

**Fundação Getúlio Vargas
Escola de Matemática Aplicada**

Gabriel Lima Novais

**Modelagem de nicho ambiental de vetores do
vírus da Dengue com métodos de Gradient
Boosting Decision Tree**

Rio de Janeiro
2021

Gabriel Lima Novais

**Modelagem de nicho ambiental de vetores do
vírus da Dengue com métodos de Gradient
Boosting Decision Tree**

Dissertação submetida à Escola de Matemática Aplicada como requisito parcial para a obtenção do grau de Mestre em Modelagem Matemática da Informação.

Área de Concentração: Modelagem e Análise da Informação

Orientador: Claudio José Struchiner

Rio de Janeiro
2021

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Novais, Gabriel Lima

Modelagem de nicho ambiental de vetores do vírus da dengue com métodos de gradient boosting decision tree. – 2021.
79 f.

Dissertação (mestrado) - Fundação Getulio Vargas, Escola de Matemática Aplicada.

Orientador: Claudio José Struchiner.

Inclui bibliografia.

1. Epidemiologia. 2. Dengue. 3. Aprendizado do computador. 4. Modelagem de dados. I. Struchiner, Claudio José. II. Fundação Getulio Vargas. Escola de Matemática Aplicada. III. Título.

CDD – 006.31

Elaborada por Kelly Ayala – CRB-7/7007

GABRIEL LIMA NOVAIS

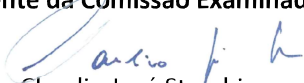
“MODELAGEM DE NICHOS AMBIENTAIS DE VETORES DO VÍRUS DA DENGUE COM MÉTODOS DE GRADIENT BOOSTING DECISION TREE”.

Dissertação apresentado(a) ao Curso de Mestrado em Modelagem Matemática do(a) Escola de Matemática Aplicada para obtenção do grau de Mestre em Modelagem Matemática.

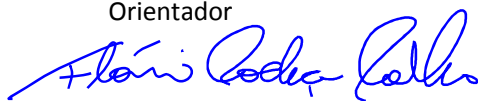
Data da defesa: 22/03/2021

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

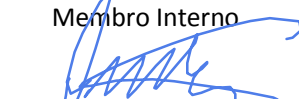
Presidente da Comissão Examinadora: Prof^o Claudio José Struchiner



Claudio José Struchiner
Orientador



Flávio Codeço Coelho
Membro Interno



Oswaldo Gonçalves Cruz
Membro Externo

Nos termos da Lei nº 13.979 de 06/02/20 - DOU nº 27 de 07/02/20 e Portaria MEC nº 544 de 16/06/20 - DOU nº 114 de 17/06/20 que dispõem sobre a suspensão temporária das atividades acadêmicas presenciais e a utilização de recursos tecnológicos face ao COVID-19, as apresentações das defesas de Tese e Dissertação, de forma excepcional, serão realizadas de forma remota e síncrona, incluindo-se nessa modalidade membros da banca e discente,



César Leopoldo Camacho Manco
Diretor



Antonio de Araujo Freitas Junior
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação FGV

Antonio Freitas, PhD
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação
Fundação Getúlio Vargas

Instrução Normativa nº 01/19, de 09/07/19 - Pró-Reitoria FGV

Em caso de participação de Membro(s) da Banca Examinadora de forma não-presencial*, o Presidente da Comissão Examinadora assinará o documento como representante legal, delegado por esta I.N.

*Skype, Videoconferência, Apps de vídeo etc

Agradecimentos

A Deus e à minha família, que sempre estão ao meu lado me apoiando e me confortando perante os inúmeros desafios da vida.

Ao professor Claudio, pela orientação, paciência e todo empenho dedicado neste trabalho.

Aos meus colegas de mestrado por compartilharem suas experiências e conhecimentos, além das lembranças de todos os intermináveis estudos nas salas da FGV.

Ao corpo docente da EMap e aos demais funcionários envolvidos, pela excelência de ensino.

À equipe do FGV CERI pela oportunidade de aplicar meus conhecimentos em um ambiente de pesquisa integrado ao mercado.

Resumo

O objetivo deste trabalho concentra-se em modelar o nicho ambiental de vetores do vírus da Dengue em nível municipal. Isto é realizado através de métodos de *Gradient Boosting* aplicados em modelos de *Ensemble* baseados em árvores de decisão. Três conhecidas técnicas foram utilizadas: XGBoost, LightGBM e CatBoost. Tais métodos são famosos pela sua capacidade de resolver problemas de classificação produzindo elevada acurácia e apresentando relevância das variáveis para o modelo. Os resultados obtidos apontam que esses modelos conseguem apresentar acurácia razoável, acima de 75%, o que associada as demais métricas de avaliação, indicam ótima performance e possível aplicação em tomadas de decisão no contexto de políticas públicas.

Palavras-chave: Epidemiologia, Dengue, Aprendizado de Máquina.

Abstract

The objective of this work is modelling the environmental niche of Dengue virus vectors at the municipal level. This is accomplished through *Gradient Boosting* methods applied to *Ensemble Tree* models. Three well-known techniques were used: XGBoost, LightGBM and CatBoost. Such methods are famous for their ability to solve classification problems, producing high accuracy and assessing the relevance of variable contribution to the model. The results obtained point out that these models manage to present reasonable accuracy, above 75%, which associated to the other evaluation metrics, indicate excellent performance and possible application in decision making in the context of public policies.

Keywords: Epidemiology, Dengue, Machine Learning.

Lista de Figuras

1	Estações meteorológicas com registros em 2001	16
2	Malhas de Temperatura criadas com KNN	17
3	Temperaturas para os pontos representativos dos municípios em 2019	18
4	Número total de infectados pela Dengue	20
5	Proporção de Municípios Infectados pela Dengue	21
6	Municípios Infectados pela Dengue em 2001	22
7	Municípios Infectados pela Dengue em 2019	23
8	Pontuação dos Municípios conforme variável <i>favelas</i>	28
9	Faixas de média dos IDH dos census de 2000 e 2010 dos Mu- nicípios	29
10	Evolução da densidade demográfica dos Municípios por Porte (Grande e Metrópoles)	31
11	Evolução da densidade demográfica dos Municípios por Porte (Pequeno I, Pequeno II e Médio)	32
12	Distribuição das Temperaturas Mínimas dos Municípios	34
13	Distribuição das Temperaturas Máximas dos Municípios	36
14	Distribuição das Temperaturas Compensadas dos Municípios . .	37
15	Distribuição das Amplitudes de Temperatura dos Municípios . .	37
16	Temperatura Mínima vs Precipitação Total	39
17	Temperatura Compensada Média vs Precipitação Total	39
18	Relação entre Temperatura Mínima e Temperatura Máxima . .	40
19	Proporção de Municípios Infectados por Bioma	42
20	Correlação entre as variáveis selecionadas	43
21	Previsão de Infecção dos Municípios em 2019 - XGBoost	67
22	Previsão de Infecção dos Municípios em 2019 - LightGBM . . .	68
23	Previsão de Infecção dos Municípios em 2019 - CatBoost	69

Lista de Tabelas

1	Média de todos os anos das variáveis para municípios litorâneos	24
2	Média de todos os anos das variáveis para as capitais	25
3	Área dos municípios por estados do Brasil	26
4	População anual média dos municípios	30
5	Temperatura Mínima Média (°C)	33
6	Temperatura Máxima Média (°C)	35
7	Precipitação Total (mm)	38
8	Avaliação dos modelos - Cross-Validation	61
9	Importância das Variáveis (%)	62
10	Resultados nos dados de Teste	62
11	Avaliação dos modelos (métricas em %) - Dados de Teste . . .	63
12	Avaliação dos modelos - Cross-Validation	64
13	Importância das Variáveis (%)	65
14	Resultados nos dados de Teste	66
15	Avaliação dos modelos (métricas em %) - Dados de Teste . . .	66

Sumário

1	Introdução	10
2	Base de Dados	12
2.1	IBGE	12
2.2	DATASUS	14
2.3	INMET	14
2.4	Construção da base de dados	15
3	Análise Exploratória dos Dados	19
3.1	Distribuição dos Municípios Infectados	19
3.2	Características Geográficas	24
3.3	Variáveis Socioeconômicas e Demográficas	26
3.4	Relações e análises das variáveis de Clima	32
3.5	Influência dos Biomas	40
3.6	Correlação entre as variáveis	42
4	Referencial Teórico	43
4.1	Decision Trees	44
4.2	Ensemble Learning	47
4.3	Gradient Tree Boosting	48
4.4	XGBoost	51
4.5	LightGBM	54
4.6	CatBoost	56
5	Metodologia	58
6	Resultados	60
6.1	Modelagem Cross-Section	60
6.2	Modelagem Espaço-Temporal	63
7	Conclusão	70

1 Introdução

A Dengue é uma doença viral transmitida aos humanos por artrópodes, e é causada pelo vírus da dengue denominado por DENV. Os sorotipos da Dengue (DENV 1-4) são transmitidos aos seres humanos principalmente pelos mosquitos *Aedes aegypti* e *Ae. Albopictus*. A Infecção causada por ela produz um espectro de apresentações clínicas que variam de debilitação aguda acompanhada de febre, até hemorragia com risco de vida, chamada de dengue hemorrágica (DHF / DSS). Existem 2,5 bilhões de pessoas com risco de infecção pela Dengue e uma estimativa de 25 a 100 milhões de casos, com 250.000 casos de DHF, ocorrendo em todo o mundo [Ross, 2010].

A Epidemia de DHF foi primeiramente reconhecida em 1950 no sudeste da Ásia e depois, em 1975, se tornou líder em casos de hospitalização e morte de crianças em diversos países nessa região [Ross, 2010]. Em 1980, DHF voltou a se expandir para a Ásia e em países nos quais ela era endêmica. Nos últimos 20 anos o DHF se espalhou pelas Américas e continua provocando surtos até hoje. Estes surtos, definidos como eventos epidemiológicos caracterizados por período prolongado, alto número de casos, e ocorrendo fora do período sazonal, ocasionam prejuízos para a saúde pública e para a qualidade da vida humana. Apesar de suas consequências, e dos estudos elaborados até o presente momento, ainda é necessário um entendimento mais aprofundado sobre a distribuição dos vetores da Dengue, com a finalidade de prevenir e controlar a doença além de estabelecer regiões de risco e auxiliar na avaliação de impactos.

As doenças infecciosas dificilmente exibem um comportamento simples e existem diversas formas de modelar espacialmente e temporalmente essa dinâmica de acordo com a literatura [Brady et al., 2015]. Para entender a dinâmica dos vetores transmissores da Dengue, devemos antes averiguar quais os nichos que são adequados para o desenvolvimento e reprodução deles. O objetivo do presente trabalho consiste em contribuir para a compreensão das variáveis que tornam o nicho dos vetores favorável e construir modelos matemáticos com técnicas de machine learning capazes de prever a ocorrência deles dadas as características da região. Exemplos de modelagem similar podem ser vistos em [Laureano-Rosario et al., 2018], [Buczak et al., 2018], [Escobar et al., 2016], [Guo et al., 2017], [Jain et al., 2019], [Ke et al., 2017] e [Mussumeci and Coelho, 2020]. O foco será dado em nível municipal para o Brasil dos anos de 2001 até 2019, ano em que se encontram disponíveis dados sobre ocorrências prováveis de dengue segundo DATASUS.

Os fatores abióticos responsáveis pela proliferação do mosquito são encontrados em abundância no Brasil [Custódio et al., 2019]. No país existem regiões com temperaturas, umidade, vento e precipitação adequadas. Tais fatores costumam estar associados diretamente com a oviposição, viabilidade dos ovos, depósitos de larvas, longevidade e dispersão dos mosquitos adultos. Além desses fatores ainda existem fatores macrodeterminantes que auxiliam a proliferação dos vetores e transmissão da Dengue, tais como condições inadequadas de saneamento, ausência de tratamento e destinos ideais para os resíduos sólidos, baixa distribuição de renda e baixa escolaridade da população, urbanização e fatores geográficos, tais como altitude, relevo e densidade de vegetação.

As variáveis mais interligadas com a adequação dos nichos para os vetores transmissores da Dengue, com resultados conhecidos na literatura, são: temperatura, umidade, precipitação e demografia. O fator temperatura é extremamente importante no ciclo de reprodução dos mosquitos, pois com a temperatura adequada a fêmea do mosquito consegue sobreviver tempo suficiente para completar o primeiro ciclo gonotrófico e então depositar os ovos. Tanto a longevidade da fêmea como a duração do primeiro ciclo gonotrófico são dependentes da temperatura. Os resultados obtidos em [Brady et al., 2013] indicam que os adultos do *Ae. albopictus* possuem uma capacidade de sobrevivência maior que o *Ae. Aegypti*, tanto em experimentos elaborados no laboratório quanto em campo, porém o *Ae. Aegypti* consegue aguentar intervalos de temperatura maior que o *Ae. albopictus*.

A disponibilidade de água é um fator relevante e está diretamente correlacionado com a precipitação e umidade da região, ver [Dickens et al., 2018]. A umidade do solo no nível da superfície está associado à disponibilidade de locais de desenvolvimento de larvas de mosquitos. Ovos e adultos precisam de umidade para sobreviver, logo lugares secos costumam afetar a mortalidade dos adultos. A precipitação está extremamente relacionada com os depósitos de água que ocorrem depois das chuvas [Dickens et al., 2018]. Estes depósitos muitas vezes são produzidos de maneira não intencional pelos seres humanos ou se originam de vegetações e outras pequenas formações naturais (microhabitats). A urbanização nesse sentido produz criadouros para os mosquitos de maneira indireta. Todos os elementos citados são interligados, e assim, modelar o nicho desses vetores podem justificar a utilização de modelos não lineares mais complexos, que possam estabelecer relações talvez pouco intuitivas sobre estas variáveis.

2 Base de Dados

As variáveis utilizadas para modelar os nichos dos vetores transmissores da Dengue vieram de diferentes fontes e precisaram de diversos ajustes. Foram necessários dados provenientes do Instituto Brasileiro de Geografia e Estatística (IBGE ¹), do departamento de informática do Sistema Único de Saúde do Brasil (DATASUS ²) e do Instituto Nacional de Meteorologia (INMET ³). As estatísticas descritivas das variáveis e os *insights* originados por elas serão detalhados de maneira mais aprofundada na seção *Análise Exploratória dos Dados*.

2.1 IBGE

Através dos estudos do IBGE e dos dados disponibilizados em seu site foi possível extrair as seguintes variáveis:

- Área (***Area***) por município (km^2)
- Pertencimento ou não do município ao litoral brasileiro (***Coast***), enquanto variável *dummy* totalizando 280 municípios de 5575
- População por município (***Pop***) por ano, segundo estimativas do próprio instituto e dos Censos realizados de maneira esporádica, em especial os que ocorreram em 2000 e 2010.
- Porte do município (***Porte***), segundo quantidade de habitantes, quantidade de Centros de Referência de Assistência Social (CRAS) e quantidade de famílias. As categorias possíveis são:
 1. **Pequeno Porte I:** Município de até 20.000 habitantes/5.000 famílias – mínimo de 1 CRAS para até 2.500 famílias referenciadas;
 2. **Pequeno Porte II:** Município de 20.001 a 50.000 habitantes/de 5.000 a 10.000 famílias – mínimo de 1 CRAS para até 3.500 famílias referenciadas;
 3. **Médio Porte:** Município de 50.001 a 100.000 habitantes/de 10.000 a 25.000 famílias – mínimo de 2 CRAS, cada um para até 5.000 famílias referenciadas;

¹<https://www.ibge.gov.br>

²<https://www.datasus.saude.gov.br>

³<https://www.portal.inmet.gov.br>

4. **Grande Porte:** Município de 100.001 a 900.000 habitantes/de 25.000 a 250.000 famílias – mínimo de 4 CRAS, cada um para até 5.000 famílias referenciadas;
 5. **Metrópoles:** Município de mais de 900.000 habitantes/mais de 250.000 famílias – mínimo de 8 CRAS, cada um para até 5.000 famílias referenciadas;
- Índice de Desenvolvimento Humano (**IDH_faixa**), que representa a média dos índices dos censos de 2000 e 2010 nos municípios e categorizados segundo os grupos de intervalos abaixo:
 1. $[0; 0, 499]$, classificado como “*Muito Baixo*”
 2. $[0, 5; 0, 599]$, classificado como “*Baixo*”
 3. $[0, 6; 0, 699]$, classificado como “*Médio*”
 4. $[0, 7; 0, 799]$, classificado como “*Alto*”
 5. $[0, 8; 1]$, classificado como “*Muito Alto*”
 - Variáveis *dummy* de Biomas existentes no município (**amazonia, caatinga, cerrado, mata_atlantica, pampa, pantanal**), sendo que um município pode conter mais de um Bioma.
 - Classificação dos Aglomerados Subnormais (**favela**), segundo 7 critérios de localização, estruturas dos imóveis e vias de acesso:
 1. Margem de córregos, rios, lagos ou lagoas;
 2. Sobre rios, córregos, lagos ou mar (palafitas);
 3. Manguezal;
 4. Aterros sanitários, lixões e outras áreas contaminadas;
 5. Beco/Travessa;
 6. Não existe via de circulação interna;
 7. Sem espaçamento;

Vale observar que a variável (**favela**) foi construída considerando uma pontuação de 0 a 8, onde o 0 representa a inexistência de aglomerado subnormal no município e valores até 8 inclusive correspondem a critérios observados nos aglomerados. Pontuações mais altas representam ambientes mais favoráveis à existência de criadouros dos vetores da Dengue.

2.2 DATASUS

Os dados são provenientes do Ministério da saúde e foram coletados de 3 seções diferentes no site do DATASUS. Elas foram:

1. Casos de Notificação de 2001 até 2006. Link acessado no dia 31 de Agosto de 2020: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinanwin/cnv/denguebr.def>
2. Casos prováveis de 2007 até 2013. Link acessado no dia 31 de Agosto de 2020: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinannet/cnv/denguebr.def>
3. Casos prováveis de 2014 até 2019. Link acessado no dia 31 de Agosto de 2020: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinannet/cnv/denguebbr.def>

Todas as partes consideraram o ano do(s) primeiro(s) sintoma(s) e a região como município de infecção. Nem todos os municípios existentes estavam registrados nessa base, de fato apenas 3306 municípios possuíam alguma ocorrência provável nos 19 anos de estudo. Por este motivo os demais municípios foram registrados na base de dados com nenhuma ocorrência provável totalizando os 5574 municípios do Brasil.

A variável coletada destes registros foi o número de casos prováveis por município por ano (*InfectNumber*), de onde foi gerada a variável dummy de classificação se ocorreu ou não infecção naquele município naquele ano (*Infected*).

2.3 INMET

O Instituto Nacional de Meteorologia (INMET) possui dados armazenados no sistema Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) que é um banco de dados para apoiar as atividades de ensino e pesquisa e outras aplicações em meteorologia, hidrologia, recursos hídricos, saúde pública, meio ambiente entre outras. No BDMEP é possível obter informações meteorológicas históricas de estações convencionais, que totalizam um total de 265 estações distribuídas no território brasileiro e diversas outras estações automáticas. As variáveis atmosféricas disponibilizadas para consultas no BDMEP, conforme descrito no próprio site, e que são importantes para o objetivo deste trabalho, foram denominadas segundo a lista abaixo:

- Temperatura Máxima Média ($^{\circ}\text{C}$) : ***TempMaximaMedia***
- Temperatura Mínima Média ($^{\circ}\text{C}$) : ***TempMinimaMedia***
- Temperatura Compensada Média($^{\circ}\text{C}$): ***TempCompensadaMedia***
- Precipitação Total (mm): ***PrecipitacaoTotal***
- Umidade Relativa Média (%): ***UmidadeRelativaMedia***
- Evaporação Piche (mm): ***EvaporacaoPiche***
- Insolação Total (W/m^2): ***InsolacaoTotal***
- Nebulosidade Média (Escala de Octas): ***NebulosidadeMedia***

A seleção dessas variáveis se deu principalmente devido a importância atribuída pela literatura, no que tange a formação de nichos favoráveis ao aparecimento do vetor do vírus da dengue, conforme mencionado na introdução. Vale ressaltar que a coleta dos dados das 265 estações se deu na periodicidade mensal, o que foi posteriormente agregado pela média para formar as variáveis em periodicidade anual. Para as variáveis oriundas das estações automáticas, não foi preciso agregar pois os dados já vinham em periodicidade anual, embora para a criação da variável Temperatura Compensada Média, uma *proxy* calculada a partir da média das temperaturas mínimas, máximas e do ar seco foi realizada.

2.4 Construção da base de dados

A construção das bases de dados anuais, utilizadas para a modelagem, precisaram passar por um tratamento para resolver os problemas de valores faltantes (*missing values*) e pela extrapolação após unificação dos dados das estações meteorológicas convencionais e automáticas.

A extrapolação dos dados teve como objetivo estabelecer as variáveis climáticas anuais para cada município. Dessa forma, a partir dos dados e das localizações das estações poder estimar valores para outros pontos geográficos representativos do município. Tais pontos, tanto para as estações quanto para os municípios, são determinados pela sua latitude e longitude. Na figura abaixo é possível verificar a distribuição das estações meteorológicas com registros em 2001.

Figura 1: Estações meteorológicas com registros em 2001

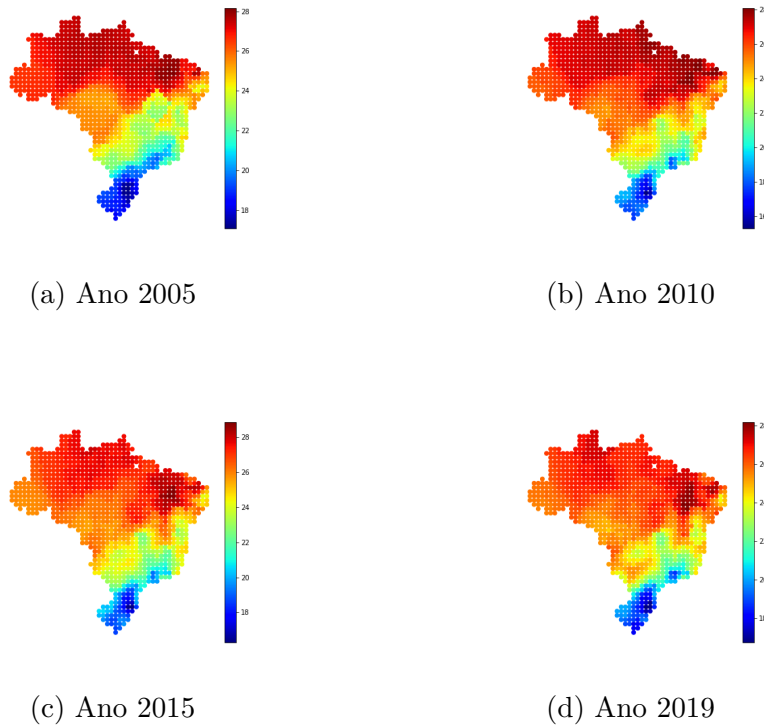


O método utilizado para este fim foi o k-nearest neighbors (KNN), método de machine learning e não paramétrico muito comum inclusive para preenchimento de valores faltantes, algo que também foi realizado nas bases anuais deste trabalho. As previsões são feitas para um novo ponto pesquisando todo o conjunto de treinamento para as 10 instâncias mais semelhantes (denominadas de vizinhos) e tomando a média deles. Para determinar quais das instâncias no conjunto de dados de treinamento são mais semelhantes a uma nova entrada, uma medida de distância é usada. Nesse trabalho foi utilizado

uma métrica de distância denominada de Distância *Minkowski* que se trata de uma generalização da distância Euclidiana e Manhattan.

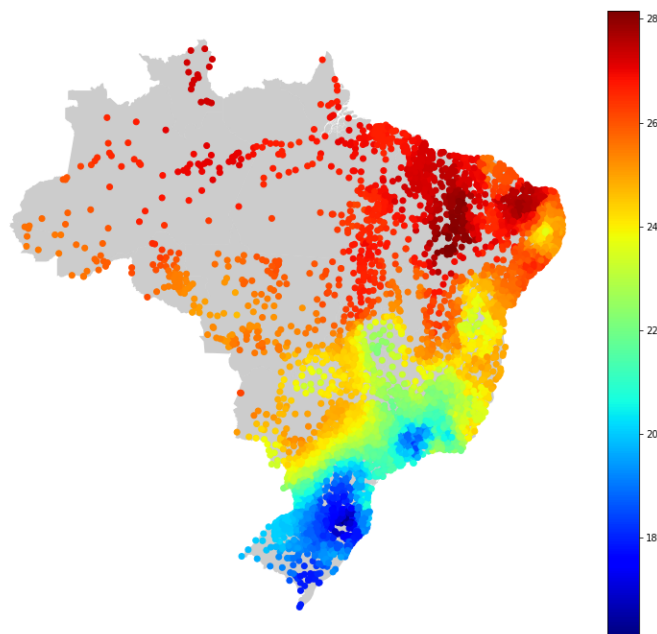
A partir desta etapa, foi possível então construir uma malha de variáveis cujos mapas para o caso das Temperaturas Compensadas Médias podem ser visualizados a seguir. Podemos verificar as mudanças na escala de temperatura, suas respectivas disposições e a suavidade das transições .

Figura 2: Malhas de Temperatura criadas com KNN



Com as malhas construídas foi então possível tomar as variáveis apenas para os pontos representativos de cada município, conforme pode ser observado para o caso da Temperatura Compensada Média exposta no gráfico abaixo.

Figura 3: Temperaturas para os pontos representativos dos municípios em 2019



A agregação consequente dos dados foi feita de maneira a abranger o maior número de observações em todos os anos. As bases de todos os anos mantiveram 5574 observações. Os scripts, produzidos na linguagem Python 3.6, com o auxílio de dois pacotes principais, o Pandas ⁴ e o Numpy ⁵ necessários para organizar e produzir as bases de dados descritas nesta seção estão listadas abaixo.

- *data_0*: script utilizado para criar uma base núcleo (core).
- *data_1*: script para ajustar os dados da Dengue para csv e por ano na base núcleo.

⁴<https://pandas.pydata.org/pandas-docs/stable/index.html>

⁵<https://numpy.org>

- *data_3*: script para juntar os dados de população por ano na base núcleo.
- *data_41*: script para organizar dados das estações convencionais.
- *data_42*: script para organizar dados das estações automáticas.
- *data_5*: script para unir dados das estações convencionais e automáticas.
- *data_6*: script para juntar os dados de clima na base núcleo.

Após realização desta etapa foram criadas duas variáveis importantes tanto para a análise exploratória dos dados quanto para a modelagem em si, a partir dos dados existentes:

1. A Densidade Populacional, denominada por ***PopArea*** calculada como a razão entre a variável ***Pop*** e ***Area***;
2. Amplitude Térmica Média, denominada por ***TempDiff***, que foi obtida a partir de diferença entre as variáveis ***TempMaximaMedia*** e ***TempMinimaMedia***

A amplitude térmica média conforme será descrito em seções posteriores, possui importância na sobrevivência de algumas espécies transmissoras do vírus da Dengue, ver [Brady et al., 2013].

3 Análise Exploratória dos Dados

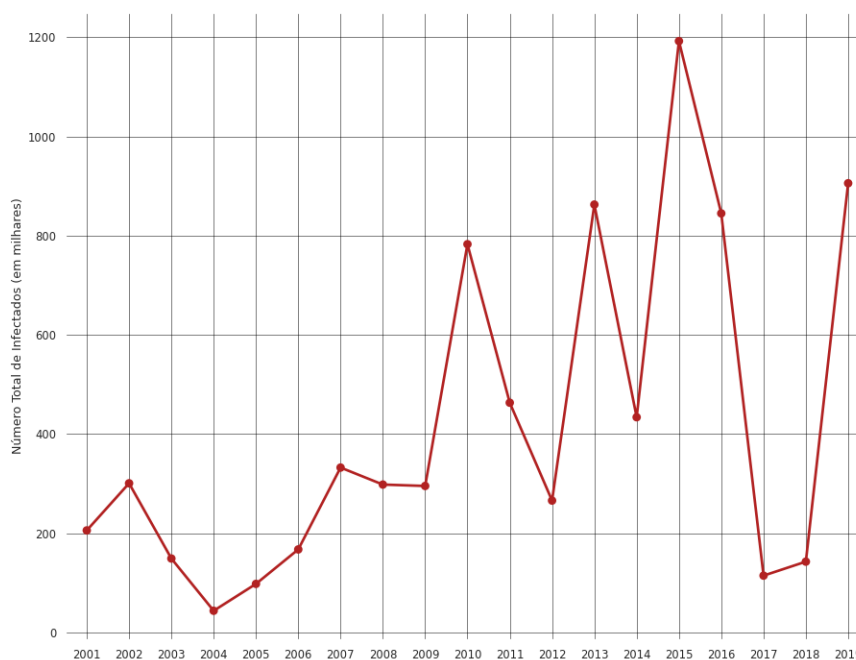
Nesta seção será realizada uma análise exploratória dos dados, cujo principal objetivo será entender as relações entre as variáveis organizadas a partir dos dados descritos na seção anterior. Pretende-se com isso, evidenciar a importância de algumas delas e buscar *insights* valiosos para o processo de modelagem do problema.

3.1 Distribuição dos Municípios Infectados

A variável ***Infected*** é a representação em *dummy* da ocorrência de infecção ou não no município pelo vírus da Dengue. A avaliação inicial para compreender os dados inicia-se pelo estudo do comportamento da infecção por município e em termos de casos prováveis, pois alguns pontos interessantes

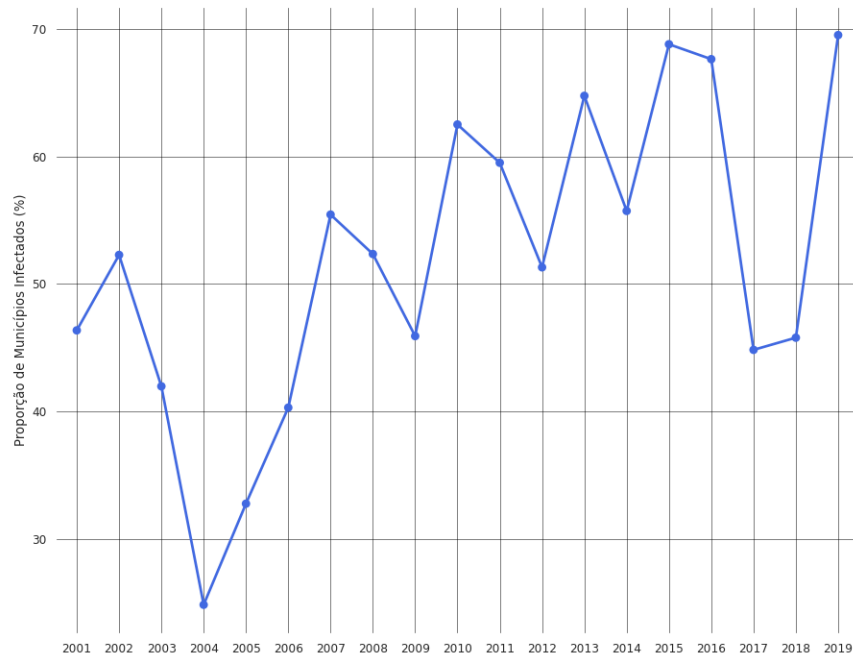
podem ser visualizados. De acordo com a Figura 4, pode-se verificar que o número de casos de infectados cresceu abruptamente desde 2001, apesar das políticas públicas de combate ao vetor da dengue e das quedas nos anos de 2016 e 2017, chegando a um nível elevado em 2019, nível este similar aos anos de 2013 e 2010.

Figura 4: Número total de infectados pela Dengue



Além disso, através da Figura 5 é possível observar curva semelhante ao da Figura 4, apontando queda em 2004 e 2017, com elevação em 2019. Dois pontos principais podem ser inferidos a partir destas figuras: (1) As séries apresentaram tendência de crescimento (2) o número de municípios infectados aumentou junto com os casos prováveis, indicando, a priori, que não ocorreu concentração dos casos em municípios específicos, mas uma dispersão deles.

Figura 5: Proporção de Municípios Infectados pela Dengue



Abaixo podemos verificar nos mapas do Brasil por municípios, em vermelho, aqueles em que foram informados casos prováveis de infecção da Dengue, traduzindo melhor o entendimento das duas figuras acima. A partir destes mapas podemos verificar que so Sudeste e o Centro-Oeste Brasileiro são os que mais apresentam municípios com ocorrências da Dengue, e que de 2001 para o ano de 2019 o sudeste apresentou um número elevado de municípios nesta situação. Por outro lado, as demais regiões ficaram relativamente estagnadas.

Figura 6: Municípios Infectados pela Dengue em 2001

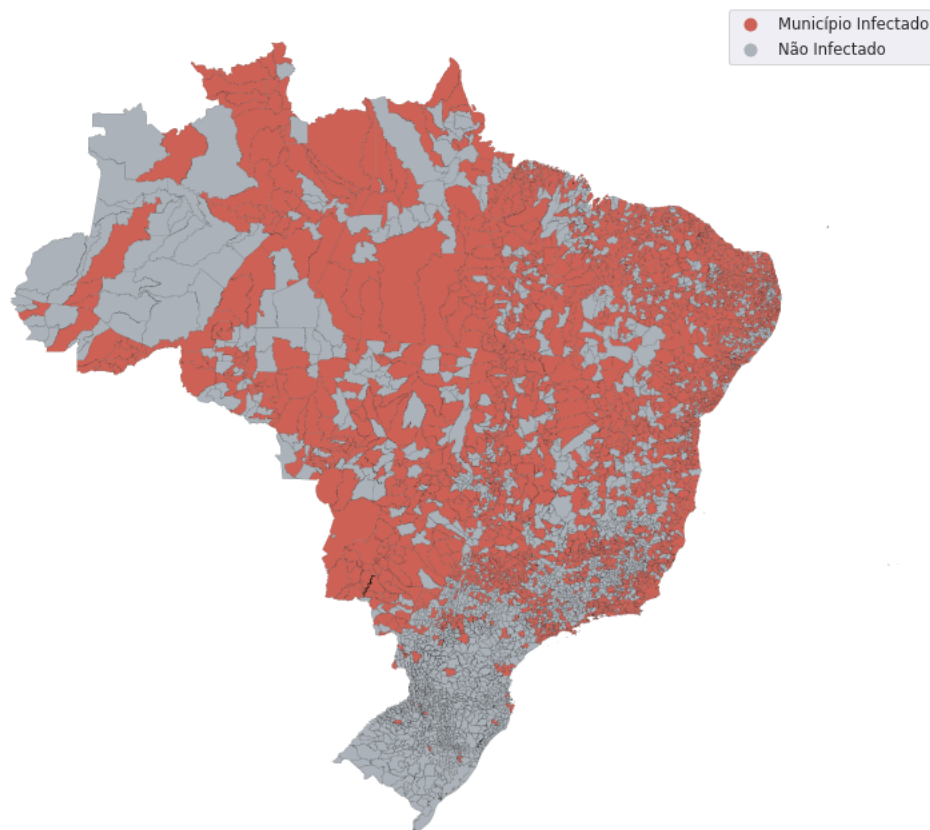
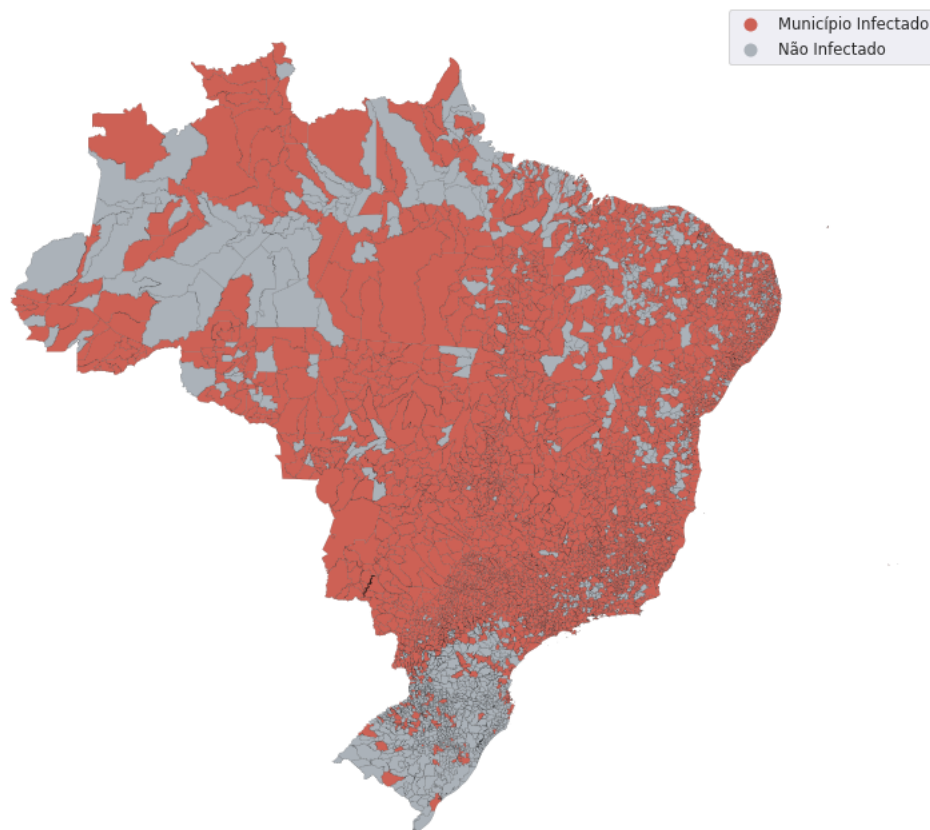


Figura 7: Municípios Infectados pela Dengue em 2019



3.2 Características Geográficas

Algumas características geográficas refletem tanto aspectos demográficos e socioeconômicos quanto se correlacionam com eles. É conhecido ⁶ que a região sudeste apresenta uma maior densidade populacional e um número maior de municípios, e isso se reflete em municípios com áreas menores e com maior infraestrutura.

O processo de colonização do Brasil privilegiou a ocupação de territórios situados no litoral. E por este motivo e o próprio desenvolvimento econômico destas regiões, a consequência hoje é percebida pela disposição geográfica dos estados e suas capitais. Inclusive a maior parte das capitais encontram-se em estados que possuem contato com o litoral, que possui mais de 7,3 mil quilômetros estendendo-se do Amapá ao estado do Rio Grande do Sul.

O litoral possui suas próprias características bioclimáticas, apresentando temperaturas elevadas, ventos constantes e intensa evaporação. As características dos municípios que possuem litoral, no que diz respeito às variáveis da base de dados selecionada, encontram-se na tabela abaixo.

Tabela 1: Média de todos os anos das variáveis para municípios litorâneos

<i>Variáveis</i>	<i>Média</i>	<i>Dp</i>	<i>Min</i>	<i>Max</i>
PrecipitacaoTotal	63,82	22,6	30,9	149,68
EvaporacaoPiche	114,11	35,38	70,4	204,1
InsolacaoTotal	199,18	23,44	157,3	247,28
NebulosidadeMedia	5,83	0,43	4,78	6,65
TempCompensadaMedia	24,04	3,1	17,52	27,57
TempMinimaMedia	21,73	2,94	15,23	25,04
TempMaximaMedia	27,03	3,34	20,35	31,75
TempDiff	5,30	0,83	3,33	7,6
UmidadeRelativaMedia	77,61	2,98	68,82	82,64

Como foi mencionado anteriormente, muitas capitais encontram-se em estados que possuem litoral e por este motivo as estatísticas obtidas na Tabela 1 apresentam pouca diferença quando comparamos com as estatísticas que estão disponíveis na Tabela 2, representando a média de todos os anos das variáveis para estas capitais. Essa diferença pequena também é fruto do cálculo de médias o que termina por suavizar as variações das regiões.

⁶<https://censo2010.ibge.gov.br/sinopse/index.php?dados=10&uf=00>

Tabela 2: Média de todos os anos das variáveis para as capitais

<i>Variáveis</i>	<i>Média</i>	<i>Dp</i>	<i>Min</i>	<i>Max</i>
PrecipitacaoTotal	72,95	28,34	21,38	149,68
EvaporacaoPiche	114,49	35,46	68,74	185,12
InsolacaoTotal	194,54	24,66	145,19	233,02
NebulosidadeMedia	5,85	0,6	4,85	6,98
TempCompensadaMedia	24,47	2,83	18,66	27,69
TempMinimaMedia	22,00	2,79	16,68	25,05
TempMaximaMedia	27,75	3,04	21,31	32,03
TempDiff	5,74	1,15	3,43	8,57
UmidadeRelativaMedia	76,22	5,11	65,07	82,99

Analizando, por outro lado, a distribuição territorial dos municípios verifica-se que além da maior parte deles estarem no sudeste, as suas áreas são bem menores que aquelas encontradas no Norte e no Centro-Oeste. Essa característica interfere significativamente na densidade populacional, gerando enormes variações desta grandeza conforme mudamos a região analisada.

Inclusive os pontos representativos dos municípios capturam menos características específicas dessas regiões, o que pode causar alguma perda de informação e prejudicar de alguma forma a modelagem nesses municípios. Pela disponibilidade dos dados não serem mais específicos, tais perdas produzidas no momento da extrapolação das variáveis climáticas, por exemplo, não podem ser minimizados mais do que foi realizado até o momento. Um fator que contribui para a atenuação desses efeitos surge pela classificação do município pelas variáveis de Bioma, que conseguem generalizar e uniformizar as regiões.

Desta maneira, para avaliar as distribuições das áreas dos municípios por estado, segue a Tabela 3 abaixo, denominada por “Área dos municípios por estados do Brasil”. Verifique que as maiores médias encontram-se de fato em estados que estão no Norte, e as menores, em regiões como Sudeste e Sul.

Tabela 3: Área dos municípios por estados do Brasil

<i>UF</i>	<i>Média</i>	<i>Dp</i>	<i>Min</i>	<i>Max</i>
AC	6.921.725,00	7.553.864	3.076	27.974.551
AL	248.084,00	189.474	48	918.208
AM	22.678.818,00	25.416.804	2.214	122.476.006
AP	7.415.380,00	8.595.890	1.579	30.971.775
BA	1.155.959,00	1.847.076	89	16.404.396
CE	740.302,00	751.422	95	4.260.455
DF	5.787.784,00	0	5.787.784	5.787.784
ES	494.756,00	518.746	89	3.501.627
GO	1.280.568,00	1.645.568	56	9.843.235
MA	1.368.083,00	1.884.379	261	13.141.688
MG	622.344,00	908.883	58	8.447.098
MS	4.470.198,00	7.756.153	1.807	64.962.836
MT	5.790.050,00	6.209.447	625	27.924.534
PA	8.410.126,00	19.618.831	489	159.533.401
PB	227.576,00	202.776	63	986.351
PE	451.313,00	647.932	73	4.558.398
PI	1.033.246,00	1.184.863	97	8.411.877
PR	465.776,00	457.958	94	3.116.313
RJ	453.853,00	497.190	76	4.026.712
RN	285.368,00	323.338	54	2.099.328
RO	4.047.467,00	5.837.852	3.442	34.096.429
RR	13.456.561,00	13.927.604	8.065	47.410.947
RS	465.814,00	825.847	60	6.950.388
SC	277.693,00	289.146	63	2.629.789
SE	265.743,00	268.983	45	1.232.117
SP	348.869,00	329.673	5	1.977.414
TO	1.924.113,00	2.256.518	621	13.423.347

3.3 Variáveis Socioeconômicas e Demográficas

O grupo de variáveis Socioeconômicas e Demográficas procuram sintetizar algumas interpretações básicas sobre o comportamento econômico, educacional e populacional dos municípios para correlacioná-los à variável ***Infected***.

As duas variáveis que representam o contexto socioeconômico dos municípios são: O índice de desenvolvimento humano (***IDH_faixa***) e a pre-

sença de aglomerados subnormais nos municípios e sua classificação segundo faixas (*favelas*). O primeiro consegue reunir três elementos interessantes: economia, saúde e educação. A renda se relaciona com o nível de moradia, e de recursos dos indivíduos; a longevidade indica em algum nível a saúde dos moradores do município; e o nível de instrução consegue estabelecer certa correlação entre conscientização e prevenção de criadouros da Dengue.

Os aglomerados subnormais, conhecidos informalmente como favelas, indicam locais prováveis de criadouros do vetor, uma vez que costumam apresentar saneamento básico incipiente, construções favorecendo micro habitats, como sulcos em telhas, ranhuras em tijolos entre outras; e por fim proporcionam em regiões concentradas altas quantidades de indivíduos, capazes de serem as fontes alimentares dos mosquitos fêmeas durante o ciclo de oviposição.

A distribuição dos aglomerados subnormais reside em sua grande parte no litoral brasileiro, se concentrando no Sudeste e em seguida no Norte. No Sudeste a característica desses aglomerados variam de moradias instaladas em terrenos de altitude elevada e próximos ao centro urbano da cidade até locais planos no interior dos estados. Neles além das possibilidades de micro habitats citadas existem ainda o agravante do clima e do bioma Mata Atlântica, favorável ao vetor. No Norte é possível verificar a presença de aglomerados situados próximos a rios e manguezais, com moradias cujo saneamento é pouco desenvolvido. Em geral, a retirada de lixos ou resíduos sólidos em aglomerados subnormais costuma ser difícil devido ao precário acesso de caminhões e funcionários das empresas dedicadas a este serviço.

Nos mapas abaixo podemos verificar a classificação dos municípios de acordo com as perspectivas apresentadas.

Figura 8: Pontuação dos Municípios conforme variável *favelas*

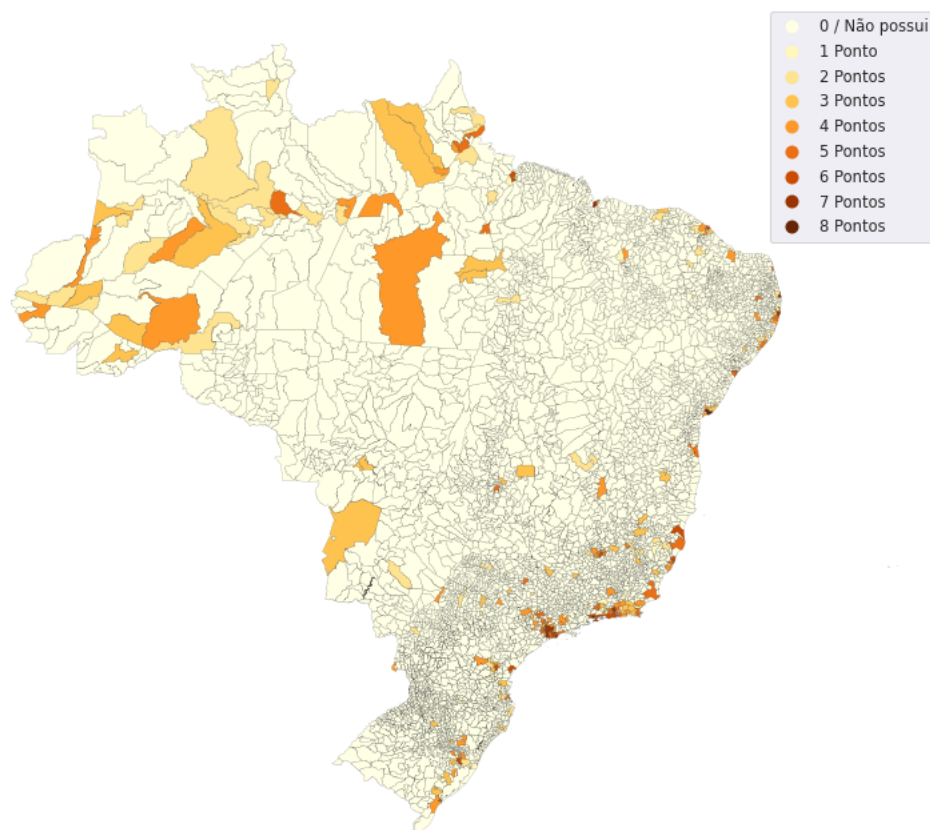
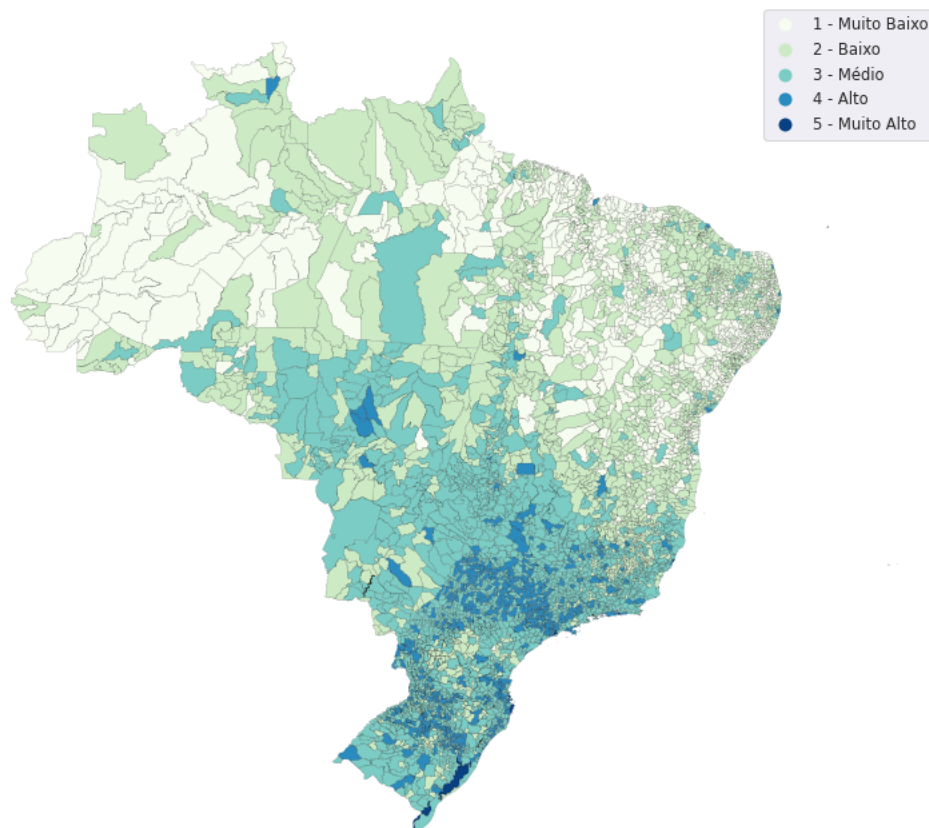


Figura 9: Faixas de média dos IDH dos census de 2000 e 2010 dos Municípios



Ainda uma última observação pode ser feita sobre os dois mapas acima, que apesar de haver a presença elevada de aglomerados subnormais no sudeste, a distribuição do IDH aponta que o mesmo lugar apresenta elevados índices de desenvolvimento humano. Isso, além de sugerir certo grau de desigualdade social, também aponta um contraste interessante e possivelmente

importante na formulação do modelo.

As variáveis demográficas conseguem reunir valiosas características. A população assim como a densidade populacional, por exemplo, observadas nos últimos anos, independente do porte do município têm crescido. Tal fato pode ser visto na Tabela 4, e nos gráficos abaixo.

Tabela 4: População anual média dos municípios

<i>Ano</i>	<i>Média</i>	<i>Dp</i>	<i>Min</i>	50%	<i>Max</i>
2001	31.070,48	187.311,63	800	10.386,5	10.499.133
2002	31.476,44	189.339,62	804	10.435	10.600.060
2003	31.881,02	191.107,5	809	10.443,5	10.677.019
2004	32.698,09	194.817,31	818	10.579,5	10.838.581
2005	33.167,73	196.897,89	823	10.634	10.927.985
2006	33.634,46	198.956,93	828	10.692,5	11.016.703
2007	33.120,45	197.747,96	804	10.684	10.886.518
2008	34.136,66	200.310,03	834	11.124	10.990.252
2009	34.473,50	201.675,08	837	11.209,5	11.037.593
2010	34.330,42	203.076,11	805	10.934	11.253.503
2011	34.640,25	204.341,97	806	11.030	11.316.119
2012	34.874,36	205.651,77	807	11.040	11.376.685
2013	36.149,32	212.621,49	825	11.388,5	11.821.873
2014	36.462,29	214.111,33	822	11.445,5	11.895.893
2015	36.765,49	215.562,6	818	11.511	11.967.825
2016	37.059,69	216.979,09	815	11.577,5	12.038.175
2017	37.344,56	218.366	812	11.635,5	12.106.920
2018	37.495,55	219.935,47	786	11.590,5	12.176.866
2019	37.793,46	221.425,11	781	11.631	12.252.023

Isso favorece o aumento de fontes alimentares dos mosquitos fêmeas no período de oviposição e aponta uma possibilidade de correlação entre capitais ou centros urbanos e o número de infectados, conforme gráficos mostrados nas seções anteriores. Outro ponto, seria o fato de que quanto maior a população de um município mais provável de ter algum indivíduo infectado, em termos probabilísticos. Os gráficos abaixo podem fornecer visualizações sobre o que foi mencionado.

Figura 10: Evolução da densidade demográfica dos Municípios por Porte (Grande e Metrôpoles)

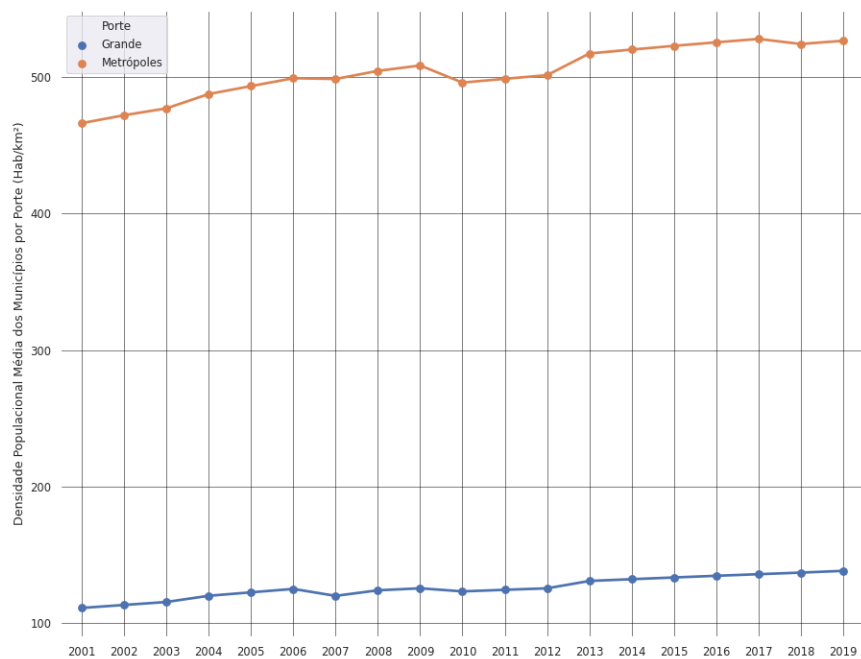
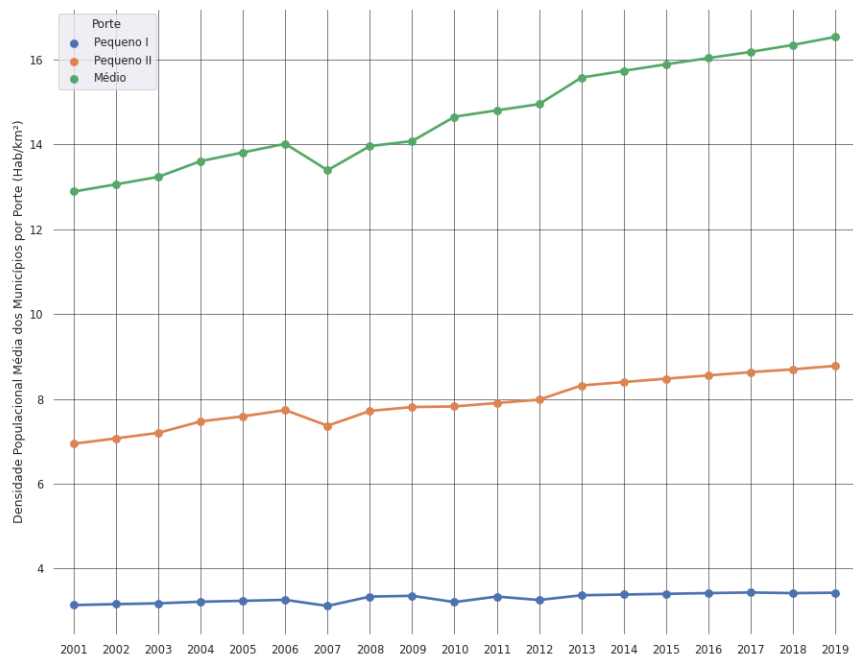


Figura 11: Evolução da densidade demográfica dos Municípios por Porte (Pequeno I, Pequeno II e Médio)



A única exceção que pode ser questionada são os municípios de pequeno porte (Pequeno I), que aparentam ter estagnado em termos de densidade populacional, com pequenas quedas nos anos de 2007, 2010 e 2012. Além disso, os dois portes que mais apresentaram o crescimento foram as Metrôpoles e os municípios de porte Médio.

3.4 Relações e análises das variáveis de Clima

As variáveis de clima, apesar de serem categorizadas conjuntamente com as variáveis de Bioma, como características macros dos municípios possuem importância elevada para a proliferação e sobrevivência dos vetores da Dengue. Em especial as variáveis que mais se aproximam das exigências biológicas

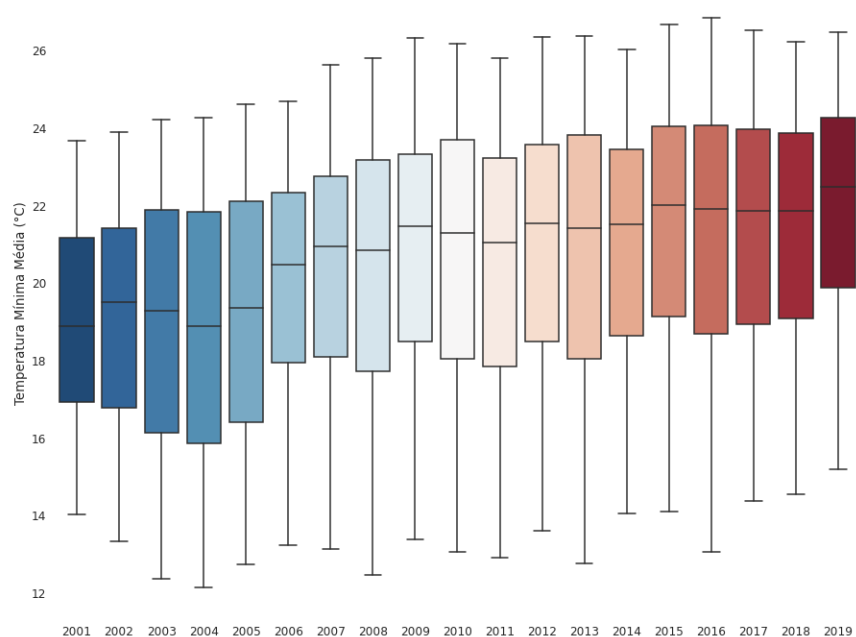
desses vetores são a temperatura, o nível de umidade e de precipitação. Variações extremas dessas características costumam produzir ambientes poucos favoráveis. Será analisada a temperatura e verificado posteriormente suas relações com a precipitação.

Iniciando-se com a Temperatura Mínima Média, pode-se inferir que de acordo com a evolução apresentada na Figura 12 os níveis médios se elevaram e a distribuição se condensou mais com o passar do tempo. Temperaturas mínimas mais elevadas indicam ambientes mais estáveis.

Tabela 5: Temperatura Mínima Média (°C)

<i>Ano</i>	<i>Média</i>	<i>Dp</i>	<i>Min</i>	25%	50%	75%	<i>Max</i>
2001	19,01	2,53	14,03	16,94	18,88	21,16	23,69
2002	19,08	2,7	13,34	16,79	19,52	21,42	23,91
2003	19,00	3,17	12,37	16,13	19,3	21,89	24,23
2004	18,85	3,22	12,14	15,86	18,89	21,85	24,27
2005	19,22	3,17	12,75	16,41	19,36	22,12	24,62
2006	20,15	2,49	13,24	17,94	20,47	22,34	24,7
2007	20,50	2,86	13,14	18,09	20,96	22,77	25,64
2008	20,50	3,16	12,46	17,72	20,85	23,18	25,8
2009	20,93	2,93	13,38	18,51	21,46	23,33	26,34
2010	20,92	3,2	13,06	18,05	21,31	23,7	26,2
2011	20,56	3,08	12,9	17,84	21,06	23,24	25,8
2012	21,09	2,92	13,6	18,5	21,55	23,58	26,35
2013	20,94	3,21	12,76	18,05	21,43	23,83	26,38
2014	21,15	2,78	14,06	18,66	21,53	23,46	26,05
2015	21,63	2,8	14,11	19,13	22,03	24,04	26,68
2016	21,38	3,18	13,06	18,69	21,92	24,08	26,85
2017	21,52	2,85	14,37	18,95	21,86	23,97	26,52
2018	21,41	2,81	14,56	19,09	21,87	23,88	26,24
2019	21,97	2,69	15,2	19,88	22,49	24,27	26,49

Figura 12: Distribuição das Temperaturas Mínimas dos Municípios



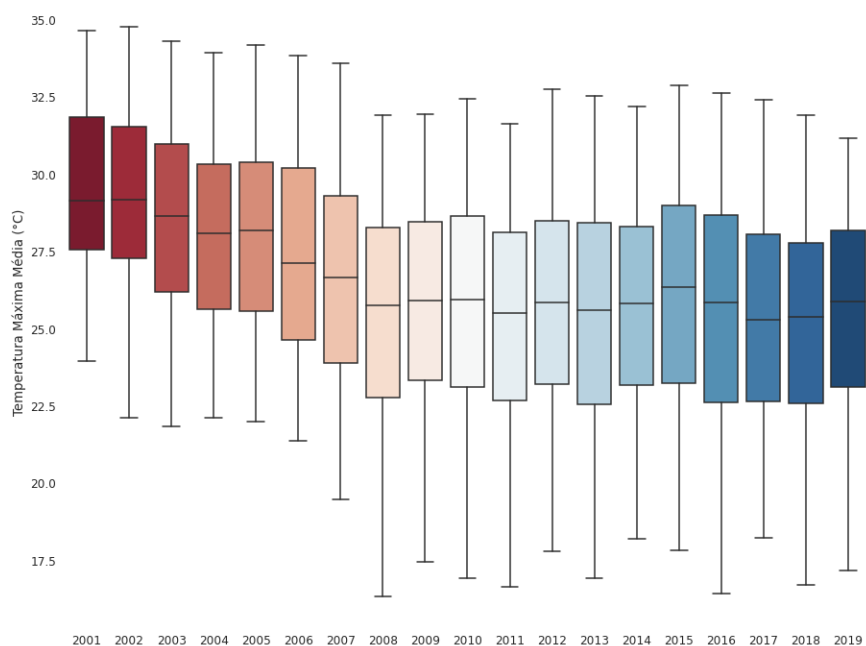
Por outro lado, enquanto as Temperaturas Mínimas Médias se elevaram, as Temperaturas Máximas Médias diminuíram de magnitude e atingiram níveis bastante interessantes para a proliferação dos vetores, constando ao final de 2019, conforme pode ser observado na Figura 13 e Tabela 6, uma temperatura média de 26 °C. E a consequência conjunta dessas evoluções é uma amplitude térmica menor e temperaturas compensadas médias estáveis, confirmando a conclusão citada.

Tabela 6: Temperatura Máxima Média (°C)

<i>Ano</i>	<i>Média</i>	<i>Dp</i>	<i>Min</i>	25%	50%	75%	<i>Max</i>
2001	29,48	2,52	23,97	27,56	29,15	31,87	34,64
2002	29,06	3,09	22,14	27,27	29,17	31,56	34,77
2003	28,49	3,13	21,85	26,19	28,65	30,99	34,31
2004	28,01	3,04	22,11	25,63	28,09	30,32	33,93
2005	28,17	3,11	22,02	25,58	28,17	30,39	34,19
2006	27,46	3,28	21,39	24,66	27,14	30,2	33,85
2007	26,65	3,5	19,5	23,9	26,67	29,31	33,58
2008	25,48	3,53	16,34	22,79	25,77	28,27	31,93
2009	25,66	3,23	17,46	23,34	25,93	28,46	31,95
2010	25,79	3,57	16,93	23,14	25,94	28,66	32,46
2011	25,30	3,36	16,65	22,69	25,53	28,14	31,64
2012	25,92	3,28	17,82	23,2	25,85	28,49	32,77
2013	25,51	3,61	16,95	22,56	25,61	28,44	32,55
2014	25,75	3,18	18,21	23,19	25,82	28,33	32,18
2015	26,14	3,46	17,82	23,23	26,36	28,98	32,89
2016	25,69	3,76	16,45	22,64	25,86	28,68	32,63
2017	25,53	3,28	18,23	22,67	25,31	28,07	32,42
2018	25,21	3,36	16,71	22,6	25,4	27,8	31,91
2019	25,54	3,19	17,19	23,11	25,9	28,18	31,18

Nas Figuras 14 e 15 podem ser observadas as evoluções das distribuições da Temperatura Compensada Média e da Amplitude Térmica dos municípios. Observe a presença de outliers e que mesmo com eles, ao final de 2019, por exemplo, a amplitude térmica não varia mais que 10 °C, algo comum em 2001. Uma queda abrupta de comportamento pode ser observado em 2006 e 2007.

Figura 13: Distribuição das Temperaturas Máximas dos Municípios



O fenômeno do El Niño ocorrido em 2006 e 2015, por outro lado, produziu seus efeitos no Brasil através da elevação da temperatura média realizada na região Sudeste. Esse efeito em 2006 terminou sendo observado com maior destaque na Figura 15, devido à grande proporção de municípios na região, cerca de 30% do total, o que eleva o impacto causado pelo fenômeno climático.

Figura 14: Distribuição das Temperaturas Compensadas dos Municípios

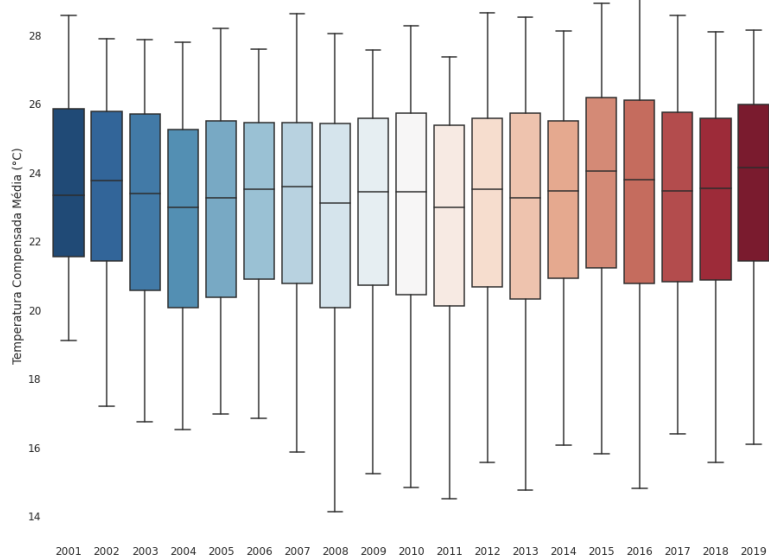
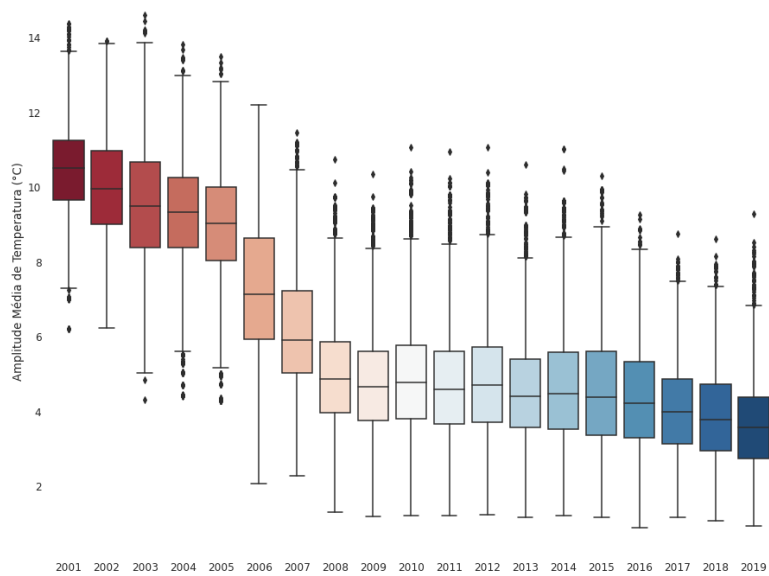


Figura 15: Distribuição das Amplitudes de Temperatura dos Municípios



Ainda sobre as variáveis climáticas é possível identificar nas figuras abaixo como estão distribuídos os pontos, que representam duas dimensões dos municípios e como estão classificados, se possuem ou não casos prováveis de infecção. A primeira relação analisada é a da Figura 16. Nela a região que compreende temperaturas mínimas entre 20 °C e 26 °C e níveis de precipitação total entre 0 mm e 150mm apontam uma presença significativa de municípios infectados. Tal fato pode ser verificado também a partir do auxílio da Tabela 7.

Tabela 7: Precipitação Total (mm)

<i>Ano</i>	<i>Média</i>	<i>Dp</i>	<i>Min</i>	25%	50%	75%	<i>Max</i>
2001	94,64	31,71	34,71	75,21	91,89	107,74	226,97
2002	105,54	35,8	44,1	79,08	95,06	128,62	244,69
2003	93,32	30,99	23,16	73,89	94,26	110,62	218,43
2004	103,49	25,26	40,06	87,22	104,45	118,69	198,82
2005	99,16	30,92	34,58	74,29	99,02	119,31	218,27
2006	73,44	32,33	9,8	55,62	66,69	80,29	226,98
2007	55,56	28,76	8,28	36,58	50,03	64,93	217,21
2008	48,47	22,4	0,13	33,19	45,16	59,56	186,58
2009	53,57	23,05	0,15	35,86	51,73	67,72	198,23
2010	41,80	17,81	0,09	29,93	39,53	52,18	173,07
2011	48,05	21,61	0,12	34,95	45,34	60,05	175,52
2012	34,03	18,57	0,13	21,53	33,28	44,1	189,06
2013	40,95	18,98	0,14	28,06	39,14	50,58	180,5
2014	36,97	19,1	0,13	24,15	33,84	46,81	168,11
2015	35,38	18,67	0,13	22,08	33,96	46,63	136,83
2016	32,82	16,4	0,11	22,66	31,05	42,07	134,31
2017	30,77	17,09	0,09	19,23	29,5	38,53	132,73
2018	30,12	17,92	0,1	17,85	29,01	38,92	135,53
2019	32,42	26,75	0,07	16,61	26,79	39,8	188,15

Números similares e raciocínio análogo respectivamente podem ser vistos nas Figuras 16 e 17, que apresentam as relações entre Temperaturas e Precipitação.

Figura 16: Temperatura Mínima vs Precipitação Total

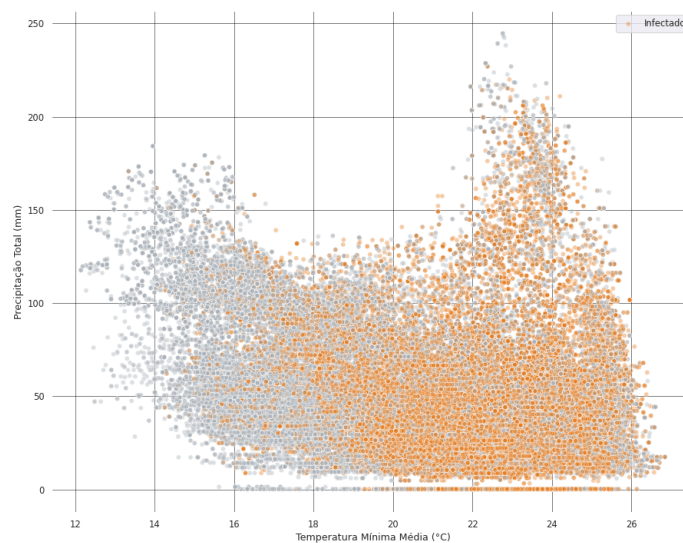
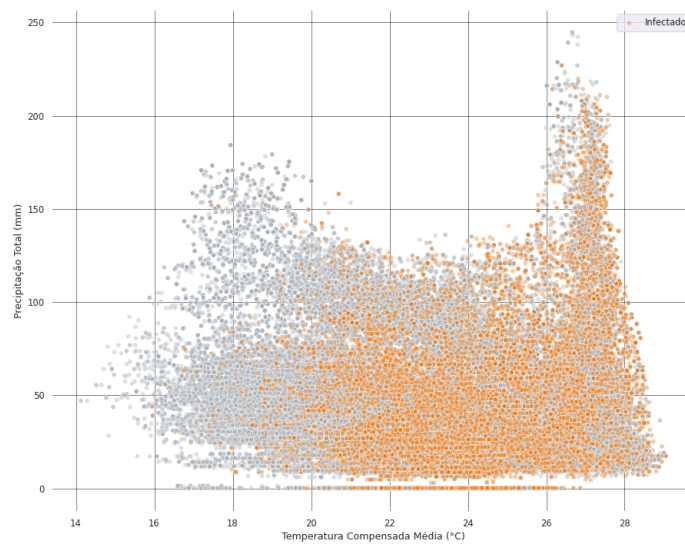


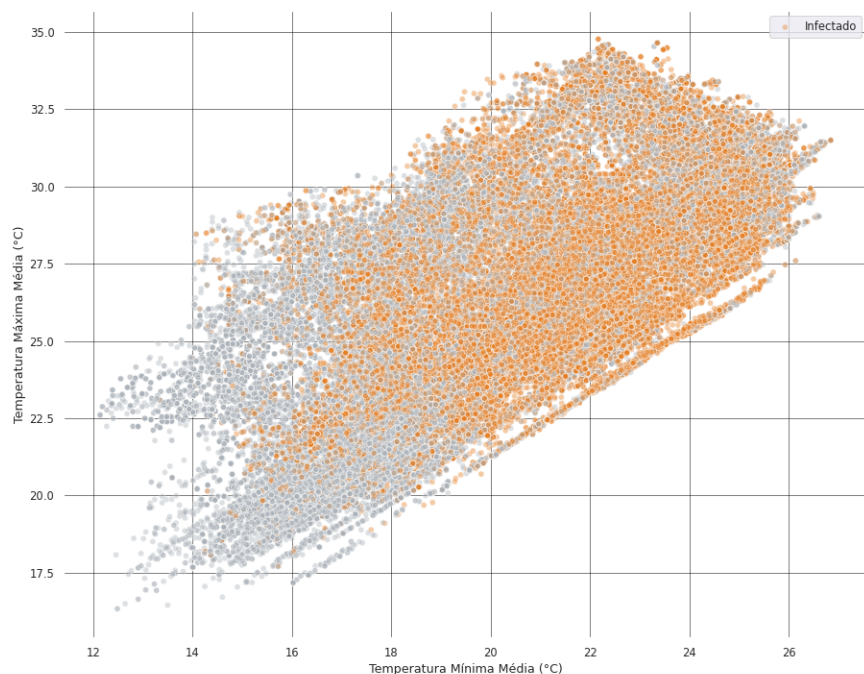
Figura 17: Temperatura Compensada Média vs Precipitação Total



A Figura 18 possui a relação entre as variáveis de Temperatura Mínima Média e Temperatura Máxima Média, cuja nuvem de pontos apresenta uma

tendência linear relativamente visível e com a massa de pontos classificados positivamente com a presença da Dengue, acumulada no canto superior a direita. Isto indica que temperaturas mínimas elevadas e temperaturas máximas medianas estão bastante associadas aos municípios onde ocorrem infecção.

Figura 18: Relação entre Temperatura Mínima e Temperatura Máxima



3.5 Influência dos Biomas

O Brasil é formado por seis biomas distintos: Amazônia, Caatinga, Cerrado, Mata Atlântica, Pampa e Pantanal. Os Biomas por sua vez são unidades biológicas ou espaços geográficos cujas características específicas são definidas pelo macroclima, a fitofisionomia, o solo e a altitude, dentre outros critérios. São categorias de ecossistemas, habitats ou comunidades biológicas com certo nível de homogeneidade

A vegetação é um dos componentes mais importantes da biota, e relevantes na classificação do Bioma. A sua utilização enquanto variável para avaliar a existência de nichos ambientais dos vetores transmissores da Dengue se torna justificada, pois os microhabitats naturais gerados por tais tipos de vegetação são possíveis criadouros desses vetores. Os dois Biomas com densidades de vegetação elevadas e mais propícios ao que foi dito anteriormente são: a Mata Atlântica e a Amazônia.

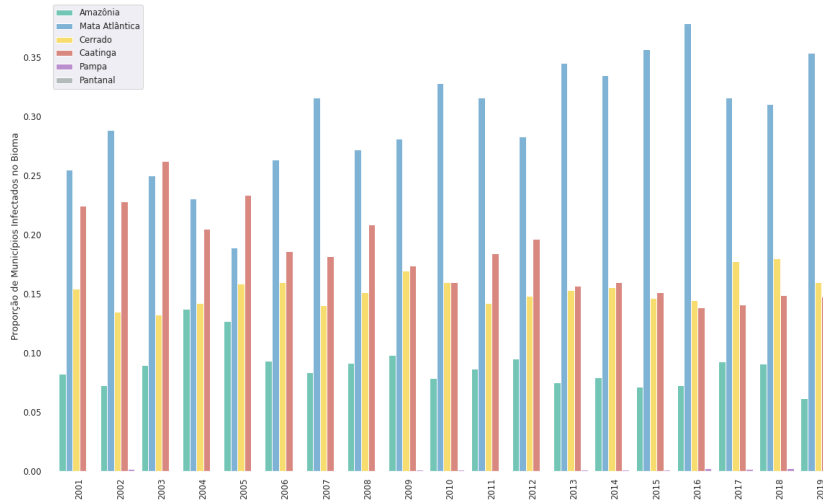
Na Figura 19 podemos verificar que, conforme esperado, a Mata Atlântica aparece com cerca de 30% dos seus municípios infectados e talvez como algo surpreendente a Caatinga junto com o Cerrado disputando o segundo lugar como Biomas com maior proporção de municípios infectados.

Este último fato provavelmente ocorre pela sua relação com a quantidade de municípios que é menor em relação aos demais biomas e seus indicadores socioeconômicos, e menos pelos seus fatores climáticos, uma vez que os Biomas podem se comportar como clusters não tradicionais que associam muito mais que informações biológicas e climáticas dos municípios. Além disso a posição da Mata Atlântica se justifica não só pelas características biológicas do Bioma, mas também pela sua correlação com o Sudeste e suas características demográficas.

A Amazônia se manteve estável na proporção de municípios infectados e abaixo dos demais Biomas citados. Os dois últimos Biomas, Pampa e Pantanal, obtiveram proporções irrelevantes e que dificilmente são observados graficamente, conforme pode ser verificado.

Os Pampas são regiões de clima temperado compostas principalmente pela presença de gramíneas e localizado no sul da América do Sul, e por estes motivos estão em ambientes pouco favoráveis aos vetores transmissores da Dengue. Por outro lado o Pantanal apresenta uma savana estépica, alagada e com clima quente e úmido, localizado no centro-oeste brasileiro. Apesar de suas características não houve uma proporção significativa de municípios infectados, e este fato deve estar relacionado ao número de municípios em sua abrangência e suas características demográficas.

Figura 19: Proporção de Municípios Infectados por Bioma



3.6 Correlação entre as variáveis

A maneira inicial mais comum de avaliar as variáveis independentes e selecioná-las para a modelagem consta em observar as correlações entre elas e com as variáveis dependentes.

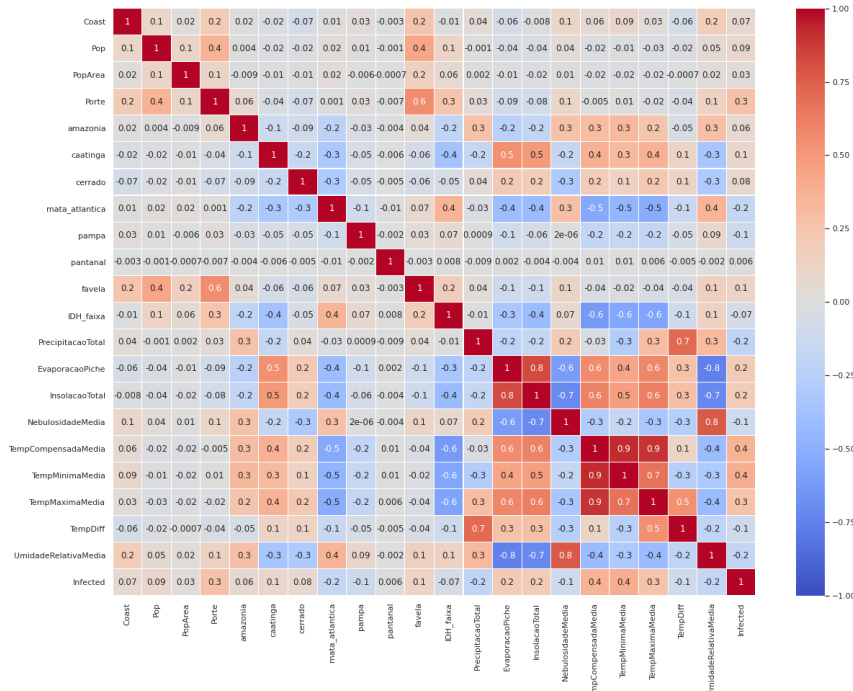
Os coeficientes de correlação de Pearson podem ser observados na Figura abaixo, e através deles observar quais variáveis podem apresentar relevância ou não para o modelo. Variáveis independentes muito correlacionadas entre si, apesar de não causar nenhum problema técnico em modelos baseados em árvores, comuns em modelos lineares, podem ser inúteis para o aprimoramento do caminho de decisão. Nesse sentido, nenhuma nova criação de ramos e novos nós serão feitos, pois outra variável já faz um caminho de decisão similar.

Segundo o que pode ser observado na Figura 20, as variáveis climáticas possuem entre si certo grau de correlação e entre elas e as variáveis de Bioma conforme era esperado. Por exemplo, insolação e umidade relativa possuem causa e efeito lógicos. Outro exemplo é a umidade relativa e Mata Atlântica, com cerca de 0.4 de correlação.

A variável dependente ***Infected*** possui um grau de correlação em torno

de 0.4 com as variáveis de temperatura e de Porte, o que pode em uma primeira análise indicar importância delas para a modelagem. As demais apresentam correlação em menor grau, o que pode ser bom em termos de controle para o modelo.

Figura 20: Correlação entre as variáveis selecionadas



4 Referencial Teórico

A área de Ciência de dados sofreu enorme avanço nos últimos anos devido a três fatores principais: Os avanços nas técnicas de Machine Learning e de outros modelos matemáticos aplicados em Big Data; Os volumes de dados disponíveis e armazenados diariamente de diferentes dispositivos; O avanço do processamento e do armazenamento de dados proporcionado pela maior capacidade computacional e desenvolvimento tecnológico da informática. As aplicações proporcionadas pela área de Ciência de Dados são inúmeras e

se estendem desde diagnósticos médicos até previsão de preço de ativos em bolsa de valores. O trabalho apresentado aqui neste estudo, utilizará técnicas de Machine Learning para a previsão de nichos ambientais do vetor do vírus da Dengue. Para tanto seis referências principais foram utilizadas: [Friedman et al., 2001], [Bishop, 2006], [Abu-Mostafa et al., 2012], [Chen and Guestrin, 2016a], [Ke et al., 2017] e [Prokhorenkova et al., 2018].

Existem dois fatores que possibilitam o sucesso dessas aplicações. O primeiro se define pelo uso correto de modelos matemáticos (ou estatísticos) que possam capturar a dependência complexa entre as variáveis de estudo, e o segundo, seria a existência de sistemas escaláveis de aprendizado que consigam treinar de forma eficiente nas grandes bases de dados. Por estes motivos foram selecionados algumas técnicas de machine learning para serem utilizadas na modelagem de classificação dos nichos dos vetores do vírus da Dengue.

Esta seção visa elucidar as técnicas selecionadas. As 3 técnicas de machine learning, cujas complexidades atendem as expectativas do problema, tanto em termos de complexidade do modelo quanto em eficiência de processamento do treinamento são:

- XGBoost (Extreme Gradient Boost) ^{7 8}
- LightGBM (Light Gradient Boost Machines) ⁹
- CatBoost (Categorical Boost) ¹⁰

Antes de explicar cada um dos três modelos separadamente é interessante nivelar e enfatizar alguns conceitos importantes que estão relacionados ao grupo de modelos ao qual eles estão contidos, que é o Gradient Boosting Regression Trees (GBRT).

4.1 Decision Trees

Existem diversos modelos simples dedicados a dividir o espaço em regiões cúbicas, cujas regiões estão alinhadas com os eixos, e então classificar ou

⁷Documentação do pacote <https://xgboost.readthedocs.io/en/latest/python>

⁸Criado a partir de projeto de pesquisa de Tianqi Chen como parte do grupo Distributed (Deep) Machine Learning Community (DMLC)

⁹Documentação do pacote <https://catboost.ai/docs>

¹⁰Documentação do pacote <https://lightgbm.readthedocs.io>

obter outra medida a partir dos pontos inseridos dentro destas regiões. O método denominado por Decision Trees, é um processo que de maneira resumida consegue selecionar um modelo específico dado um input x , através de decisões sequenciais correspondendo a estrutura de uma árvore, no qual a cada nó (node) ocorrem a divisão de novos ramos (branches) gerando novos nós e ramos novamente. Para cada novo input x , é determinado qual região ele será incluído, começando pelo nó inicial, no topo da árvore e seguindo o caminho dos nós e ramos que são estabelecidos pelo modelo e pelos critérios de decisão em cada nó.

As Decision Trees não são modelos gráficos probabilísticos, mas modelos não paramétricos e não lineares que conseguem capturar certo nível de complexidade entre as variáveis de input e output, também denominada de target. Uma grande vantagem desse modelo é a sua interpretabilidade tanto por apresentar de maneira clara as sequências de decisões para a classificação ou regressão do input, e também por conseguir apresentar as importâncias das variáveis na modelagem. Uma observação válida a respeito dessa importância das variáveis deve ser mencionada, que reside no fato de que ela informa apenas que a variável foi ou não escolhida para o processo de decisão e não informa de acordo com a sua variação o sentido da variação do output, tal como é feito nos modelos lineares. Outra vantagem dos modelos de Decision Trees reside no fato de que não é necessário realizar algum tipo de normalização nos dados, pois o modelo é invariante a escala, inclusive o algoritmo funciona de maneira adequada com features que são completamente diferentes que podem ser contínuas ou dummies.

O processo de construção do modelo de Decision Trees segue algumas etapas. Para realizar o treinamento desse modelo a partir de um conjunto de dados de treino, precisa-se determinar a estrutura da árvore, incluindo qual variável deve ser escolhida primeiro, quando dividir em mais nós, qual parâmetro escolher para essa divisão e quando parar a divisão e encerrar o crescimento da árvore.

Considere um problema de regressão. Neste problema determinar a estrutura da árvore a partir de um número fixo de nós seria muito custoso computacionalmente, pois calcular a estrutura ótima através das diversas possibilidades de minimização da soma dos quadrados dos erros é combinatoriamente impraticável. Para contornar este problema é utilizado uma *greedy optimization* começando pelo nó inicial, correspondendo ao espaço total dos inputs e então acrescentando nós gradualmente. A adição de nós costuma ser feita por procura exaustiva e é feita de maneira eficiente, sempre

procurando dividir o espaço dos inputs e gerando nós e novos ramos a partir deles. Com a estratégia da *greedy optimization* o próximo problema consta em quando parar o crescimento da árvore. Uma forma de lidar com isso consiste em parar assim que o erro residual das divisões ficar contido em alguma vizinhança predeterminada. Entretanto, empiricamente, é percebido que a inserção do erro nessa vizinhança não ocorre com frequência sendo sempre estimulado o crescimento da árvore, sem parar. Por este motivo é sempre indicado crescer a árvore formando uma estrutura de decisão bem longa, com um critério mais fraco de parada como o número de pontos associados com a folha (leaf) e então realizar o procedimento de *Pruning*.

O procedimento de *Pruning* consiste em estabelecer um critério que possa balancear o erro residual com alguma medida de complexidade do modelo. Os modelos de Decision Trees costumam tender ao overfitting e esse método consegue tratar parcialmente este problema. Considere uma previsão da variável alvo (target) y , indexada pela folha t , a partir de um vetor D -dimensional $x = (x_1, \dots, x_D)^t$ composta por variáveis independentes (inputs). Denote também por T_0 a árvore inicial que deve ser submetida ao *Pruning*, chegando à árvore T . Considere que as folhas são indexadas por $\tau = 1, \dots, |T|$, com τ representando uma região R_τ do espaço dos inputs contendo N_τ pontos e $|T|$ a quantidade de folhas da árvore. Então a previsão ótima da região é dada por :

$$y_t = \frac{1}{N_\tau} \sum_{x_n \in R_\tau} t_n \quad (1)$$

E sendo definido $Q_\tau(T) = \sum_{x_n \in R_\tau} (t_n - y_\tau)^2$ temos que o critério para realizar o *Pruning* é dado por:

$$c(T) = \sum_{\tau=1}^{|T|} Q_\tau(T) + \lambda |T| \quad (2)$$

O parâmetro λ pode ser entendido como um parâmetro de regularização que determina o trade-off entre o resíduo da soma dos quadrados dos erros e a complexidade do modelo. Para os modelos de classificação o problema é similar e o único ponto importante que é modificado é a medida de erro, que é substituída pela Cross-entropy ou pelo Índice de Gini. Se definirmos por $p_{\tau k}$ a proporção de dados na região R_τ classificados na classe k então tais métricas são descritas respectivamente nas equações abaixo:

$$Q_{\tau}(T) = \sum_{k=1}^K p_{\tau k} \ln p_{\tau k} \quad (3)$$

$$Q_{\tau}(T) = \sum_{k=1}^K p_{\tau k} (1 - p_{\tau k}) \quad (4)$$

Essas métricas incentivam a formação de regiões no qual uma proporção grande dos dados são incluídas em apenas uma classe. Além disso, elas são métricas melhores que a taxa de classificação errada por serem mais sensíveis às probabilidades dos nós e por serem diferenciáveis, permitindo a atuação de algoritmos de otimização baseados em gradientes.

4.2 Ensemble Learning

Ensemble Learning é o processo pelo qual diversos modelos são gerados de forma estratégica e combinados para resolver um problema particular em Inteligência Artificial. A estratégia que agrega, combina e diversifica modelos para obter um como resultado final consegue associar de maneira eficiente diferenças nas escolhas dos dados, diferenças nas hipóteses, diferenças nas técnicas de modelagem e diferenças na escolha dos parâmetros iniciais, que os modelos possuem intrinsecamente.

O erro de qualquer modelo genérico $f(x)$, com previsões denominadas por $\hat{f}(x)$, pode ser separado em termos segundo o que é representado na equação abaixo:

$$Erro(x) = (E[\hat{f}(x)] - f(x))^2 + E(\hat{f}(x) - E[\hat{f}(x)])^2 + \sigma_{\epsilon}^2 \quad (5)$$

O Erro acima corresponde ao primeiro termo sendo o viés ao quadrado somado ao segundo termo, a variância, e um último termo, o erro irreduzível do modelo. O Viés alto representa um modelo que está performando abaixo do ideal, sendo possivelmente originada pelas hipóteses adotadas que são fracas. A Variância elevada implica que as extrapolações do modelo para dados de fora do treinamento possuirão uma performance péssima. Este problema é denominado por Overfitting. Os modelos de *Ensemble Learning* conseguem balancear esse trade-off de forma que nem o Viés e Variância sejam demasiadamente elevados.

E esse gerenciamento de complexidade versus generalização é uma vantagem desse tipo de técnica. Outros benefícios dessa técnica residem nas

elevadas capacidades preditivas, sendo refletidas em acurácias melhores, e a estabilidade do modelo que consegue evitar uma seleção de modelo viesada. Por outro lado, as desvantagens desses modelos são observados na redução da interpretabilidade dos resultados do modelo, algo que pode ser compensado caso seja baseado em modelos de árvores de decisão, e pelo tempo de produção e computação dos modelos, o que dificulta sua aplicação em problemas que exigem modelagens em tempo real.

Os algoritmos mais comuns de aplicação da técnica de *Ensemble Learning*, além das aplicações de regras simples tais como votação ou média dos resultados, são o Bagging (*bootstrap aggregating*) e o Boosting. Entretanto, existem uma série de outros métodos tais como Bayesian model combination, Bayes optimal classifier, Bucket of models, Stacking e *etc.* O Bagging relaciona-se mais a uma combinação dos modelos de forma paralela e o Boosting de uma maneira sequencial.

O Bagging costuma atribuir para cada modelo constituinte do conjunto um peso igual, mas para promover a melhora de performance, na generalização, são utilizadas amostras diferentes da base de dados de treino para cada treinamento de cada modelo. As respostas de cada modelo são então guardadas e é possível obter uma resposta final única a partir tanto da média das respostas quanto outras formas de agregação. Isto diminui a variância do modelo final e reduz o overfitting, gerando resultados satisfatórios em termos de acurácia [Breiman, 1996]. O exemplo mais famoso de Bagging é o Random Forest.

O Boosting é similar ao Bagging, no sentido que cria um conjunto de classificadores treinados a partir de reamostragem dos dados. Entretanto, ele o faz para gerar classificadores fracos (*weak learners*) associados a pesos (pesos maiores para classificadores com maior erro), que serão de forma sequencial melhorados conforme as iterações do algoritmo. Ao final são produzidos classificadores fortes (*strong learners*). O algoritmo mais comum é o AdaBoost (*Adaptive Boosting*), ver [Schapire, 1999]. A interpretação do Boosting enquanto um algoritmo de otimização sequencial de um modelo aditivo sob uma função perda bem definida (como por exemplo o erro exponencial) consegue fornecer extensões interessantes para a área, pois são bem flexíveis.

4.3 Gradient Tree Boosting

O *Gradient Tree Boosting* (abreviado por GTB), assim como denominações similares, tais como o *Gradient Boosting Regression Tree* ou *Gradient Bo-*

osting Decision Tree , é um tipo de técnica sob o contexto de *Ensemble Learning*. Nele ocorrem a combinação de diversas árvores de decisão ou a melhora de uma única árvore de maneira sequencial para criar um modelo mais performático.

A funcionalidade mais comum dos métodos reside resumidamente em organizar as árvores de decisão de maneira sequencial, fazendo com que as próximas árvores corrijam os erros das anteriores. Por padrão nenhuma randomização é realizada para tratar do overfitting, mas sim usa-se o *Prunning*. A idéia importante que consegue destacar o método é justamente o fato de combinar modelos simples como árvores superficiais e gerar um resultado complexo que possui excelente performance em termos de acurácia.

Segundo a notação e estudo de [Kanamori et al., 2007] é possível estabelecer um algoritmo genérico para o Boosting derivado do metodo do gradiente a partir de uma função perda genérica L . Defina uma amostra dada por $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, onde $x_i \in X$ e $y_i \in \{1, -1\}$. Considere um conjunto de hipóteses dado por $H = \{h_t : X \rightarrow \{1, -1\} | t = 1, 2, 3, \dots\}$ e um preditor linear como combinação dos conjuntos de hipóteses dado por $H_\alpha(x) = \sum_{t=1}^{\infty} \alpha_t h_t(x)$, com $\alpha_t \in R$ para todo t . A previsão da classe do input $x \in X$ é dada por $sign(H_\alpha(x))$. Seja Z igual a $X \times \{1, -1\}$, então para uma medida de probabilidade Q em Z o risco pode ser escrito como $R_L(Q, H) = \int_Z Q(dz)(-yH(x))$. Em situações práticas a medida de probabilidade Q é substituída por uma distribuição empírica \hat{P}^{11} de um conjunto de amostras, originando o seguinte risco: $R_L(\hat{P}, H) = \frac{1}{n} \sum_{i=1}^n (-y_i H(x_i))$. O problema se torna então minimizar o $R_L(Q, H)$ com respeito a H . Esse problema se traduz então em calcular $argmin_{h \in H} \int_Z Q(dz) L'(-yH(x)) I(-y \neq h(x))$. Com isso o preditor $H(x)$ é atualizado para $H(x) + \alpha h(x)$. O resumo do algoritmo pode ser visualizado no esquema abaixo.

¹¹Para maiores detalhes das definições e cálculos averiguar a fonte citada [Kanamori et al., 2007]

Algoritmo 1: Boosting pelo método do gradiente

Inputs: Distribuição conjunta Q em $Z = X \times \{1, -1\}$ e um preditor inicial H^0 . Na prática o Q é a distribuição empírica da amostra.

for $m = 1, \dots, M$ **do**

1. Encontre a hipótese $h^{(m)} \in H$ tal que
 $h^{(m)} = \operatorname{argmin}_{h \in H} \int_Z Q(dz) L'(-yH^{(m-1)}(x)) I(-y \neq h(x)).$
2. Encontre o coeficiente $\alpha^{(m)} \in R$ tal que
 $\alpha^{(m)} = \operatorname{argmin}_{\alpha \in R} R_L(Q, H^{(m-1)} + \alpha h^{(m)}).$
3. Atualizar o preditor $H^{(m)}(x) = H^{(m-1)}(x) + \alpha^{(m)} h^{(m)}(x).$

end

Output: Preditor estimado $H^{(M)}$

Fonte: Modificado a partir de [Kanamori et al., 2007]

A aplicação mais comum da técnica de Boosting pode ser vista no *Ada-Boost* (Adaptative Boost). Segue a explicação resumida do método. Nela, as amostras recebem um peso inicial igual, denotado por w_i que posteriormente vai se modificando a cada treinamento caso não ocorra a classificação correta na variável dependente.

O preditor é descrito por $H(x) = \sum_{m=1}^M c_m \delta_m(x)$, no qual c_m é um parâmetro de importância dos classificadores fracos $\delta_0, \delta_1, \dots, \delta_M$, no caso árvores de decisão. A função perda do método é exponencial, isto é, $L(y, f(x)) = \exp\{-yf(x)\}$. O preditor na prática se torna $\hat{f}(x) = \sum_{m=1}^M c_m \delta_m(x, \gamma_m)$, com sua respectiva atualização dada por $\hat{f}_m(x) = \hat{f}_{m-1}(x) + c_m \delta_m(x, \gamma_m)$. O problema de minimização, para fornecer os valores dos parâmetros (c_m, γ_m) é o encontrado na equação 6 abaixo.

$$\begin{aligned} (c_m, \gamma_m) &= \operatorname{argmin}_{c, \gamma} \sum_{i=1}^n \exp\{y_i(\hat{f}_{m-1}(x_i) + c\delta(x_i, \gamma))\} \Rightarrow \\ (c_m, \gamma_m) &= \operatorname{argmin}_{c, \gamma} \sum_{i=1}^n w_i^m \exp\{-cy_i(\delta(x_i, \gamma))\} \end{aligned} \quad (6)$$

As soluções explícitas da equação acima bem como o algoritmo detalhado podem ser encontrados respectivamente em [Friedman et al., 2001] e [Bishop, 2006]. Concluindo o método, os pesos são então atualizados segundo a regra dada por $w_i^m = w_i^{m-1} \exp\{-c_m\}$.

A implementação do *Gradient Boosting Tree* baseado no *AdaBoost* realiza algumas modificações no processo de atualização, incluindo o gradiente conforme equação 7 abaixo, e ao invés de encontrar o melhor classificador procura encontrar a melhor taxa de aprendizado γ , conforme equação 8.

$$\hat{f}_m(x) = \hat{f}_{m-1}(x) + \gamma_m \sum_{i=1}^n \nabla_{f_{m-1}} L(y_i, f_{m-1}(x_i)) \quad (7)$$

$$\min_{\gamma} \sum_{i=1}^n L(y_i, f_{m-1}(x) - \gamma \frac{(y_i, f_{m-1}(x))}{\partial f_{m-1}(x_i)}) \quad (8)$$

A observação final sobre o método reside na sua inicialização, que é efetivada por uma constante dada por $\hat{f}_0 = \operatorname{argmin}_{\gamma} \sum_{i=1}^n L(y_i, \gamma)$, significando que de início é criada uma árvore com apenas uma única folha.

4.4 XGBoost

O XGBoost, possui a complexidade dos modelos de *Ensemble Learning*, mas consegue ter uma performance melhor para bases de dados maiores, pois é um método escalável de aprendizado. Este método consegue executar com uma velocidade dez vezes mais rápido do que diversos modelos existentes (por exemplo o Random Forest, ver [Chen and Guestrin, 2016a]) e consegue utilizar de forma eficiente os recursos de memória da máquina, que costumam ser limitantes naturais destas técnicas.

As inovações geradas por este método além dos procedimentos de manipulação dos pesos dos quantis de cada instância (denominados de *Weighted Quantile Sketch*), que surgem no aprendizado do modelo, e também do processamento em blocos dos dados, se concentra na introdução de uma regularização na função custo. Para entender melhor como ela é realizada, considere um conjunto de dados D , com n observações e m variáveis, que pode ser descrito como $D = (x_i, y_i)$ onde $|D| = n$ e $x_i \in R^m$ e $y_i \in R$. Um *Ensemble Tree Model*¹² utiliza K funções aditivas para prever o output. Isto

¹²Esses modelos também são denominados por CART, termo cunhado por Leo Breiman, ver [Grajski et al., 1986]

é representado pela seguinte equação:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^K f_k(x_i) \quad (9)$$

onde temos que o espaço das árvores de regressão é denominado por $F = \{f(x) = w_{q(x)}\}$ com a estrutura de cada árvore responsável por mapear um exemplo para a folha correspondente denominada por q sendo $q : R^m \rightarrow T, w \in R^T$. De maneira resumida ¹³, cada f_k corresponde a uma estrutura de árvore independente q com pesos das folhas determinado por w . Esse peso w_i representa a pontuação da folha i . Para realizar o treinamento do modelo é comum resolver o seguinte problema com função perda regularizada:

$$\min_{w_k} [L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \gamma T + \frac{1}{2} \lambda |w_k|^2] \quad (10)$$

sendo que $l(\hat{y}_i, y_i)$ costuma ser definida com uma função perda convexa para cada output e sua devida observação real nos dados, e o termo incluído na segunda soma costuma ser considerado o termo de regularização $R(f_k) = \gamma T + \frac{1}{2} \lambda |w_k|^2$ que penaliza a complexidade do modelo. Retirando-se o fator de regularização o método torna-se igual ao *Gradient Tree Boosting*.

Através de [Chen and Guestrin, 2016a] segue uma generalização didática do modelo. A partir do problema de minimização apresentado na equação 10 é possível verificar que o mesmo não consegue ser otimizado utilizando-se apenas de métodos tradicionais de otimização no espaço euclidiano. Desta maneira o modelo é treinado de uma forma aditiva. Formalizando a sentença, seja \hat{y}_i^t a previsão da instância i na iteração t , torna-se necessário adicionar a função f_t para minimizar a seguinte função objetivo:

$$L^t = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + R(f_t) \quad (11)$$

essa função adicionada é estipulada de forma gradual, para que seja verificada uma melhora no modelo segundo o problema de minimização citado. É tipicamente uma estratégia de *greedy algorithms*. Uma aproximação de segunda ordem pode ser feita para tratar o problema de maneira mais geral:

¹³No artigo principal sobre XGBoost tais definições são abordadas de maneira mais ampla, ver [Chen and Guestrin, 2016a]

$$\begin{aligned}
f_t(x_i) &\approx \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} f_t(x_i) + \frac{1}{2} \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} f_t^2(x_i) = p_{1i} f_t(x_i) + \frac{1}{2} p_{2i} f_t^2(x_i) \Rightarrow \\
\Rightarrow L^t &\approx \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + p_{1i} f_t(x_i) + \frac{1}{2} p_{2i} f_t^2(x_i)] + R(f_t) \quad (12)
\end{aligned}$$

onde foi denominado p_{1i} e p_{2i} como as derivadas parciais de primeira e segunda ordem respectivamente. Desta maneira podemos expandir a equação acima para a seguinte equação abaixo:

$$\begin{aligned}
L^t &\approx \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + p_{1i} f_t(x_i) + \frac{1}{2} p_{2i} f_t^2(x_i)] + \gamma T + \frac{1}{2} \lambda |w_t|^2 \Rightarrow \\
\Rightarrow L^t &\approx \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + p_{1i} f_t(x_i) + \frac{1}{2} p_{2i} f_t^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (13)
\end{aligned}$$

agora excluindo o primeiro termo do somatório pois ele é constante para o nosso problema de minimização teremos que nossa nova função perda pode ser entendida como:

$$\tilde{L}^t = \sum_{i=1}^n [p_{1i} f_t(x_i) + \frac{1}{2} p_{2i} f_t^2(x_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (14)$$

fazendo alguns ajustes nos somatórios após definir a instância na folha j como $I_j = \{i | q(x_i) = j\}$ chegamos em:

$$\tilde{L}^t = \sum_{j=1}^T [(\sum_{i \in I_j} p_{1i}) w_j + \frac{1}{2} (\sum_{i \in I_j} p_{2i}) w_j^2] + \gamma T \quad (15)$$

ao fixarmos uma estrutura $q(x)$ podemos computar o peso ótimo w_j^* da folha j por:

$$w_j^* = - \frac{\sum_{i \in I_j} p_{1i}}{\sum_{i \in I_j} p_{2i} + \lambda} \quad (16)$$

e o seu valor ótimo é então dado por:

$$\tilde{L}^t(q) = -\frac{1}{2} \sum_{j=1}^T \frac{(\sum_{i \in I_j} p_{1i})^2}{\sum_{i \in I_j} p_{2i} + \lambda} + \gamma T \quad (17)$$

A equação 17 pode ser então utilizada para avaliar a qualidade da estrutura de árvore q . Essa equação termina sendo caracterizada como função *score* e assim essa pontuação pode ser entendida como uma medida de impureza para avaliar a árvore de decisão. Geralmente as árvores de decisão não são enumeradas, apenas as que apresentam uma certa qualidade são consideradas e são então melhoradas no processo de Boosting.

Sobre a escolha dos possíveis candidatos para o split, vale ressaltar que o método não utiliza as maneiras comuns como o *Cross-entropy* ou *Gini*, mas sim uma forma especial de ganho informacional descrito na equação 18. A divisão ocorre da seguinte maneira exemplificada: suponha que existam duas divisões da instância em dois nós após o split: I_r e I_l . Então temos que a função de redução da função perda será dada por:

$$L_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_r} p_{1i})^2}{\sum_{i \in I_r} p_{2i} + \lambda} + \frac{(\sum_{i \in I_l} p_{1i})^2}{\sum_{i \in I_l} p_{2i} + \lambda} - \frac{(\sum_{i \in (I_r \cup I_l)} p_{1i})^2}{\sum_{i \in (I_r \cup I_l)} p_{2i} + \lambda} \right] - \gamma \quad (18)$$

Apesar de ser um método de Boosting, e portanto possuir certa tendência para o overfitting, existem estratégias alternativas a regularização para diminuir a variância do modelo. Uma técnica possível é a de "shrinkage" apresentada por [\[Friedman, 2002\]](#). Trata-se da multiplicação de um fator similar a taxa de aprendizado (na otimização estocástica) nos novos pesos calibrados pelo processo de otimização anterior, o que reduz a influência de cada espaço de árvore e folhas para que as futuras árvores possam treinar e melhorar o modelo. Uma outra técnica possível é utilizar uma amostragem das features, algo similar ao que é feito no Random Forest.

4.5 LightGBM

Trata-se de uma implementação produzida pela Microsoft, pertencente ao grupo de modelos de GBDT, que procura resolver o problema de eficiência existente no momento do split da árvore de decisão gerado nos outros modelos similares. Esse problema surge quando para cada variável (*feature*) os modelos existentes de GBDT necessitam realizar uma leitura para todas as

instâncias dos dados com o objetivo de estimar o ganho informacional (*information gain*) de todos os possíveis pontos de separação (*split*), e realizar tal empreendimento é demasiado custoso em termos de tempo.

O método em questão procura resolver esse problema com duas técnicas denominadas de *Gradient-based One-Side Sampling* e *Exclusive Feature Bundling*, cujas siglas são respectivamente GOSS e EFB.

Em um modelo GBDT o aprendizado de uma função do espaço X^s ocorre por meio do espaço de gradientes G . Supondo um conjunto de treino denominado por $\{x_1, x_2, \dots, x_n\}$ onde cada x_i é um vetor com dimensão s no espaço X^s , temos que a cada iteração do método, os gradientes negativos da função perda com respeito aos *outputs* do modelo são denotados por $\{g_1, g_2, \dots, g_n\}$. O modelo então realiza o *split* de cada nó na variável mais informativa, aquela com maior ganho informacional, e isso é feito a partir da métrica de variância após *split*, definida na equação abaixo:

$$V_{j|O}(d) = \frac{1}{n_O} \left(\frac{(\sum_{x_i \in O; x_{ij} \leq d} g_i)^2}{n_{l|O}^j(d)} + \frac{(\sum_{x_i \in O; x_{ij} \geq d} g_i)^2}{n_{r|O}^j(d)} \right) \quad (19)$$

onde $n_O = \sum I[x_i \in O]$, $n_{l|O}^j(d) = \sum I[x_i \in O : x_{ij} \leq d]$, $n_{r|O}^j(d) = \sum I[x_i \in O : x_{ij} \geq d]$. Para a variável j , a árvore de decisão seleciona o valor ótimo $d_j^* = \operatorname{argmax}_d V_j(d)$ e calcula o maior ganho informacional $V_j(d_j^*)$.

Através do GOSS é possível excluir partes das instâncias geradas pelos dados com pequenos gradientes, e utilizar apenas o restante para a computação do ganho informacional. Verifica-se que como as instâncias dos dados com gradientes elevados desempenham um papel importante no cálculo do ganho informacional, o método de GOSS consegue obter uma acurácia estimada razoável do ganho informacional com apenas uma quantidade pequena dos dados.

O método GOSS de forma mais específica faz o que segue. Primeiro ele estabelece um *rank* nas instâncias de treinamento de acordo com os valores absolutos dos gradientes em ordem decrescente. Então são separados dois subconjuntos, um com as primeiras a instâncias denominado de A , e outro complementar com o restante. Desse último subconjunto é selecionada uma amostra aleatória de tamanho $b \times |A^c|$, o *split* ocorre então de acordo com a Variância descrita abaixo:

$$\tilde{V}_j(d) = \frac{1}{n} \left(\frac{(\sum_{x_i \in A_l} g_i + \frac{1-a}{b} \sum_{x_i \in B_l} g_i)^2}{n_l^j(d)} + \frac{(\sum_{x_i \in A_r} g_i + \frac{1-a}{b} \sum_{x_i \in B_r} g_i)^2}{n_r^j(d)} \right) \quad (20)$$

onde $A_l = \{x_i \in A : xij \leq d\}$, $A_r = \{x_i \in A : xij \geq d\}$, $B_l = \{x_i \in B : xij \leq d\}$, $B_r = \{x_i \in B : xij \geq d\}$, e o coeficiente $\frac{1-a}{b}$ é utilizado para normalizar a soma dos gradientes no conjunto B proveniente do conjunto A^c . Com a aplicação desta variância no momento do *split* o custo computacional e o tempo são reduzidos significativamente ¹⁴.

O outro método procura solucionar os problemas gerados por dados de alta dimensão que normalmente são bem esparsos. Através do EFB é possível agrupar de maneira mutuamente exclusiva as variáveis para reduzir o seu número. Apesar de que encontrar o agrupamento ótimo seja NP-hard, um *greedy algorithm* consegue se aproximar relativamente bem desse cálculo. Isso consegue então reduzir eficientemente o número de variáveis sem ferir a acurácia dos pontos de *split* demasiadamente.

4.6 CatBoost

Este é mais um método de machine learning inserido no grupo dos GBDT, mas originalmente destinado principalmente à modelagem de problemas de classificação (inclusive seu nome possui o termo *Cat* de *Categorical*), e que busca a utilização do método de boosting não-viesado para as variáveis categóricas. Foi produzido por um grupo de pesquisadores russos, pertencentes a Yandex, companhia de serviços de internet, responsável pelo maior motor de busca na internet russa.

Dois algoritmos que permitem o sucesso do CatBoost são o *ordered boosting*, uma alternativa ao algoritmo clássico, construído e conduzido por permutações, e um outro algoritmo de processamento das variáveis categóricas.

¹⁴Para a prova detalhada desta afirmação verificar página 4 de [Ke et al., 2017]

Algoritmo 2: *Ordered Boosting*

Inputs: $\{(x_k, y_k)\}_{k=1}^n, I$;
 $\sigma \leftarrow$ *permutação aleatória de* $[1, n]$;
 $M_i \leftarrow$ *for* $i = 1, 2, \dots, n$;
for $t \leftarrow$ *to* I **do**
 for $i \leftarrow 1$ *to* I **do**
 $r_i \leftarrow y_i - M_{\sigma(i)-1}(x_i)$;
 end
 for $i \leftarrow 1$ *to* I **do**
 $\Delta M \leftarrow \text{LearnModel}((x_j, r_j) : \sigma(j) \leq i)$;
 $M_i \leftarrow M_i + \Delta M$;
 end
end
Output: M_n

Fonte: Modificado a partir de [Prokhorenkova et al., 2018]

Ambos os algoritmos são uma forma de combater o problema denominado de *prediction shift*, problema este que ocorre quando a variável *target* costuma acompanhar os dados da base, mas estão distantes do valor real que deveriam possuir. Esse problema é comum nas implementações de GBDT causado pelo fenômeno de *target leakage*. O algoritmo de *Ordered Boosting* pode ser descrito como o que se encontra na página 6 de [Prokhorenkova et al., 2018] e destacado acima.

Através de uma pequena adaptação no algoritmo original apresentado na referência mencionada acima, é possível produzir um pseudo-algoritmo que descrevem bem o funcionamento da criação de árvores com CatBoost através do modo *Ordered*. Este se encontra no esquema abaixo. O CatBoost ao utilizar diversas permutações consegue diminuir a variância e isso oferece uma solução não apenas para o problema de *prediction shift*, mas também permite uma diminuição no *Overfitting*, gerando aprendizados mais capazes de generalizar os resultados.

Algoritmo 3: Criação de árvore com CatBoost - modo Ordered

Inputs: $M, \{(x_k, y_k)\}_{k=1}^n, I, \alpha, L, \{\sigma_i\}_{i=1}^s$;
 $\text{grad} \leftarrow \text{CalcGradiente}(L, M, y)$;
 $r \leftarrow \text{número aleatório de } (1, s)$;
 $G \leftarrow (\text{grad}_{r, \sigma_r(i)-1}(i) \text{ for } i = 1, 2, \dots, n)$;
 $T \leftarrow \text{árvore vazia}$;
for passo do procedimento top-down **do**
 for candidato ao split c **do**
 $T_c \leftarrow \text{adicionar split em } T$;
 $\Delta(i) \leftarrow \text{avg}(\text{grad}_{r, \sigma_r(i)-1}(p) \text{ for } p : \text{folha}_r(p) = \text{folha}_r(i), \sigma_r(p) \leq \sigma_r(i)) \text{ for } i = 1, 2, \dots, n$;
 $\text{perda}(T_c) \leftarrow \text{cos}(G)$;
 end
 $T \leftarrow \text{argmin}_{T_c}(\text{perda}(T_c))$;
end
 $M_{r', j}(i) \leftarrow M_{r', j}(i) - \alpha \text{avg}(\text{grad}_{r', j}(p) \text{ for } p : \text{folha}_{r'}(p) = \text{folha}_{r'}(i), \sigma_{r'}(p) \leq j) \text{ for } r' = 1, 2, \dots, s, i = 1, 2, \dots, n, j \geq \sigma_{r'}(i) - 1$;
Output: T, M

Fonte: Modificado a partir de [Prokhorenkova et al., 2018]

5 Metodologia

A metodologia utilizada no presente trabalho foi inspirado no modelo denominado por *Cross Industry Standard Process for Data Mining*, ou CRISP-DM. Nele, o processo se inicia pela formulação do problema seguido pela coleta, análise exploratória dos dados e tratamento. Logo após estarem completamente preparados, os dados são então inseridos em um modelo e avaliados segundo métricas de performance. Se forem abaixo do desejado, novos procedimentos de coleta e tratamento nos dados e consequente modificação no modelo ou nos seus parâmetros são realizados. O ciclo só se finaliza com a obtenção de um modelo que atinja a meta estabelecida anteriormente na formulação do problema.

Os objetivos deste estudo consistem em observar características diretas e indiretas na consecução de um nicho ambiental mais favorável aos vetores transmissores da Dengue. Para tanto até esta seção foram feitas as coletas e tratamento dos dados, que se utilizaram de diversas fontes e métodos tais como o KNN para lidar com problemas como ausência de dados e a necessidade de extrapolação.

Com os dados organizados e disponibilizados em formato tanto de cross-section como em painel e junto do referencial teórico exposto, o que se propõe é a utilização de métodos de Gradient Boosting Decision Trees para efetuar o objetivo.

Desta maneira foram adotadas duas estratégias principais:

1. Modelagem **Cross-Section**: Sua finalidade é a de entender as importâncias das variáveis na decisão sobre se haverá infecção pela Dengue no município ou não e prever esse fato. Para isso serão desconsideradas consequências temporais e cada município em um ano será uma nova observação, agregando uma base de dados única de todos os anos e municípios nesses anos.
2. Modelagem **Espaço-Temporal**: Com objetivo similar à Modelagem Cross-Section, essa modelagem possui de forma divergente uma abordagem de série temporal e se utiliza dos dados de anos anteriores como variáveis. Assim deseja-se realizar a previsão da infecção pela Dengue no município para os anos futuros.

Ambas as modelagens de classificação binária são construídas com um mesmo núcleo de variáveis escolhidas segundo critérios baseados em revisão de literatura ¹⁵, análise exploratória dos dados, realizada em seções anteriores, disponibilidade e granularidade de dados confiáveis.

A divisão dos dados em todos os modelos foi realizada na seguinte proporção: 30% para o teste e 70% para o treino. O método de avaliação dos modelos foi baseado na utilização da técnica de *Cross-Validation K-Fold* apenas nos dados separados para treino, com 10 *Folds*, onde temos a implementação do modelo para cada uma das $K - 1$ partes e testado na última parte restante. Apenas os dados de treino foram utilizados para escolhas de hiperparâmetros e para melhora do modelo. Em seguida, o modelo treinado

¹⁵Para maiores detalhes ver as referências, a introdução deste trabalho e o referencial teórico

foi aplicado no conjunto de dados de teste de maneira a avaliar a performance sem ocorrências de *Data Leakage*. Logo, a avaliação realizada apresentou uma característica mais robusta e menos sensível a seleções viesadas.

As métricas de erro utilizadas no processo de *boosting* dos modelos foi o *Log-Loss* ou também conhecido por *Binary Cross-Entropy*. As métricas conforme serão vistas nas seções posteriores se basearam não somente na acurácia, mas também na pontuação da métrica *F1-score*, *Recall*, *Precision*, *ROC/AUC* e por tempo de execução do algoritmo. Com o intuito de sempre trabalhar com as mesmas inicializações foi utilizada a *seed* fixa de valor 17.

Os resultados e detalhes mais específicos da implementação de cada uma podem ser observados na seção adiante, denominada de Resultados.

6 Resultados

Nesta seção serão abordadas as modelagens realizadas para prever os municípios com casos prováveis de infecção da Dengue, através das variáveis que representam em algum nível os melhores nichos dos vetores da Dengue. Para a modelagem *Cross-Section* foram totalizadas 105906 observações com 21 variáveis independentes, enquanto para modelagem *Espaço-Temporal* foram totalizadas 5574 observações com 27 variáveis independentes.

As modelagens foram realizadas segundo o auxílio de implementações feitos pelos pacotes em *Python* mencionados na seção Referencial Teórico. Os pacotes para organização dos dados, visualização dos dados, computação de matrizes e vetores e para as estatísticas descritivas e métricas dos modelos foram respectivamente *Pandas* ¹⁶, *Matplotlib* ¹⁷, *Seaborn* ¹⁸, *Geopandas* ¹⁹, *Numpy* ²⁰, *Scikit-Learn* ²¹.

6.1 Modelagem Cross-Section

As 21 variáveis estudadas na análise exploratória dos dados foram então aplicadas para avaliar a existência de casos nos municípios. A tabela abaixo

¹⁶<https://pandas.pydata.org/>

¹⁷<https://matplotlib.org/3.1.0/index.html>

¹⁸<https://seaborn.pydata.org/index.html>

¹⁹<https://geopandas.org/>

²⁰<https://numpy.org/>

²¹<https://scikit-learn.org/stable/>

representa os resultados dos modelos, com suas respectivas métricas de avaliação. Vale ressaltar que essas são métricas do *Cross-Validation* utilizadas para encontrar os melhores hiperparâmetros como a profundidade da árvore, número de árvores de decisão, função perda adequada entre outras.

Tabela 8: Avaliação dos modelos - Cross-Validation

<i>Métricas</i>	<i>XGBoost</i>	<i>LightGBM</i>	<i>CatBoost</i>
Fit Time (s)	308,36	26,15	694,48
Score Time (s)	2,69	2,50	7,67
Acurácia de Teste (%)	77,26	76,60	77,79
Acurácia de Treino (%)	80,77	77,75	80,42
F1 de Teste (%)	78,93	78,62	79,45
F1 de Treino (%)	82,15	79,68	81,87
Precisão de Teste (%)	75,66	74,51	76,20
Precisão de Treino (%)	78,84	75,51	78,56
Recall de Teste (%)	82,50	83,20	82,99
Recall de Treino (%)	85,76	84,34	85,47
ROC/AUC de Teste (%)	86,22	85,33	86,58
ROC/AUC de Treino (%)	89,69	86,70	89,14

Esses modelos no treino apresentaram resultados razoáveis com acurácias girando em torno de 77%, com pouco overfitting, quando comparados com as acurácias de treino, que se estabeleceram em torno de 80%. Pelas métricas de F1 pode-se verificar que o modelo apresentou uma média de 78%, indicando excelente trade-off entre as métricas de precisão e recall. Sobre os resultados do ROC AUC verifica-se que o grau de separabilidade dos dados é alta, em torno de 87%, compatível com as demais métricas obtidas.

A lista das importâncias das variáveis podem ser vistas na tabela a seguir. Verifica-se nela que os modelos apresentam uma variável climática (tipicamente alguma medida de temperatura) e uma demográfica (Porte ou o tamanho da população) como sendo as mais importantes para a construção da árvore. As variações das importâncias dos modelos são comuns, pois os algoritmos são diferentes. Este fato ocorre em especial, quando comparamos o XGBoost com os demais.

Tabela 9: Importância das Variáveis (%)

<i>Variável</i>	<i>XGBoost</i>	<i>LightGBM</i>	<i>CatBoost</i>
Coast	2,24	0,76	0,44
Pop	9,58	12,23	12,88
PopArea	1,36	6,30	3,05
Porte	23,31	1,37	5,98
amazonia	6,10	0,29	0,15
caatinga	1,38	0,22	0,15
cerrado	2,57	0,58	0,54
mata_atlantica	1,80	0,87	0,74
pampa	4,00	1,23	2,22
pantanal	0,00	0,00	0,01
favela	1,26	0,83	0,85
IDH_faixa	6,96	3,62	4,27
PrecipitacaoTotal	2,79	10,96	8,24
EvaporacaoPiche	1,79	9,37	6,80
InsolacaoTotal	1,70	8,65	5,46
NebulosidadeMedia	1,56	5,28	4,07
TempCompensadaMedia	8,61	7,78	12,42
TempMinimaMedia	17,64	9,23	14,96
TempMaximaMedia	1,68	5,54	4,56
TempDiff	1,62	6,58	4,72
UmidadeRelativaMedia	2,05	8,32	7,50

Após a realização do treinamento e validação, o próximo passo foi avaliar os resultados das classificações feitas nos dados de Teste. Como são 30% das 105906 observações, temos que o conjunto de testes se resumiu a 31772 observações. De acordo com a matriz de confusão explicitada pelos elementos descritos na tabela abaixo podemos verificar que grande parte dos erros se deram devido aos denominados ***False Positives***.

Tabela 10: Resultados nos dados de Teste

<i>Modelos</i>	<i>True Pos.</i>	<i>True Neg.</i>	<i>False Pos.</i>	<i>False Neg.</i>
XGBoost	10961	13623	4347	2841
LightGBM	10545	13679	4827	2721
CatBoost	11097	13574	4296	2805

A tabela abaixo consegue a partir dos valores da matriz de confusão acima, estabelecer as métricas de avaliação do modelo para os dados de teste.

Tabela 11: Avaliação dos modelos (métricas em %) - Dados de Teste

<i>Modelos</i>	<i>Acurácia</i>	<i>Precisão</i>	<i>Recall</i>	<i>F1</i>
XGBoost	77,38	71,60	79,42	75,31
LightGBM	76,24	68,60	79,49	73,64
CatBoost	77,65	72,09	79,82	75,76

Algumas observações podem ser feitas a partir desta tabela. A primeira reside no fato de que a Precisão é a menor das métricas o que termina diminuindo o valor do F1 e distanciando da acurácia. Por outro lado o poder preditivo ainda continua com uma magnitude aceitável, mas menor que aquele apresentado nos dados de treinamento. O melhor modelo dadas as métricas em todas as tabelas apresentadas é o CatBoost, em segundo lugar o XGBoost e por último o LightGBM.

6.2 Modelagem Espaço-Temporal

Além das 21 variáveis principais mais outras 6, construídas a partir de dados anteriores a 2019, foram então aplicadas para avaliar a existência de casos nos municípios. As 6 variáveis são *dummies* utilizadas para verificar se houve naquele ano infecção no município. Elas são: Infected_2013, Infected_2014 , Infected_2015 , Infected_2016 , Infected_2017 , Infected_2018.

A tabela abaixo, similar ao que foi apresentado no modelo de *Cross-Section* representa os resultados dos modelos, com suas respectivas métricas de avaliação. Vale ressaltar que essas também são métricas do *Cross-Validation* utilizadas para encontrar os melhores hiperparâmetros e outros aprimoramentos.

Tabela 12: Avaliação dos modelos - Cross-Validation

<i>Métricas</i>	<i>XGBoost</i>	<i>LightGBM</i>	<i>CatBoost</i>
Fit Time (s)	4,76	2,10	29,25
Score Time (s)	0,60	0,77	2,62
Acurácia de Teste (%)	84,41	83,65	83,82
Acurácia de Treino (%)	85,41	84,29	84,44
F1 de Teste (%)	89,28	88,74	88,76
F1 de Treino (%)	89,96	89,17	89,17
Precisão de Teste (%)	86,06	84,71	85,45
Precisão de Treino (%)	86,70	85,16	85,91
Recall de Teste (%)	92,78	93,21	92,40
Recall de Treino (%)	93,48	93,58	92,70
ROC/AUC de Teste (%)	90,68	90,12	90,32
ROC/AUC de Treino (%)	91,98	90,77	91,04

Esses modelos no treino apresentaram resultados razoáveis com acurácias girando em torno de 84%. Pelas métricas de F1 pode-se verificar que o modelo apresentou uma média de 89%, indicando excelente trade-off entre as métricas de precisão e recall. Sobre os resultados do ROC AUC verifica-se que o grau de separabilidade dos dados é alta, em torno de 90%, compatível com as demais métricas obtidas.

A lista das importâncias das variáveis podem ser vistas na tabela a seguir. Verifica-se nela que os modelos apontaram como sendo as mais importantes para a construção das árvores, alguma variável climática, uma das 6 variáveis novas e alguma de demografia. As variações das importâncias dos modelos são comuns e possuem explicação similar ao que foi dado na subseção anterior.

Tabela 13: Importância das Variáveis (%)

<i>Variável</i>	<i>XGBoost</i>	<i>LightGBM</i>	<i>CatBoost</i>
Coast	0,20	0,00	0,00
Pop	0,66	16,00	7,11
PopArea	0,38	3,00	1,06
Porte	0,00	0,00	0,00
amazonia	0,00	0,00	0,00
caatinga	0,33	0,00	0,56
cerrado	0,38	0,00	0,37
mata_atlantica	0,00	0,00	0,00
pampa	0,18	0,00	0,00
pantanal	0,00	0,00	0,00
favela	0,00	0,00	0,00
IDH_faixa	0,47	7,00	0,56
PrecipitacaoTotal	0,20	0,00	0,00
EvaporacaoPiche	1,60	4,00	6,69
InsolacaoTotal	1,18	5,00	1,75
NebulosidadeMedia	1,13	9,00	5,79
TempCompensadaMedia	0,36	8,00	15,81
TempMinimaMedia	0,61	6,00	0,45
TempMaximaMedia	14,62	5,00	0,43
TempDiff	0,21	0,00	0,18
UmidadeRelativaMedia	0,81	10,00	4,55
Infected_2013	21,69	3,00	7,95
Infected_2014	7,82	3,00	4,50
Infected_2015	30,29	3,00	15,16
Infected_2016	9,42	4,00	9,81
Infected_2017	2,01	4,00	3,10
Infected_2018	5,49	10,00	14,18

Uma observação decorre do fato de que muitas variáveis ficaram com importâncias insignificantes, tais como a variável *favela* ou *pantanal*, mas a remoção delas do modelo não implicou em melhora nas métricas de avaliação (esperado para modelos de árvores), e suas permanências foram para fins de comparação e controle dos modelos.

Após a realização do treinamento e validação, o próximo passo foi avaliar os resultados das classificações feitas nos dados de Teste. Como são 30%

das 5574 observações, temos que o conjunto de testes se resumiu a 1673 observações. De acordo com a matriz de confusão explicitada pelos elementos descritos na tabela abaixo podemos verificar que grande parte dos erros se deram devido aos denominados ***False Positives***.

Tabela 14: Resultados nos dados de Teste

<i>Modelos</i>	<i>True Pos.</i>	<i>True Neg.</i>	<i>False Pos.</i>	<i>False Neg.</i>
XGBoost	347	1064	183	79
LightGBM	315	1095	182	81
CatBoost	329	1077	169	98

A variável *Infected* para o ano de 2019, não está balanceada, apresentando algo próximo dos 68% de municípios infectados pela Dengue. Isso pode afetar o aprendizado do modelo e fornecer métricas de avaliação que não são muito confiáveis. Com a finalidade de resolver este problema existe uma métrica de avaliação denominada por *Balanced Accuracy* que mede a acurácia segundo a equação abaixo. Mais detalhes sobre esta métrica podem ser verificados em [Brodersen et al., 2010].

$$Bal.Acc. = \frac{\frac{True\ Pos.}{True\ Pos.+False\ Neg.} + \frac{True\ Neg.}{True\ Neg.+False\ Pos.}}{2} \quad (21)$$

Além disso, como a variável é pouco desbalanceada, nenhuma outra técnica mais avançada de reamostragem ou adaptação do processo de *Cross-Validation* se demonstrou necessária ou justificada e apenas a aplicação de uma nova métrica de avaliação foi suficiente. Inclusive na tabela abaixo é possível verificar que os valores dos dois tipos de acurácias pouco se diferem.

Tabela 15: Avaliação dos modelos (métricas em %) - Dados de Teste

<i>Modelos</i>	<i>Acurácia</i>	<i>Bal.Acc.</i>	<i>Precisão</i>	<i>Recall</i>	<i>F1</i>
XGBoost	84,34	83,39	65,47	81,46	72,59
LightGBM	84,28	82,65	63,38	79,55	70,55
CatBoost	84,04	81,74	66,06	77,05	71,14

Algumas observações podem ser feitas a partir desta tabela. A primeira reside no fato de que a Precisão é a menor das métricas o que termina dimi-

nuindo o valor do F1 e distanciando da acurácia. O poder preditivo ainda continua com suas magnitudes similares ao apresentado nos dados de treinamento. O melhor modelo dadas as métricas em todas as tabelas apresentadas é o XGBoost, em segundo lugar o CatBoost e por último o LightGBM.

Figura 21: Previsão de Infecção dos Municípios em 2019 - XGBoost

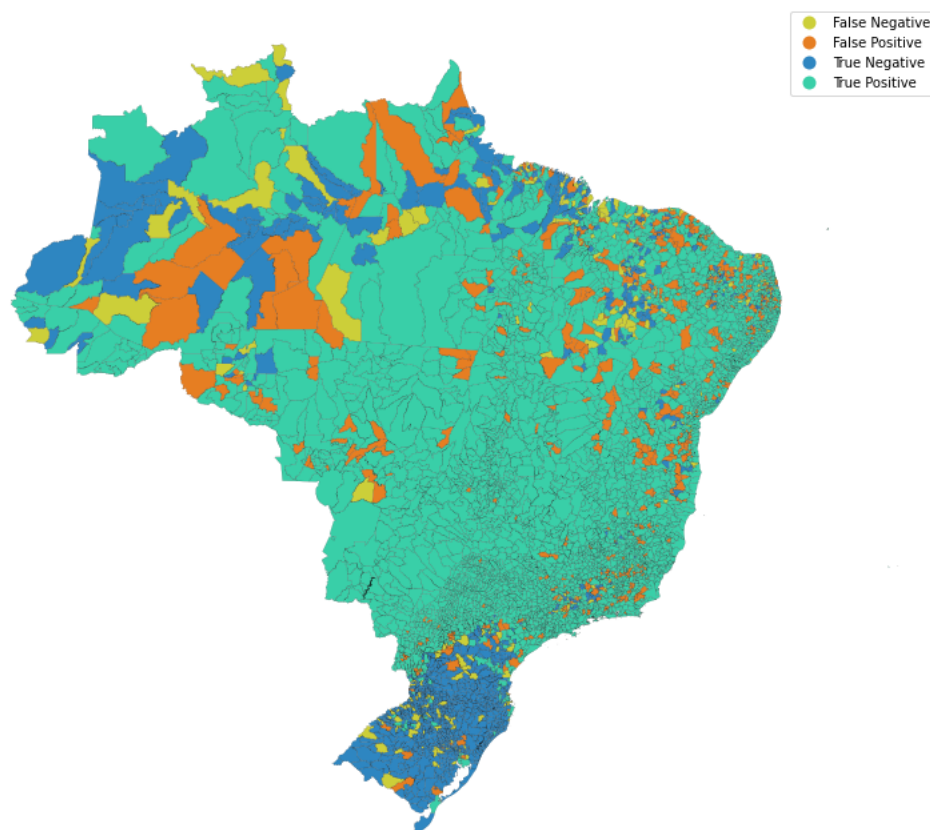


Figura 22: Previsão de Infecção dos Municípios em 2019 - LightGBM

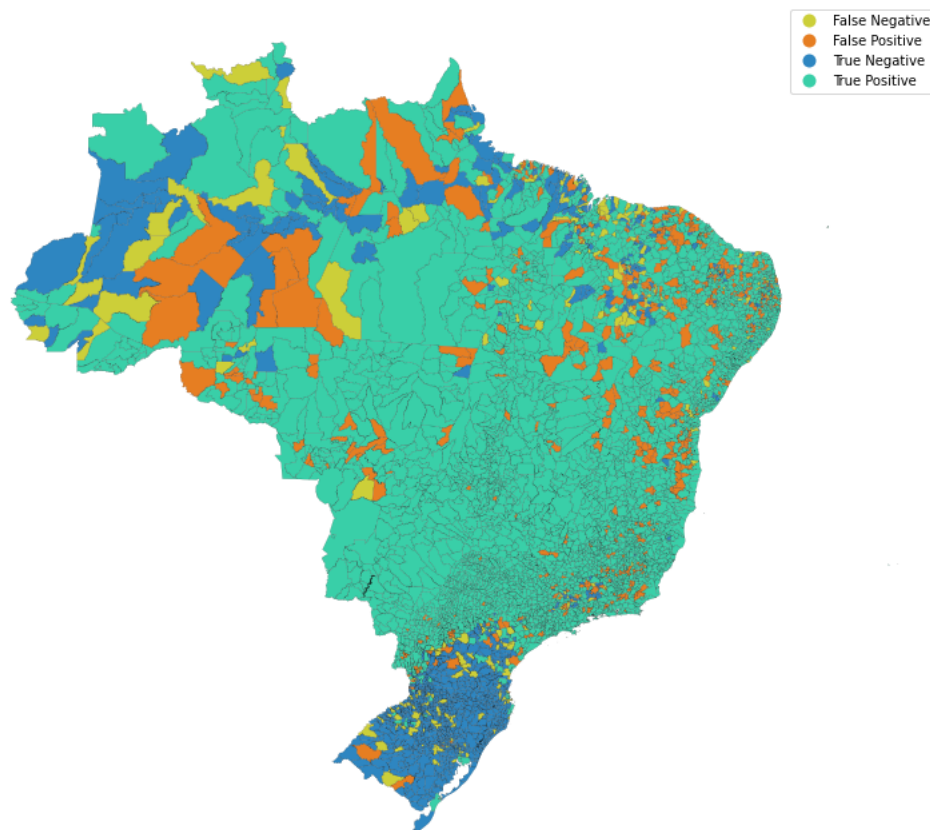
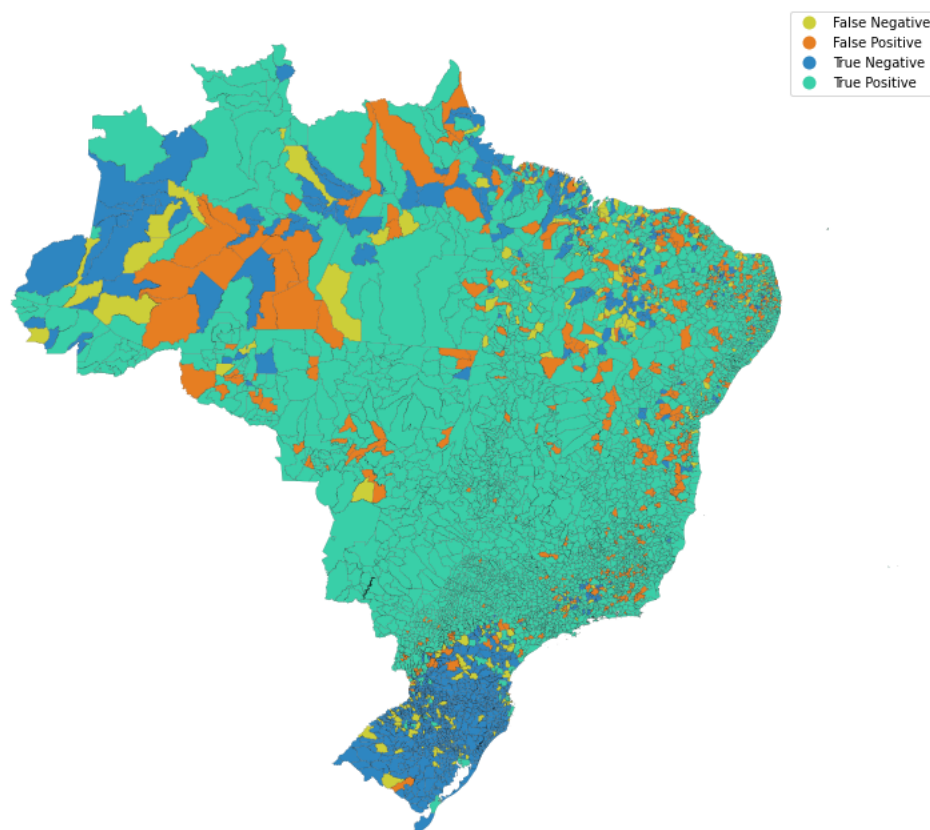


Figura 23: Previsão de Infecção dos Municípios em 2019 - CatBoost



7 Conclusão

Através dos estudos realizados foi possível selecionar variáveis importantes que afetam o ciclo de reprodução dos vetores do vírus da Dengue e tornam favorável o ambiente para maior proliferação deles. Além disso, alguns fatores indiretos relacionados ao escopo demográfico, socioeconômico e geográfico dos municípios também foram levados em consideração, com a finalidade de complementar as demais variáveis.

A maior dificuldade no processo de organização dos dados foi tratar a base de dados de clima, devido à inexistência de valores distribuídos por ano e por município. Este problema, entretanto, foi contornado via utilização de um método de extrapolação baseado em uma técnica conhecida por seu uso para esta finalidade, o KNN.

Através das análises das séries verificou-se o aumento de número de casos após 2007 tanto em quantidade de infecções quanto na proporção de municípios infectados. Este fato que está associado não apenas aos fatores bioclimáticos e demográficos destes municípios, mas também à maneira como os casos foram registrados no DATASUS, demonstram preocupação relevante para a saúde pública. Além disso, em [Escobar et al., 2016] é possível averiguar que em diversos países inicialmente a doença possui um nível de contágio pequeno e de maneira gradual ela eleva os números de infectados.

Ainda em [Escobar et al., 2016], constata-se os efeitos negativos para os estudos e previsão dos modelos decorrentes da ausência de registros constantes e formais das ocorrências de infecção, o que é inclusive agravado pela dificuldade de se estabelecer um diagnóstico, segundo [Bordignon et al., 2008]. A metodologia desses registros idealmente deveriam não se modificar, mas na prática tais situações ocorrem devido a capacidade de organização e resposta do sistema de saúde.

A construção de dois tipos de modelos pôde fornecer uma maneira ampla de visualizar o tema. Um focado em entender os aspectos espaciais dos municípios e outro mais focado em como prever futuramente os municípios que apresentam uma maior capacidade de risco de infecção pela Dengue. Em ambos foi possível estabelecer uma previsão com acurácia superior a 75%. Esse resultado se assemelha com modelos encontrados em [Laureano-Rosario et al., 2018] e [Buczak et al., 2018].

Algumas modelagens, como por exemplo [Messina et al., 2019], consideram efeitos dos fluxos migratórios sobre a disseminação da doença, uma vez que fatores relacionados a imunização de indivíduos presentes no município

tendem a alterar os números de infecção. Neste trabalho, devido a dificuldade de obtenção de dados migratórios entre os municípios não foi possível incluir tais efeitos na análise, de forma que a melhor medida desses fluxos ficou a cargo das variáveis demográficas e socioeconômicas.

Além disso, as variáveis com maior importância para a construção das árvores de decisão foram aquelas citadas pela literatura da área, em especial, aquelas relacionadas a demografia e a temperaturas. Isso indica que variáveis socioeconômicas e de biomas não foram tão eficientes em alterar a classificação do município como infectado ou não.

Em termos de previsibilidade do modelo, o maior problema foi resultante dos Falsos Positivos, caso em que se prevê a infecção do município naquele ano e tal fato não ocorre. Isso se verifica em maior quantidade para o nordeste e norte, talvez pela diferenças intrínsecas dessas regiões.

Algumas limitações do modelo são associadas a alguns fatores. Primeiro na forma de caracterização do município infectado, considerando apenas se houve ao menos um caso de infecção ou não. Isso pode aumentar as dimensões das proporções de municípios infectados. Talvez uma pequena mudança em estudos futuros, como a caracterização de surtos com um número percentual mínimo de infecção seja mais interessante.

Por outro lado, existe uma certa dificuldade em padronizar os municípios que variam em todas suas características (área, população, biomas entre outras). Isso faz com que as relações e os caminhos de decisão do modelo sejam afetados e percam um pouco de sua previsibilidade. A própria variável relacionada aos aglomerados subnormais termina favorecendo ocorrências no sudeste, uma vez que aquelas presentes no norte estão associadas mais a um aglomerado horizontal do que vertical, por exemplo. Entretanto, tentativas de clusterizar o estado de forma não tradicional como o uso de regiões geográficas ou segundo as classificações de biomas segundo Köppen-Geiger, não se demonstraram superiores em estudos prévios, e desta maneira os biomas presentes no IBGE foram utilizados.

Observando os resultados dos modelos, é visível a consistência de municípios do sudeste e do litoral nordeste em termos de presença de infecção. O centro-oeste terminou após 2007 apresentando tal constância e assim como é possível verificar nas figuras 21, 22 e 23, os modelos não falharam em apontar tal fato.

Algumas modificações e extensões futuras podem ser realizadas para o aprimoramento dos modelos. Uma delas poderia ser a inserção de outras variáveis, tais como aquelas relacionadas ao saneamento básico, ou a aplicação

de técnicas de classificação realizadas por Redes Neurais. As possibilidades dependem também da disponibilidade dos dados para a janela temporal selecionada.

A importância deste trabalho reside em possibilitar maior capacidade de gerenciamento de políticas públicas e auxiliar o combate aos vetores transmissores da Dengue. E com os modelos apresentados esse objetivo principal foi em certa medida alcançado. Os modelos conseguem apontar variáveis relevantes e através de suas simples manipulações informar quais municípios estão em situação de maior risco, favorecendo e antecipando ações preventivas ou de resposta imediata para o enfrentamento da doença.

Referências

- [Abu-Mostafa et al., 2012] Abu-Mostafa, Y. S., Magdon-Ismael, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook New York, NY, USA:.
- [Aguirre-Obando et al., 2017] Aguirre-Obando, O. A., Martins, A. J., and Navarro-Silva, M. A. (2017). First report of the phe1534cys kdr mutation in natural populations of aedes albopictus from brazil. *Parasites & vectors*, 10(1):1–10.
- [Alaniz et al., 2019] Alaniz, A. J., Carvajal, M. A., Bacigalupo, A., and Cattán, P. E. (2019). Global spatial assessment of aedes aegypti and culex quinquefasciatus: a scenario of zika virus exposure. *Epidemiology & Infection*, 147.
- [Anyamba et al., 2019] Anyamba, A., Chretien, J.-P., Britch, S. C., Soebiyanto, R. P., Small, J. L., Jepsen, R., Forshey, B. M., Sanchez, J. L., Smith, R. D., Harris, R., et al. (2019). Global disease outbreaks associated with the 2015–2016 el niño event. *Scientific reports*, 9(1):1–14.
- [Aswi et al., 2019] Aswi, A., Cramb, S., Moraga, P., and Mengersen, K. (2019). Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. *Epidemiology & Infection*, 147.
- [Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.
- [Bordignon et al., 2008] Bordignon, J., Probst, C. M., Mosimann, A. L. P., Pavoni, D. P., Stella, V., Buck, G. A., Satproedprai, N., Fawcett, P., Zanata, S. M., de Noronha, L., et al. (2008). Expression profile of interferon stimulated genes in central nervous system of mice infected with dengue virus type-1. *Virology*, 377(2):319–329.
- [Brady et al., 2013] Brady, O. J., Johansson, M. A., Guerra, C. A., Bhatt, S., Golding, N., Pigott, D. M., Delatte, H., Grech, M. G., Leisnham, P. T., Maciel-de Freitas, R., et al. (2013). Modelling adult aedes aegypti and aedes albopictus survival at different temperatures in laboratory and field settings. *Parasites & vectors*, 6(1):1–12.

- [Brady et al., 2015] Brady, O. J., Smith, D. L., Scott, T. W., and Hay, S. I. (2015). Dengue disease outbreak definitions are implicitly variable. *Epidemics*, 11:92–102.
- [Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.
- [Brodersen et al., 2010] Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE.
- [Buczak et al., 2018] Buczak, A. L., Baugher, B., Moniz, L. J., Bagley, T., Babin, S. M., and Guven, E. (2018). Ensemble method for dengue prediction. *PloS one*, 13(1):e0189988.
- [Carlson et al., 2016] Carlson, C. J., Dougherty, E. R., and Getz, W. (2016). An ecological assessment of the pandemic threat of zika virus. *PLoS neglected tropical diseases*, 10(8):e0004968.
- [Chen and Guestrin, 2016a] Chen, T. and Guestrin, C. (2016a). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.
- [Chen and Guestrin, 2016b] Chen, T. and Guestrin, C. (2016b). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA. ACM.
- [Cunze et al., 2019] Cunze, S., Kochmann, J., Koch, L. K., Genthner, E., and Klimpel, S. (2019). Vector distribution and transmission risk of the zika virus in south and central america. *PeerJ*, 7:e7920.
- [Custódio et al., 2019] Custódio, J. M. d. O., Nogueira, L. M. S., Souza, D. A., Fernandes, M. F., Oshiro, E. T., Oliveira, E. F. d., Piranda, E. M., and Oliveira, A. G. d. (2019). Abiotic factors and population dynamic of aedes aegypti and aedes albopictus in an endemic area of dengue in brazil. *Revista do Instituto de Medicina Tropical de São Paulo*, 61.

- [da Cruz Ferreira et al., 2017] da Cruz Ferreira, D. A., Degener, C. M., de Almeida Marques-Toledo, C., Bendati, M. M., Fetzner, L. O., Teixeira, C. P., and Eiras, Á. E. (2017). Meteorological variables and mosquito monitoring are good predictors for infestation trends of *aedes aegypti*, the vector of dengue, chikungunya and zika. *Parasites & vectors*, 10(1):78.
- [Dickens et al., 2018] Dickens, B. L., Sun, H., Jit, M., Cook, A. R., and Carrasco, L. R. (2018). Determining environmental and anthropogenic factors which explain the global distribution of *aedes aegypti* and *ae. albopictus*. *BMJ global health*, 3(4).
- [Escobar et al., 2016] Escobar, L. E., Qiao, H., and Peterson, A. T. (2016). Forecasting chikungunya spread in the americas via data-driven empirical approaches. *Parasites & vectors*, 9(1):1–12.
- [Faria et al., 2017] Faria, N. R., Quick, J., Claro, I., Theze, J., de Jesus, J. G., Giovanetti, M., Kraemer, M. U., Hill, S. C., Black, A., da Costa, A. C., et al. (2017). Establishment and cryptic transmission of zika virus in brazil and the americas. *Nature*, 546(7658):406–410.
- [Friedman et al., 2001] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.
- [Friedman, 2002] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.
- [Garcia et al., 2018] Garcia, G. d. A., David, M. R., Martins, A. d. J., Maciel-de Freitas, R., Linss, J. G. B., Araújo, S. C., Lima, J. B. P., and Valle, D. (2018). The impact of insecticide applications on the dynamics of resistance: The case of four *aedes aegypti* populations from different brazilian regions. *PLoS neglected tropical diseases*, 12(2):e0006227.
- [Gloria-Soria et al., 2016] Gloria-Soria, A., Ayala, D., Bheecarry, A., Calderon-Arguedas, O., Chadee, D. D., Chiappero, M., Coetzee, M., Elahee, K. B., Fernandez-Salas, I., Kamal, H. A., et al. (2016). Global genetic diversity of *aedes aegypti*. *Molecular ecology*, 25(21):5377–5395.
- [Grajski et al., 1986] Grajski, K. A., Breiman, L., Di Prisco, G. V., and Freeman, W. J. (1986). Classification of eeg spatial patterns with a tree-

- structured methodology: Cart. *IEEE transactions on biomedical engineering*, (12):1076–1086.
- [Guo et al., 2017] Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., Luo, G., Li, Z., He, J., Zhang, Y., et al. (2017). Developing a dengue forecast model using machine learning: A case study in china. *PLoS neglected tropical diseases*, 11(10):e0005973.
- [Haddawy et al., 2019] Haddawy, P., Wettayakorn, P., Nonthaleerak, B., Yin, M. S., Wiratsudakul, A., Schöning, J., Laosiritaworn, Y., Balla, K., Euaungkanakul, S., Quengdaeng, P., et al. (2019). Large scale detailed mapping of dengue vector breeding sites using street view images. *PLoS neglected tropical diseases*, 13(7):e0007555.
- [Harris et al., 2020] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585:357–362.
- [Hunter, 2007] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *IEEE Annals of the History of Computing*, 9(03):90–95.
- [Jain et al., 2019] Jain, R., Sontisirikit, S., Iamsirithaworn, S., and Prendinger, H. (2019). Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. *BMC infectious diseases*, 19(1):1–16.
- [Jordahl et al., 2020] Jordahl, K., den Bossche, J. V., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., Wolf, L. J., Journois, M., Wilson, J., Greenhall, A., Holdgraf, C., Filipe, and Leblanc, F. (2020). geopandas/geopandas: v0.8.1.
- [Kamal et al., 2018] Kamal, M., Kenawy, M. A., Rady, M. H., Khaled, A. S., and Samy, A. M. (2018). Mapping the global potential distributions of two

- arboviral vectors aedes aegypti and ae. albopictus under changing climate. *PloS one*, 13(12):e0210122.
- [Kanamori et al., 2007] Kanamori, T., Takenouchi, T., Eguchi, S., and Murata, N. (2007). Robust loss functions for boosting. *Neural computation*, 19(8):2183–2244.
- [Ke et al., 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154.
- [Kikuti et al., 2015] Kikuti, M., Cunha, G. M., Paploski, I. A., Kasper, A. M., Silva, M. M., Tavares, A. S., Cruz, J. S., Queiroz, T. L., Rodrigues, M. S., Santana, P. M., et al. (2015). Spatial distribution of dengue in a brazilian urban slum setting: Role of socioeconomic gradient in disease risk. *PLoS Negl Trop Dis*, 9(7):e0003937.
- [Kotsakiozi et al., 2017] Kotsakiozi, P., Gloria-Soria, A., Caccone, A., Evans, B., Schama, R., Martins, A. J., and Powell, J. R. (2017). Tracking the return of aedes aegypti to brazil, the major vector of the dengue, chikungunya and zika viruses. *PLoS neglected tropical diseases*, 11(7):e0005653.
- [Kraemer et al., 2019] Kraemer, M. U., Reiner, R. C., Brady, O. J., Messina, J. P., Gilbert, M., Pigott, D. M., Yi, D., Johnson, K., Earl, L., Marczak, L. B., et al. (2019). Past and future spread of the arbovirus vectors aedes aegypti and aedes albopictus. *Nature microbiology*, 4(5):854–863.
- [Kraemer et al., 2015] Kraemer, M. U., Sinka, M. E., Duda, K. A., Mylne, A., Shearer, F. M., Brady, O. J., Messina, J. P., Barker, C. M., Moore, C. G., Carvalho, R. G., et al. (2015). The global compendium of aedes aegypti and ae. albopictus occurrence. *Scientific data*, 2(1):1–8.
- [Laureano-Rosario et al., 2018] Laureano-Rosario, A. E., Duncan, A. P., Mendez-Lazaro, P. A., Garcia-Rejon, J. E., Gomez-Carro, S., Farfan-Ale, J., Savic, D. A., and Muller-Karger, F. E. (2018). Application of artificial neural networks for dengue fever outbreak predictions in the northwest coast of yucatan, mexico and san juan, puerto rico. *Tropical medicine and infectious disease*, 3(1):5.

- [Leta et al., 2018] Leta, S., Beyene, T. J., De Clercq, E. M., Amenu, K., Kraemer, M. U., and Revie, C. W. (2018). Global risk mapping for major diseases transmitted by *aedes aegypti* and *aedes albopictus*. *International Journal of Infectious Diseases*, 67:25–35.
- [Lorenz et al., 2017] Lorenz, C., Azevedo, T. S., Virginio, F., Aguiar, B. S., Chiaravalloti-Neto, F., and Suesdek, L. (2017). Impact of environmental factors on neglected emerging arboviral diseases. *PLoS neglected tropical diseases*, 11(9):e0005959.
- [Lourengo et al., 2017] Lourenço, J., de Lima, M. M., Faria, N. R., Walker, A., Kraemer, M. U., Villabona-Arenas, C. J., Lambert, B., de Cerqueira, E. M., Pybus, O. G., Alcantara, L. C., et al. (2017). Epidemiological and ecological determinants of zika virus transmission in an urban setting. *Elife*, 6:e29820.
- [McKinney et al., 2010] McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.
- [Messina et al., 2019] Messina, J. P., Brady, O. J., Golding, N., Kraemer, M. U., Wint, G. W., Ray, S. E., Pigott, D. M., Shearer, F. M., Johnson, K., Earl, L., et al. (2019). The current and future global distribution and population at risk of dengue. *Nature microbiology*, 4(9):1508–1515.
- [Messina et al., 2016] Messina, J. P., Kraemer, M. U., Brady, O. J., Pigott, D. M., Shearer, F. M., Weiss, D. J., Golding, N., Ruktanonchai, C. W., Gething, P. W., Cohn, E., et al. (2016). Mapping global environmental suitability for zika virus. *Elife*, 5:e15272.
- [Metsky et al., 2017] Metsky, H. C., Matranga, C. B., Wohl, S., Schaffner, S. F., Freije, C. A., Winnicki, S. M., West, K., Qu, J., Baniecki, M. L., Gladden-Young, A., et al. (2017). Zika virus evolution and spread in the americas. *Nature*, 546(7658):411–415.
- [Mussumeci and Coelho, 2020] Mussumeci, E. and Coelho, F. C. (2020). Large-scale multivariate forecasting models for dengue-lstm versus random forest regression. *Spatial and Spatio-temporal Epidemiology*, page 100372.

- [Panda et al., 2009] Panda, B., Herbach, J. S., Basu, S., and Bayardo, R. J. (2009). Planet: massively parallel learning of tree ensembles with mapreduce.
- [Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.
- [Posen et al., 2016] Posen, H. J., Keystone, J. S., Gubbay, J. B., and Morris, S. K. (2016). Epidemiology of zika virus, 1947–2007. *BMJ global health*, 1(2).
- [Prokhorenkova et al., 2018] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In *Advances in neural information processing systems*, pages 6638–6648.
- [Romero et al., 2019] Romero, D., Olivero, J., Real, R., and Guerrero, J. C. (2019). Applying fuzzy logic to assess the biogeographical risk of dengue in south america. *Parasites & vectors*, 12(1):428.
- [Ross, 2010] Ross, T. M. (2010). Dengue virus. *Clinics in laboratory medicine*, 30(1):149–160.
- [Samy et al., 2016] Samy, A. M., Thomas, S. M., Wahed, A. A. E., Cohoon, K. P., and Peterson, A. T. (2016). Mapping the global geographic potential of zika virus spread. *Memorias do Instituto Oswaldo Cruz*, 111(9):559–560.
- [Schapire, 1999] Schapire, R. E. (1999). A brief introduction to boosting. In *Ijcai*, volume 99, pages 1401–1406.
- [Tjaden et al., 2017] Tjaden, N. B., Suk, J. E., Fischer, D., Thomas, S. M., Beierkuhnlein, C., and Semenza, J. C. (2017). Modelling the effects of global climate change on chikungunya transmission in the 21 st century. *Scientific reports*, 7(1):1–11.
- [Valle et al., 2019] Valle, D., Bellinato, D. F., Viana-Medeiros, P. F., Lima, J. B. P., and Martins Junior, A. d. J. (2019). Resistance to temephos

and deltamethrin in aedes aegypti from brazil between 1985 and 2017. *Memórias do Instituto Oswaldo Cruz*, 114.

[Waskom and the seaborn development team, 2020] Waskom, M. and the seaborn development team (2020). `mwaskom/seaborn`.

[Watts et al., 2017] Watts, A. G., Miniota, J., Joseph, H. A., Brady, O. J., Kraemer, M. U., Grills, A. W., Morrison, S., Esposito, D. H., Nicolucci, A., German, M., et al. (2017). Elevation as a proxy for mosquito-borne zika virus transmission in the americas. *PloS one*, 12(5):e0178211.

[Wiratsudakul et al., 2018] Wiratsudakul, A., Suparit, P., and Modchang, C. (2018). Dynamics of zika virus outbreaks: an overview of mathematical modeling approaches. *PeerJ*, 6:e4526.