UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

FCT – Faculdade de Ciências e Tecnologia
DMC – Departamento de Matemática e Computação
Bacharelado em Ciência da Computação

INTRODUÇÃO À CIÊNCIA DE DADOS

PROJETO FINAL

ANÁLISE E MODELO PREDITIVO DE UM CONJUNTO DE DADOS DE CASOS DE DENGUE DO BRASIL DE 2001 A 2019

ALUNO: LUCAS BERNARDO DE SOUZA
PROFESSOR: PROF.º DANILO ROBERTO PEREIRA

Presidente Prudente
2024

Sumário

1 INTRODUÇÃO

O Brasil por ser um país tropical tem uma certa relação com uma doença característica desse tipo de região, que é a dengue. Contudo será mesmo que a dengue só se relaciona com o clima? Sabemos que água parada é o habitat para que as larvas do mosquito se reproduzem e que a temperatura da água não pode ser abaixo dos vinte graus celsius. Uma país onde a temperatura mínima média gira em torno dos 19 graus e com uma certa recorrência de chuvas em algumas regiões é o lugar perfeito para esse mosquito. Porém de acordo com as análises feitas por esse conjunto de dados o clima não tem tanto impacto assim na quantidade de casos de dengue como se imagina e foi inferido que outros fatores socioeconômicos como saneamento básico, coleta de lixo, entre outros podem estar influenciando as altas quantidades de casos em algumas regiões do país.

O objetivo é criar um modelo de regressão capaz de prever a quantidade de casos de dengue no futuro para que as cidades possam se preparar para períodos em que o modelo previu um maior número de casos. Evitando o que o sistema público de saúde sobrecarregue ou tenha recursos essenciais para o tratamento dessa patologia esgotados.

A base de dados utilizada foi retirada do Kaggle.

2 METODOLOGIA

Para criação do modelo preditivo primeiramente foi realizado uma *análise exploratória* com o intuito de conhecer o conjunto de dados, seus atributos, procurar dados nulos, ‘outliers’, entender o padrão dos dados e como ele se comporta. Por fim, é selecionada as features mais aptas que descrevem bem o conjunto de dados e que servem o propósito. Além de preparar aquelas que não estão adequadas para o treinamento do modelo.

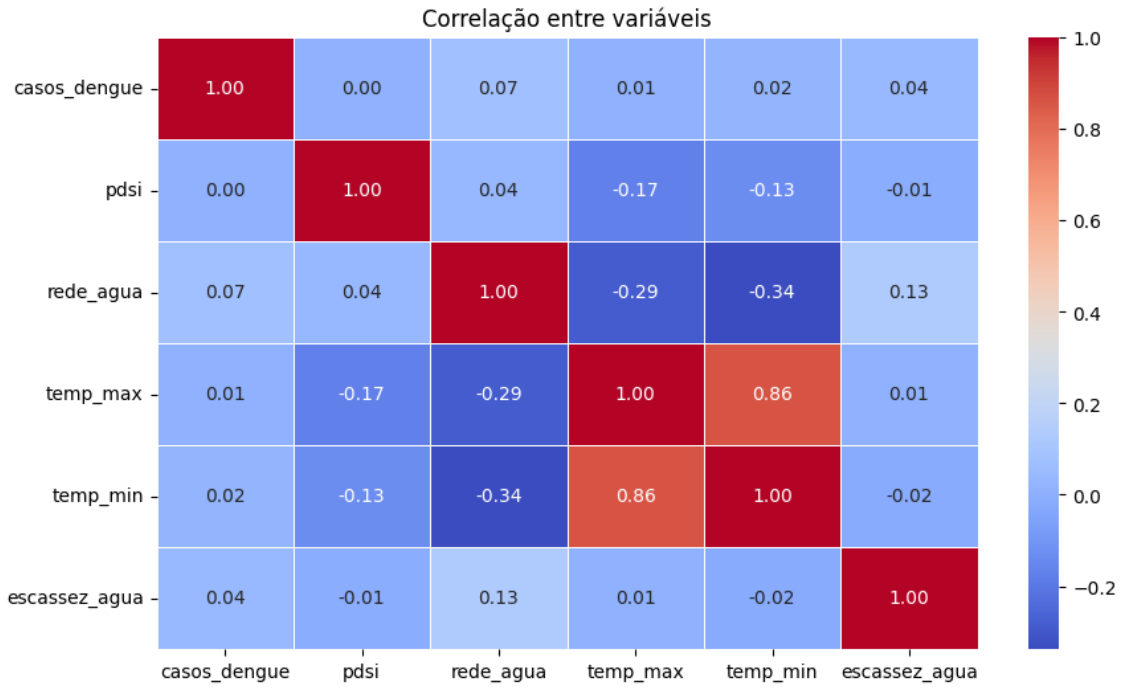
2.1 Análise Exploratória

O conjunto de dados em questão é formado por vinte e seis atributos e pouco mais de cento e trinta mil linhas. Ele é formado por dados de diversas cidades do Brasil, não pude atestar se de todas as cidades, mas temos muitas cidades.

Além da sua grande dimensão, o conjunto contém diversos atributos como nome e código das cidades, estado, região, bioma, ecozona, clima principal da cidade, temperatura mínima e máxima, índice de seca de Palmer (PDSI), rede de água, escassez, ano, mês e hora da coleta, casos de dengue, população e densidade populacional.

Os atributos casos de dengue, população e densidade populacional apresentaram uma taxa de 0.05% de dados faltantes. Como o conjunto de dados é grande e temos uma taxa muito baixa de dados nulos nesses atributos a solução aplicada foi a remoção das linhas que tinham esses dados nulos. Após a remoção o conjunto de dados passou a ter pouco mais de cento e vinte e cinco mil linhas.

Por fim, foi observado que os dados estão entre 2001 e 2019. A quantidade de casos de dengue que mais aparece no conjunto de dados está entre 0 e 29. A média de casos no Brasil entre esse período foi de aproximadamente cem casos, com um desvio padrão de mais de novecentos. As médias das temperaturas máximas e mínimas são de aproximadamente 29 e 19 respectivamente.

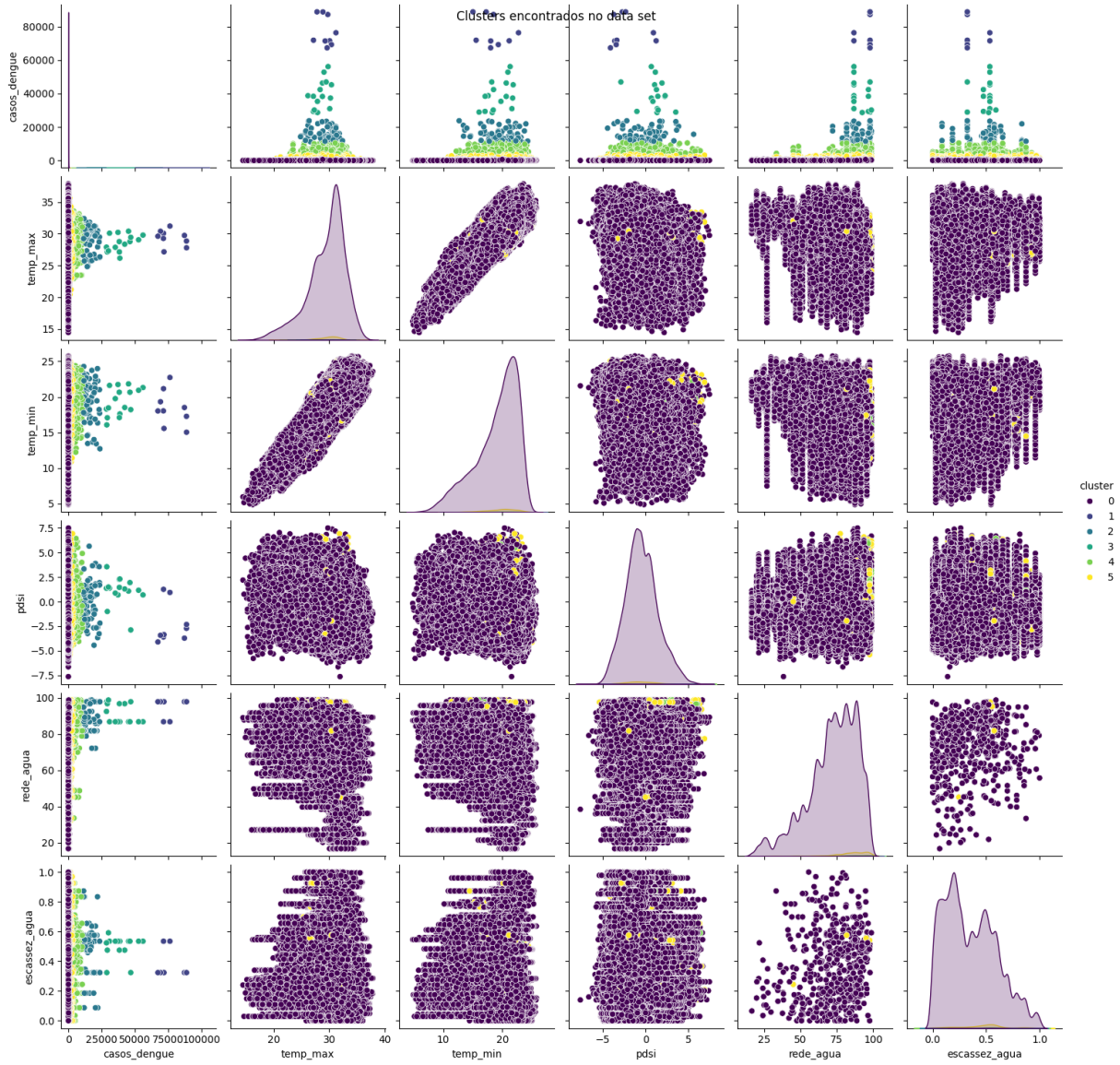


Outro fato que chama atenção é que nenhum atributo se relaciona fortemente com o outro além dos que já eram esperados como temperatura mínima e máxima. O que mais interessa aqui é analisar a correlação dos casos de dengue com outras variáveis e como é visto no gráfico acima ela não está fortemente relacionada com nenhum outro atributo.

2.2 Clusterização

Com o intuito de verificar se os dados já têm alguma separação natural foi aplicado o algoritmo de clusterização. Para isso, foi utilizado o método do cotovelo para

encontrar um número ideal de clusters e o algoritmo utilizado foi o K-Means. Como resultado o algoritmo encontrou seis agrupamentos.



2.3 Modelo de Regressão

Para o treinamento do modelo foi aplicado a técnica OneHotEncoder para o tratamento de dados não numéricos como nome da cidade, estado, bioma e clima principal. Além disso, foi realizado também um segundo treinamento onde não foi utilizado o OneHotEncoder e sim os próprios códigos de cidade, estado entre outros que havia no data-set, somente a feature clima principal que foi necessário a utilização dessa técnica.

Para o treinamento foi definido as features removendo a coluna que seria prevista que é os casos de dengue e definimos o label (Y) com a coluna casos de dengue.

O conjunto de dados foi dividido em treino e teste, onde vinte por cento dos dados foi destinado para teste e oitenta para treino.

Os algoritmos utilizados foram a regressão linear (LinearRegression); KNN regressor (KNeighborsRegressor); decision tree (DecisionTreeRegressor) e o Random forest regressor (RandomForestRegressor).

3 RESULTADOS

A principal métrica utilizada para avaliação dos modelos foi o erro médio absoluto (MAE).

Aplicando os algoritmos descritos acima, com a técnica do OneHotEncoder para tratar dados não numéricos, foi obtido o seguinte resultado:

	LinearRegression	KNeighborsRegressor	DecisionTreeRegressor	RandomForestRegressor
Erro Médio Absoluto (MAE)	165.2986470436055	80.35052671024691	95.84415205895824	85.16121580547112

Para obtenção desses resultados o KNeighborsRegressor, foi instanciado com o valor cinco no parâmetro de quantidade de vizinhos (‘n_neighbors’). E o RandomForestRegressor foi instanciado com valor cem no parâmetro ‘n_estimators’. DecisionTreeRegressor e LinearRegression não foi definido parâmetro algum durante a instanciação e treinamento.

Utilizando os códigos de nome de cidade, estado entre outros que já estavam presentes no conjunto de dados e aplicando o OneHotEncoder somente para o atributo clima principal os resultados foram ligeiramente melhores.

	LinearRegression	KNeighborsRegressor	DecisionTreeRegressor	RandomForestRegressor
Erro Médio Absoluto (MAE)	156.32126993363545	76.53227297331057	86.4635466544531	75.62615605612692

4 DISCUSSÃO

O ponto forte dos modelos obtidos foi a baixa faixa de erro absoluto que obtemos. Isso quer dizer que cidades onde tem ocorrências de mil casos de dengue por exemplo o modelo treinado vai trazer um palpite entre 1076 e 924 em seu melhor desempenho que foi por meio do RandomForest. Porém em cidades onde a ocorrência de casos é menor por exemplo uma média de cem casos, uma faixa de erro de aproximadamente 76 é muito alta. Resumindo, o ponto forte do modelo são cidades maiores com maior média de ocorrência de casos de dengue e o ponto fraco dele são as cidades menores com menos ocorrência de casos de dengue.

Caso o conjunto de dados abarcasse outros atributos como fatores socioeconômicos das cidades, investimento em saneamento básico, esgoto a céu aberto, entre outros. Seria possível identificar outras correlações com os casos de dengue e não ficaria restrito somente ao clima. Isso garantiria um modelo com um poder de abstração muito maior.

5 VIABILIDADE PARA APLICAÇÃO PRÁTICA

Avaliando os pontos fracos e fortes do modelo gerado, ele é mais indicado para cidades maiores com maior ocorrência de casos de dengue, nesse caso o modelo seria útil para identificar períodos de maior ocorrência com base em dados e evidências científicas. Com isso, as cidades podem se preparar melhor para os picos de casos.

Entretanto, o modelo não seria muito viável para cidades menores onde a ocorrência de casos também é menor, ele não teria uma precisão tão alta devido ao erro médio absoluto.