados semanalmente, como temperatura, precipitação e umidade, foram implementados modelos capazes de inferir a previsão dos casos de dengue nestas duas cidades.

Este trabalho é dividido em 6 seções. Na próxima seção serão discutidas as técnicas encontradas na literatura para o problema de previsão de doenças endêmicas. Na terceira seção, é descrito o problema proposto neste trabalho. A quarta seção discorre sobre a abordagem realizada para a resolução do problema, contendo as análises de pré-processamento realizadas e os algoritmos utilizados para a previsão de casos de dengue. Em seguida, na quinta seção são apresentados os resultados obtidos. Por fim, a última seção trata das conclusões e trabalhos futuros propostos.

2. Revisão Bibliográfica

A modelagem da dinâmica do mosquito e a propagação da doença é um assunto bastante pesquisado atualmente. Diferentes estudos nessa área, através de modelos baseados em Machine Learning, visam obter e melhorar um modelo preditivo de dengue para diferentes regiões.

Em [Scavuzzo et al. 2018], são propostos algoritmos de *Machine Learning*, como *Support Vector Machine* (SVM), *Multi-Layer Perceptron* (MLP) e *K-Nearest Neighbors* (KNN), para a modelagem de uma população de mosquitos na província de Salta, Argentina. Em [Kwon et al. 2015], é proposto um modelo utilizando MLP para a previsão de população de mosquitos urbanos na Coréia do Sul.

[Halide and Ridd 2008] incorporam variáveis climáticas, como temperatura, precipitação e umidade, em modelo estatístico para a previsão de casos de dengue hemorrágica na cidade de Makassar, Indonésia.

Além disso, em [Gubler et al. 2001] são utilizados fatores sociais e demográficos para uma obtenção de melhores resultados na previsão de casos de dengue.

O estudo de fenômenos da natureza também é estudado como um fator de grande relevância na propagação de vírus como a dengue. Em [Fuller et al. 2009], foi estudado a

[1]DengAI - Predicting Disease Spread. Disponível em: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/ [2 de Julho, 2019]

influência do fenômeno El Niño e índices de vegetação para a previsão de casos de dengue na cidade da Costa Rica.

Em [Sathler and Luciano 2017], foram utilizados diversos modelos, como Redes Neurais Recorrentes, Redes Neurais *Feed-Forward* e regressão bayesiana.

Em [Chua et al. 2017] são incluídos outros métodos de pré-processamento, como correlação cruzada. Além disso, são também propostos algoritmos, como *Multiple Linear Regression* (MLR), para a previsão de casos de dengue.

3. Descrição do problema

Este trabalho é baseado na competição '*DengAI - Predicting Disease Spread*' visando a previsão de casos de dengue para as cidades de San Juan e Iquitos. A avaliação dos resultados da previsão é feita através da métrica MAE (Erro Médio Absoluto).

Usando métodos de pré-processamento e implementação de algoritmos de aprendizado, este trabalho tem como objetivo implementar modelos preditivos para melhorar a previsão de casos de dengue para a cidade de San Juan e Iquitos.

A competição estabelece um *Benchmark* para efeitos comparativos, utilizando algoritmo de Regressão Binomial para a previsão de casos de dengue. Desta forma, este trabalho também visa o desenvolvimento de algoritmos que obtenham performance melhor do que o *Benchmark* da competição.

4. Abordagens realizadas

4.1. Base de Dados

Os dados utilizados neste trabalho são disponibilizados pela plataforma *DrivenData* e estão divididos em três arquivos. O primeiro arquivo contém informações climáticas e geográficas de ambas as cidades, totalizando 23 atributos (divididos em atributos de temperatura, como temperatura do ar e temperatura de ponto de orvalho, atributos de precipitação e atributos de vegetação). Para a cidade de San Juan foram coletados dados semanais de 1990 a 2008, totalizando 936 registros. Já para a cidade de Iquitos, estão disponíveis dados de 2000 a 2010, totalizando 520 registros. A Tabela 1 mostra as variáveis disponíveis, com uma breve descrição.

Além disso, o site *DrivenData* disponibiliza um arquivo correspondente aos resultados de casos totais de dengue em cada uma das semanas do treinamento. Esses dados são os '*targets*' para o nosso problema. Por fim, o terceiro arquivo disponível corresponde aos atributos para a previsão de casos de dengue. No total, são 262 registros disponíveis para a cidade de San Juan. Logo, para a competição deve-se realizar a previsão de casos totais de dengue para 262 semanas.

4.2. Pré-processamento

Baseado no estudo teórico acerca do ciclo de vida do mosquito e nos dados disponíveis, segue-se com a análise de dados.

4.2.1. Valores faltantes

Conforme observado anteriormente, os dados disponibilizados contém dados da cidade de San Juan e de Iquitos. Contudo, as cidades possuem características climáticas diferentes e,

Tabela 1. Descrição dos atributos das bases para San Juan e Iquitos

Tipo de atributo	Descrição	Nome do atributo
Datas	Data do início da semana	week-start-date
	Semana do ano	week-of-year
	Ano	year
Dados climáticos	Temperatura máxima	station-max-temp_c
	Temperatura mínima	station-min-temp-c
	Temperatura média	station-avg-temp-c
	Total de precipitação	station-precip-mm
	Variação de temperatura diurna	station-diur-temp-rng-c
Precipitação via satélite	Total de precipitação	precipitation-amt-mm
Climate Forecast System Reanalysis	Total de precipitação	reanalysis-sat-precip-amt-mm
	Temperatura média de ponto de orvalho	reanalysis-dew-point-temp-k
	Temperatura média do ar	reanalysis-air-temp_k
	Umidade relativa média	reanalysis-relative-humidity-percent
	Umidade específica média	reanalysis-specific-humidity-g-per-kg
	Total de precipitação	reanalysis-precip-amt-kg-per-m2
	Temperatura máxima do ar	reanalysis-max-air-temp-k
	Temperatura mínima do ar	reanalysis-min-air-temp-k
	Temperatura média do ar	reanalysis-avg-temp_k
	Variação de temperatura diurna	reanalysis-tdtr-k
Índice de diferença de vegetação	Centróide do pixel da região sudeste	ndvi-se
	Centróide do pixel da região sudoeste	ndvi-sw
	Centróide do pixel da região nordeste	ndvi-ne
	Centróide do pixel da região noroeste	ndvi-nw

portanto, o ciclo de vida do mosquito e a dinâmica de proliferação são diferentes. Deste modo, os dados foram separados em dois conjuntos, um para cada cidade. Portanto, a análise dos atributos será feita separadamente.

O primeiro problema observado foram os valores faltantes de alguns atributos. No total, existem 380 valores não preenchidos para a cidade de San Juan, nos quais os atributos 'ndvi_ne' e 'ndvi_nw' podem ser destacados com 191 e 49 dados faltantes, respectivamente. Já para a cidade de Iquitos, 168 valores não são disponibilizados, sendo os atributos 'station_avg_temp_c' e 'station_diur_temp_rng_c' com maior quantidade de dados faltantes.

Dois métodos de preenchimento de dados faltantes foram testados: *Forward Fill* e *Attribute Average* [Bennett 2001]. A primeira abordagem propaga o último valor válido do atributo para o próximo, e portanto os valores faltantes são preenchidos com o dado válido mais recente. O segundo método calcula a média para cada um dos atributos, preenchendo os valores faltantes com as médias calculadas. Neste trabalho, o primeiro método foi escolhido, visto que é razoável supor que as condições climáticas da semana no qual existem valores faltantes são parecidas com o clima da semana anterior.

4.2.2. *Outliers* e normalização

Além disso, foi gerado também o histograma e um *boxplot* para verificação de valores extremos. Neste caso, a exclusão destes valores não é previsto, pois estes casos podem ser importantes para a modelagem do problema. A Figura 1 mostra o *boxplot* e o histograma para as variáveis de precipitação e casos totais de dengue.

A observação da distribuição dos dados auxilia no tipo de normalização a ser realizada para obter uma melhor previsão. Neste caso, para variáveis que possuem um comportamento de cauda longa, como presente na Figura 1, foi utilizada uma

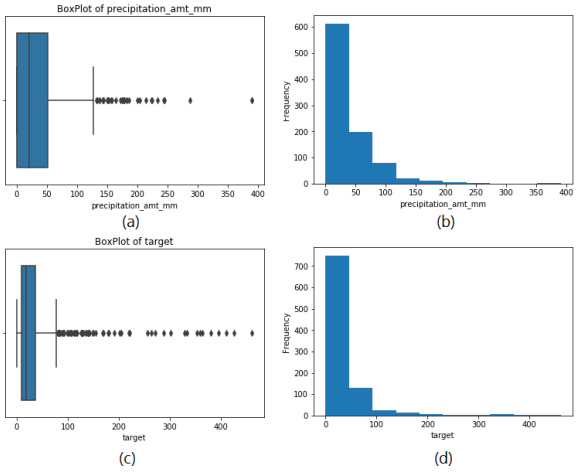

Figura 1. Boxplots: (a) e (c); Histogramas (b) e (d) para as variáveis precipitação e casos totais de dengue na cidade de San Juan

normalização por partes. Para as variáveis que possuem uma distribuição mais uniforme, foi utilizada a normalização min-max.

4.3. Seleção de variáveis

Após a observação individual dos atributos relativos ao problema, foram estabelecidos alguns métodos para seleção de variáveis. Esta subseção visa eliminar as variáveis que são redundantes ou potencialmente irrelevantes para este problema.

O primeiro método de seleção de variáveis utilizado foi a correlação de Pearson [Benesty et al. 2009]. Este método verifica o grau de correlação entre duas variáveis, onde o coeficiente próximo de 1 significa uma forte correlação positiva e -1 uma forte correlação negativa. A Figura 2 mostra o resultado obtido pela correlação de Pearson.

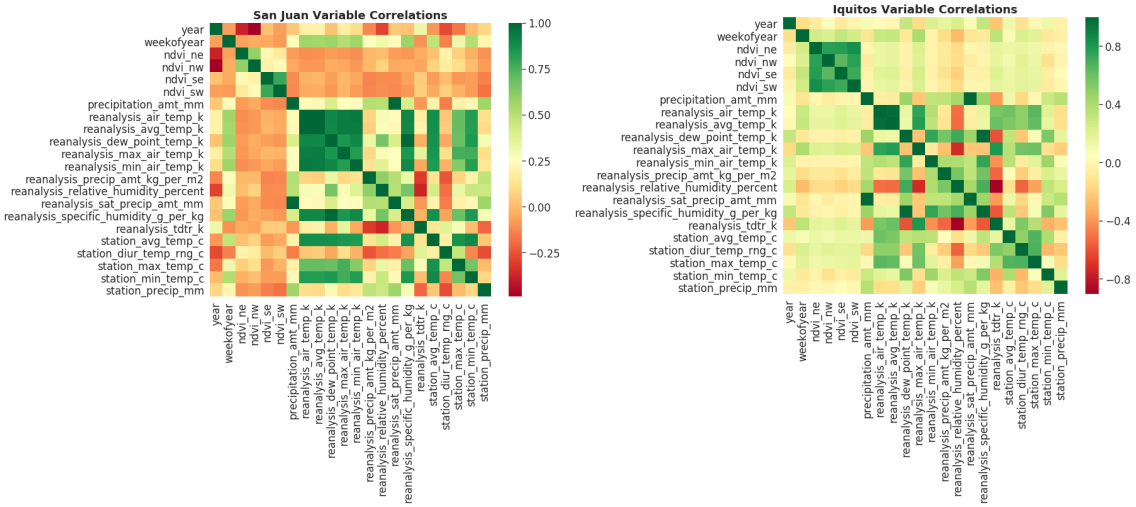

Figura 2. Correlação de Pearson para a cidade de San Juan (esq.) e Iquitos (dir.)

Analisando o resultado, é possível observar uma forte correlação entre algumas variáveis do problema. Este método auxilia a indicar variáveis que são redundantes neste problema e que posteriormente podem ser retiradas para o treinamento do algoritmo.

Neste trabalho, as variáveis com graus de correlação maiores que 0.9 foram consideradas redundantes.

Na sequência, foram implementados paralelamente outros dois métodos de seleção de variáveis, a saber: *Recursive Feature Elimination* (RFE) [Granitto et al. 2006] e *Least Squared Estimator* (LSE) [Marquardt 1963].

O método RFE foi utilizado para selecionar os 6 atributos mais relevantes. Esse algoritmo efetua um ranqueamento das variáveis selecionadas, mostrando numericamente a relevância de cada atributo para o problema.

O método LSE avalia o comportamento da variação da saída Δy em função das variações das entradas Δx do problema. Conforme pode ser observado pela Eq. 1, este método calcula os coeficientes b_i, que indicam a relevância do atributo x_i em função da saída y.

$$\Delta y = \sum_{i=1}^{m} \mathbf{b_i} \mathbf{\Delta x_i}, \tag{1}$$

Para uma melhor análise dos resultados obtidos na seleção de variáveis com os algoritmos implementados, a Tabela 2 mostra os atributos com melhores avaliações.

Tabela 2. Atributos com as 6 melhores avaliações para os métodos RFE e LSE em ordem de relevância.

Método de seleção	RFE		LSE	
Cidade	San Juan	Iquitos	San Juan	Iquitos
Atributos	ndvi_se	ndvi_ne	reanalysis_specific_humidity_g_per_kg	reanalysis_max_air_temp_k
	ndvi_sw	ndvi_nw	reanalysis_air_temp_k	reanalysis_tdtr_k
	ndvi_nw	ndvi_se	year	weekofyear
	reanalysis_dew_point_temp_k	ndvi_sw	weekofyear	ndvi_ne
	reanalysis_max_air_temp_k	reanalysis_avg_temp_k	ndvi_se	reanalysis_dew_point_temp_k
	realanysis_specific_humidity_g_per_kg	reanalysis_specific_humidity_g_per_kg	precipitation_amt_mm	ndvi_se

Baseado nos resultados do processo de pré-processamento, é possível observar que alguns atributos possuem grande redundância (Fig. 2) e outros aparentam ter maior relevância (Tab. 2). Deste modo, os seguintes atributos foram selecionados, para ambas as cidades: quatro atributos relacionados a vegetação (ndvi_ne, ndvi_nw, ndvi_se, ndvi_sw), um atributo de precipitação (precipitation_amt_mm), três atributos de temperatura (reanalysis_dew_point_temp_k, reanalysis_max_air_temp_k e reanalysis_tdtr_k), um atributo relacionado a umidade (reanalysis_specific_humidity_g_per_kg) e a semana do ano (weekofyear).

4.4. Algoritmos de regressão

Para este trabalho, foram pesquisados alguns modelos capazes de obter desempenhos satisfatórios em problemas de regressão. Deste modo, foram selecionados diferentes modelos, a saber: *K-Nearest Neighbors* [Dudani 1976], *Random Forest*[Breiman 2001], Redes Neurais MLP [Gardner and Dorling 1998], *Bayesian Ridge Regressor* [Shi et al. 2016] e *Facebook's Prophet*[2]. Posteriormente, foi realizado um teste utilizando uma combinação entre dois algoritmos, o *Random Forest* e Redes Neurais,

[2]Prophet: Automatic Forecasting Procedure. Disponível em: https://github.com/facebook/prophet [2 de Julho, 2019]

de forma a tentar obter um desempenho melhor. Com exceção do algoritmo *Facebook's Prophet*, todos os outros modelos utilizam como atributos de entrada as variáveis selecionadas destacadas na etapa de pré-processamento.

Antes da implementação de cada um dos algoritmos, foi realizado uma divisão da base de dados com 85% dos dados para o conjunto de treinamento e 15% dos dados para o conjunto de validação. Esta divisão é utilizada para treinar os modelos com apenas parte da base de dados e avaliar suas performances através do conjunto de validação.

A subseção a seguir descreve cada um destes modelos utilizados neste trabalho.

4.4.1. *K-Nearest Neighbors* (KNN)

O KNN é um algoritmo comumente utilizado em problemas de classificação e regressão devido a sua simplicidade e baixa complexidade computacional. Neste trabalho, o algoritmo KNN foi utilizado para a tarefa de regressão. Desta forma, o KNN busca, na base de dados de treinamento, os K registros mais próximos para o cômputo da quantidade prevista de casos de dengue. A saída da regressão é a média dos casos de dengue para os K vizinhos mais próximos.

4.4.2. *Random Forest* (RF)

Random Forest é um método Ensemble amplamente utilizado em problemas de classificação e regressão. Este método consiste na criação e treinamento de diversas árvores de decisão, nos quais as árvores mais fortes participam no processo de decisão da tarefa, provendo uma melhor performance se comparado a apenas uma árvore de decisão.

Existem alguns hiperparâmetros relevantes para o algoritmo RF, como o tamanho da floresta (número de estimadores), a profundidade das árvores e o *Bootstrap*. O tamanho da floresta contribui para a criação de uma regressão suave, enquanto que a profundidade das árvores evita problemas de *underfitting* e *overfitting*. Além disso, o *Bootstrap* é um importante operador que possibilita a escolha de dados aleatórios sem repetição durante os passos de treinamento.

4.4.3. Redes Neurais MLP

Redes Neurais possuem ampla aplicação em problemas de regressão. Deste modo, foi implementado uma Rede Neural MLP para a previsão de casos de dengue em San Juan e outra para Iquitos.

Em uma topologia de Redes Neurais, existem alguns hiperparâmetros que podem ser modificados para obter um melhor resultado. Neste trabalho, foram realizados testes de MLP com apenas uma camada escondida e com duas camadas escondidas. Para cada uma das duas configurações foram testados diferentes quantidades de neurônios nas camadas escondidas.

Além disso, foi modificada também a quantidade de entradas do problema, experimentando diferentes janelas temporais para os atributos. No total, foram utilizadas 4 variáveis com janelas temporais, sendo as variáveis de umidade e temperatura de ponto de orvalho com uma janela igual a 6, precipitação com janela igual a 3 e temperatura máxima do ar com janela igual a 3.

Por fim, foi utilizado o método de parada antecipada para evitar o *overfitting* da rede. Para este método, foi definido um valor de paciência de 40 épocas.

4.4.4. *Bayesian Ridge Regressor* (BRR)

O método BRR obtém um modelo probabilístico usado para resolver problemas de regressão. É incluído a regularização de parâmetros no processo de regressão, além da otimização de uma função de mínimos quadrados como função de perda. Alguns dos hiperparâmetros são α_1, α_2 e λ_1, nos quais são escolhidos para serem não-informativos, conforme descrito em [Shi et al. 2016].

4.4.5. *Facebook's Prophet* (FP)

Este modelo é usualmente utilizado para previsão de séries temporais que contém comportamento não-linear. Conforme descrito na documentação, esse modelo possui uma alta performance em problema de séries temporais no qual a sazonalidade é presente. Além disso, este modelo é robusto a *outliers* e valores faltantes.

A maior vantagem deste modelo é a necessidade de apenas um atributo para a previsão: a informação da data no formato Ano-Mês-Dia. Com isso, a informação do número total de casos passa a estar diretamente relacionada à sua data de ocorrência, desconsiderando possíveis variáveis cujos valores podem se mostrar bastantes ruidosos. Observando os atributos disponíveis, é possível perceber que a variável week_start_date corresponde ao atributo necessário para a utilização deste modelo.

Neste modelo, dois hiperparâmetros são relevantes: o nível de sazonalidade e a escala, que provém efeitos de regularização para reduzir o *overfitting*.

4.4.6. *Ensemble*

Ocasionalmente, apenas um modelo para realizar a previsão ou a classificação não é capaz de representar da melhor forma o problema inteiro. Diante desta limitação, uma combinação de resultados gerados por diferentes algoritmos pode captar de melhor forma o problema.

No presente trabalho, foi implementado um modelo Ensemble para a previsão de casos de dengue. Para tanto, foram combinados os resultados obtidos pela rede MLP e RF. Para o cômputo da saída, foi utilizado a média dos resultados dos modelos individuais.

4.5. Experimentos

Os experimentos realizados neste trabalho foram desenvolvidos na linguagem *Python*, possibilitando a implementação de todos os algoritmos descritos anteriormente.

De forma a obter uma melhor configuração para cada modelo, foram modificados os hiperparâmetros de cada um dos algoritmos utilizados, conforme pode ser observado na Tabela 3.

Em cada experimento, foi avaliado o impacto das variações dos hiperparâmetros, usando o conjunto de treinamento para obter os melhores modelos de regressão. Após a obtenção dos melhores modelos, utilizou-se o conjunto de validação para testar a generalização de cada modelo. A métrica utilizada para a avaliação da performance dos algoritmos foi o MAE, dada pela Eq. 2. Esta métrica também é utilizada pela competição para a avaliação da previsão dos algoritmos.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{2}$$

Tabela 3. Variação dos hiperparâmetros

Algoritmo	Hiperparâmetros
KNN	Tamanho da vizinhança: K = [1,2 ..., 100]
RF	Quantidade de estimadores = [8, 16, 32, 64, 128, 256, 512]
	Profundidade = [3,5,7]
	BootStrap = [True,False]
MLP	Camadas escondidas: [1,2]
	Neurônios por camada: [10, ..., 20]
	Taxa de Aprendizado = [0.1, 0.01, 0.001]
FP	Escala = [0.1,0.2]
	Sazonalidade = [5, ..., 10]
BRR	Tolerância = [0.1, 0.01, 0.001]
	Número de iterações = [100, ..., 500]

Onde f_i representa o valor real, y_i é o valor previsto pelo modelo e n é o número de instâncias.

Vale ressaltar que para os algoritmos KNN e RF, os hiperparâmetros foram modificados utilizando um método de *grid search*, no qual testa diferentes combinações de hiperparâmetros e retorna a configuração que obteve a melhor performance.

Após a obtenção da melhor configuração dos algoritmos, estes foram utilizados para a previsão de casos de dengue utilizando o conjunto de teste disponibilizado pela competição *DengAI*.

5. Discussão de resultados

Com o objetivo de comparar e avaliar a performance de cada um dos algoritmos, para o conjunto de treinamento, validação e teste, forma dispostos os melhores resultados obtidos, conforme mostra a Tabela 4. Diferentemente dos resultados de treinamento e validação, os resultados de teste são verificados de forma conjunta para ambas as cidades. Além disso, nesta tabela é disposta o erro médio absoluto para o *Benchmark* proposto pela competição, que utiliza o algoritmo de Regressão Binomial Negativo (RBN).

Tabela 4. Resultados para o conjunto de treinamento, validação e teste (submissão) para as cidades SJ e Iq.

Algoritmo	Erro (MAE)		
	Treinamento	Validação	Teste (DataDriven)
KNN	SJ: 27.09 ; Iq: 6.93	SJ: 25.02 ; Iq: 5.84	(SJ + Iq): 28.63
RF	SJ: 17.81 ; Iq: 5.23	SJ: 20.95 ; Iq: 5.32	(SJ + Iq): 24.93
MLP	SJ: 18.96 ; Iq: 3.58	SJ: 16.81 ; Iq: 5.40	(SJ + Iq): 25.40
FP	SJ: 36.24 ; Iq: 7.57	SJ: 34.79 ; Iq: 6.69	(SJ + Iq): 25.66
BRR	SJ: 16.72 ; Iq: 6.24	SJ: 23.66 ; Iq: 6.87	(SJ + Iq): 25.85
Ensemble	SJ: 17.33 ; Iq: 4.21	SJ: 19.50 ; Iq: 5.16	**(SJ + Iq): 24.56**
(Benchmark) RBN	-	-	(SJ + Iq): 25.82

Conforme pode ser observado, o algoritmo KNN não obteve um resultado satisfatório. Embora o erro obtido para o conjunto de validação não seja alto, o MAE obtido na previsão foi a maior entre os algoritmos propostos.

Para o modelo probabilístico usando BRR, é possível observar que a performance pode ser equiparada ao modelo de regressão binomial negativo.

O modelo utilizando RF obteve um resultado final melhor do que o modelo proposto pela competição. Conforme discutido em [Robnik-Šikonja 2004], o algoritmo RF ainda é um bom algoritmo para resolver problemas de regressão não-linear.

Além disso, é possível perceber que o modelo utilizando MLP também obteve melhor performance do que o algoritmo de *Benchmark* da competição. Um dos fatores que contribuíram para uma melhor performance deste algoritmo foi a inserção de valores de semanas anteriores, como precipitação e temperatura do ar, para a previsão de casos de dengue.

O algoritmo FP também obteve um resultado melhor do que o modelo proposto pela competição, mesmo utilizando apenas um atributo. Deste modo, é possível observar a robustez deste modelo em problemas de séries temporais, principalmente quando há sazonalidade.

É possível perceber que o melhor resultado obtido, neste trabalho, foi implementando uma combinação de resultados obtidos pelos algoritmos RF e MLP, isto é, o algoritmo Ensemble. Deste modo, é possível perceber que este método é capaz de fornecer um resultado mais estável. Este algoritmo obteve uma melhora de 4.5% em relação ao *Benchmark* da competição.

Para uma melhor verificação dos resultados para cada modelo, a Figura 3 mostra a previsão de casos de Dengue de todos os modelos implementados.

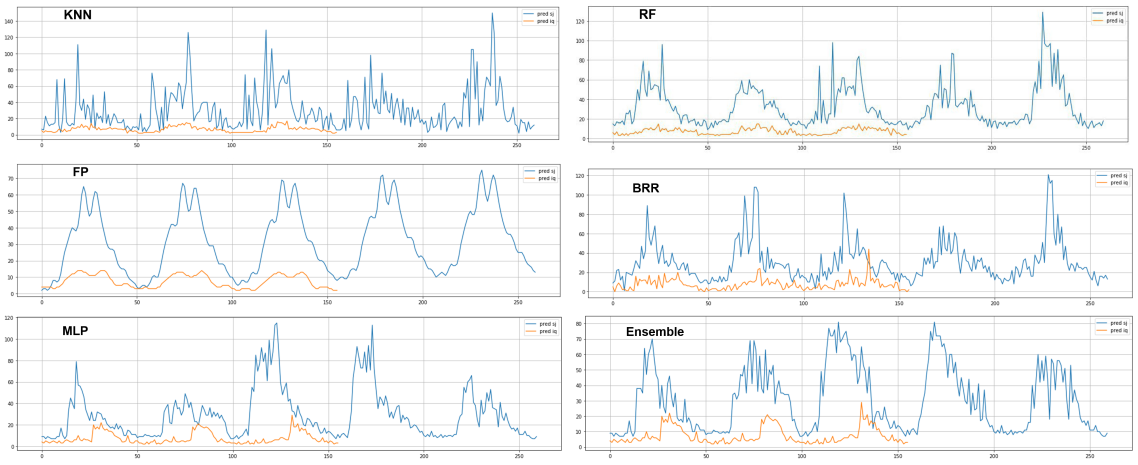

Figura 3. Gráficos das previsões para todos os algoritmos implementados.

Através dos gráficos, é possível perceber que os melhores resultados obtidos, usando RF e a combinação de resultados, possuem uma menor oscilação de previsão entre uma semana e outra.

Por fim, outra constatação observada é o resultado obtido pelo algoritmo FP, que possui repetições aproximadas para cada uma das cidades, com uma tendência

de crescimento dos valores. Esse efeito é decorrente da sazonalidade detectada pelo algoritmo.

6. Conclusão e Trabalhos Futuros

Através do trabalho realizado, é possível observar a relação existente entre as condições climáticas e vegetativas na previsão de casos de dengue, corroborando com os estudos teóricos sobre esse tema.

Além disso, os algoritmos propostos obtiveram resultados satisfatórios para o problema. O melhor algoritmo proposto neste trabalho está entre os 10% melhores resultados da competição (671 de 7053 competidores).

Contudo, existem algumas tarefas a serem realizadas como trabalho futuro. A primeira delas é obter uma melhor combinação de entradas que consigam explicar melhor o problema proposto, através de métodos de redução de variáveis, por exemplo.

Além disso, é esperado no futuro a implementação de novos modelos que possuem boa performance em trabalhos de regressão e previsão de séries, como *Long Short-Term Memory* [Gers et al. 1999] e *Undecimated Fully Convolutional Neural Networks* [Mittelman 2015].

Referências

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Bennett, D. A. (2001). How can i deal with missing data in my study? *Australian and New Zealand journal of public health*, 25(5):464–469.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chua, M., Deb, S., and Acebedo, C. M. (2017). An ensemble prediction approach to weekly dengue cases forecasting based on climatic and terrain conditions. *Journal of Health and Social Sciences*, 2:257–272.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4):325–327.

Focks, D. A., Daniels, E., Haile, D. G., and Keesling, J. E. (1995). A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results. *The American journal of tropical medicine and hygiene*, 53(5):489–506.

Fuller, D. O., Troyo, A., and Beier, J. C. (2009). El nino southern oscillation and vegetation dynamics as predictors of dengue fever cases in costa rica. *Environmental Research Letters*, 4(1):014011.

Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.

Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: Continual prediction with lstm.

Granitto, P. M., Furlanello, C., Biasioli, F., and Gasperi, F. (2006). Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90.

Gubler, D. J., Reiter, P., Ebi, K. L., Yap, W., Nasci, R., and Patz, J. A. (2001). Climate variability and change in the united states: potential impacts on vector-and rodent-borne diseases. *Environmental health perspectives*, 109(suppl 2):223–233.

Halide, H. and Ridd, P. (2008). A predictive model for dengue hemorrhagic fever epidemics. *International journal of environmental health research*, 18(4):253–265.

Kuno, G. (1997). Factors influencing the transmission of dengue viruses. *Dengue and dengue hemorrhagic fever*, 1:23–39.

Kwon, Y.-S., Bae, M.-J., Chung, N., Lee, Y.-R., Hwang, S., Kim, S., Choi, Y., and Park, Y.-S. (2015). Modeling occurrence of urban mosquitos based on land use types and meteorological factors in korea. *International journal of environmental research and public health*, 12(10):13131–13147.

Lambrechts, L., Paaijmans, K. P., Fansiri, T., Carrington, L. B., Kramer, L. D., Thomas, M. B., and Scott, T. W. (2011). Impact of daily temperature fluctuations on dengue virus transmission by aedes aegypti. *Proceedings of the National Academy of Sciences*, 108(18):7460–7465.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441.

Mittelman, R. (2015). Time-series modeling with undecimated fully convolutional neural networks. *arXiv preprint arXiv:1508.00317*.

Robnik-Šikonja, M. (2004). Improving random forests. In *European conference on machine learning*, pages 359–370. Springer.

Rodhain, F. R. (1997). Mosquito vectors and dengue virus-vector relationships. *Dengue and dengue hemorrhagic fever*, pages 45–60.

Sathler, C. and Luciano, J. (2017). Predictive modeling of dengue fever epidemics: A neural network approach.

Scavuzzo, J. M., Trucco, F., Espinosa, M., Tauro, C. B., Abril, M., Scavuzzo, C. M., and Frery, A. C. (2018). Modeling dengue vector population using remotely sensed data and machine learning. *Acta tropica*, 185:167–175.

Shi, Q., Abdel-Aty, M., and Lee, J. (2016). A bayesian ridge regression analysis of congestion's impact on urban expressway safety. *Accident Analysis & Prevention*, 88:124–137.