Estimativa de impacto da dengue usando dados climatológicos

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

06/09/2020

MBA Analytics em Big Data

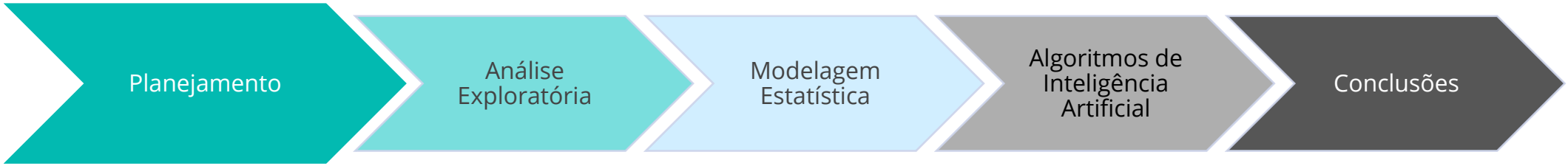
Nome do Aluno:
Pedro Henrique Quadros Alves

Coordenadores:
Prof.ª Dr.ª Alessandra de Álvila Montini
Prof. Dr. Adolpho Walter Pimazoni Canton

Agenda

1. Objetivo do Trabalho
2. Contextualização do Problema
3. Base de Dados
    i. Bases originais e principais variáveis
    ii. Filtros
4. Análise Exploratória de Dados
5. Modelagem com Estatística Tradicional
6. Inteligência Artificial e Otimização
7. Conclusões

Metodologia de análise de dados



Definição do problema

- Objetivos
- Contextualização do problema
- Base de Dados

Variáveis Qualitativas

- Tabela de Frequência

Variáveis Quantitativas

- Medidas de Posição
- Medidas de Dispersão
- Boxplot

Estatística Tradicional

- Regressão Linear
- Árvore de Decisão

Inteligência Artificial e Otimização

- Random Forest
- XGBoost
- LGBM
- Cross Validation

Metodologia de análise de dados

Planejamento → Análise Exploratória → Modelagem Estatística → Algoritmos de Inteligência Artificial → Conclusões

Definição do problema

- Objetivos
- Contextualização do problema
- Base de Dados

Variáveis Qualitativas

- Tabela de Frequência

Variáveis Quantitativas

- Medidas de Posição
- Medidas de Dispersão
- Boxplot

Estatística Tradicional

- Regressão Linear
- Árvore de Decisão

Inteligência Artificial e Otimização

- Random Forest
- XGBoost
- LGBM
- Cross Validation

1. Objetivo do Trabalho

Todos os anos diversas cidades brasileiras sofrem impacto significativo com casos de dengue, causando aumento de atendimento em hospitais, impacto social e em casos mais graves, óbitos.

O objetivo do trabalho é proporcionar através de análise de dados e inteligência artificial, uma maneira para ajudar prefeituras a preverem a dimensão dos casos de dengue em suas cidades.

O problema a ser resolvido é auxiliar as prefeituras e demais órgãos governamentais a direcionarem melhor a verba de prevenção e campanhas sociais e também a verba para recursos relacionados à saúde. Tudo isso para deixar as regiões mais propensas ao impacto da doença melhor preparadas.

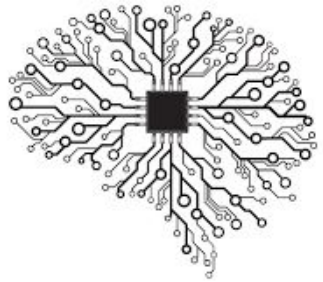
Para isso serão usados dados de estações climatológicas do INMET (Instituto Nacional de Meteorologia), e registros de notificação de dengue do SINAN (Sistema de Informação de Agravos de Notificação).

1. Objetivo do Trabalho

O fluxograma abaixo demonstra o objetivo do trabalho em forma de aplicação

Coleta de Dados de Estações Meteorológicas através de API disponibilizada pelo INMET

Processamento dos Dados em cloud para organização das bases, mantendo atualização serverless



Uso de Inteligência Artificial para previsão de Notificações

Disponibilização dos resultados de forma a facilitar análises e insights para órgãos governamentais e hospitais

Obs: Foi realizado um trabalho semelhante pela FIOCRUZ em parceria com a FGV. O trabalho pode ser consultado em https://info.dengue.mat.br
Apesar do objetivo similar, nenhuma base ou modelo foi consultado do trabalho em questão.

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

2. Contextualização do Problema

Sobre a dengue, sabe-se que a doença é transmitida através da picada do mosquito **Aedes *aegypti*** que se reproduz com maior facilidade em ambientes de alta umidade e em época de chuvas e calor.

Como a reprodução do mosquito acontece muito rapidamente (entre 7 a 10 dias), é necessário cuidado constante para evitar proliferação do mesmo e consequentemente aumento da probabilidade de dispersão da dengue.

Para evitar óbitos e dificuldades no atendimento hospitalar, será necessário desenvolver um modelo preditivo para entender qual o impacto da dengue em diversas cidades brasileiras de forma a possibilitar que as prefeituras façam uma correta distribuição de recursos e melhor planejamento de capacidade para atendimento aos habitantes, além de direcionar esforços preventivos nas regiões mais propensas a impacto.

Como a reprodução do mosquito transmissor está fortemente associada à mudanças do clima, a base usada para explicar o número de casos será uma base de dados gerada por estações climatológicas.

Estas estações medem variáveis como temperatura, precipitação, umidade relativa do ar, etc. O INMET disponibiliza os dados em seu site, e através de uma API.

Os casos de dengue foram retirados do SINAN, neste link[1].

Como apoio foi usada uma base de municípios do IBGE, disponível neste link.

[1] Os dados de dengue são disponibilizados no formato DBC. Antes do trabalho com as bases, foi usada a biblioteca readdbc do R para ler os dados, os dados foram posteriormente exportados em CSV para uso no Python.

3.i Bases de Dados Originais

Volume: 30.739.080 registros (dados a cada 1 hora entre 2014 e 2020)

Fonte: INMET (Instituto Nacional de Meteorologia)

Principais Variáveis

- PRECIPITAÇÃO TOTAL, HORÁRIO (mm)
- PRESSÃO ATMOSFÉRICA AO NÍVEL DA ESTAÇÃO, HORÁRIA (mB)
- PRESSÃO ATMOSFÉRICA MAX.NA HORA ANT. (AUT) (mB)
- PRESSÃO ATMOSFÉRICA MIN. NA HORA ANT. (AUT) (mB)
- RADIAÇÃO GLOBAL (W/m²)
- TEMPERATURA DO AR - BULBO SECO, HORÁRIA (°C) [1]
- TEMPERATURA DO PONTO DE ORVALHO (°C) [2]
- TEMPERATURA MÁXIMA NA HORA ANT. (AUT) (°C)
- TEMPERATURA MÍNIMA NA HORA ANT. (AUT) (°C)
- TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (°C)
- TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (°C)
- UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)
- UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)
- UMIDADE RELATIVA DO AR, HORÁRIA (%)
- VENTO, DIREÇÃO HORÁRIA (gr) (° (gr))
- VENTO, RAJADA MÁXIMA (m/s)
- VENTO, VELOCIDADE HORÁRIA (m/s)

Conceitos

[1] A temperatura do bulbo seco (t) do ar é a temperatura medida com um termômetro comum

[2] Ponto de orvalho é a temperatura até a qual o ar deve ser resfriado para que o vapor de água presente condense na forma de orvalho ou geada.

3.i Bases de Dados Originais

Fonte: SINAN (Sistema de Informação de Agravos de Notificação)

Volume : 8.907.774 registros (dados de todas as notificações no Brasil entre 2014 e 2019)

Principais Variáveis

- NU_NOTIFIC - Número da Notificação
- DT_NOTIFIC - Data da Notificação
- NU_ANO - Ano da Notificação
- ID_MUNICIP - Código IBGE do Município onde ocorreu a notificação

Obs: Devido ao número muito grande de variáveis e falta de clareza na documentação fornecida pelo SINAN, optou-se por abordar no trabalho apenas as principais variáveis (que foram usadas para atingir o objetivo)

3.i Bases de Dados Originais

Fonte: IBGE

Volume : 10.496 registros (Tabela com códigos IBGE desde o nível estado até o nível distrito)

Principais Variáveis

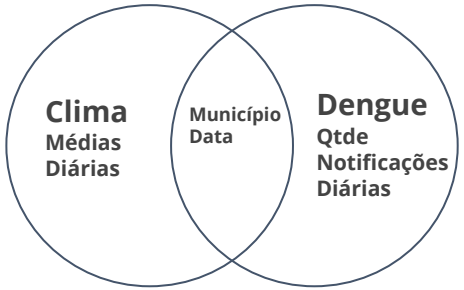
- UF
- Nome_UF
- Mesorregião Geográfica
- Nome_Mesorregião
- Microrregião Geográfica
- Nome_Microrregião
- Município
- Código Município Completo
- Nome_Município
- Distrito
- Código de Distrito Completo
- Nome_Distrito

3.ii Filtros Realizados nas Bases

Processo Executado até Chegar na Tabela ABT

O objetivo final foi atingir uma base de dados onde tivéssemos dados diários de clima agregados pela média como variáveis explicativas e a quantidade de notificações de casos de dengue como variáveis resposta.

Não foram realizados filtros básicos nas bases originais, porém o processo de agregação e cruzamento de bases já acarretou em uma redução brusca no volume de dados.



Volume final de registros: 157.268

Passo 1

Contagem das notificações de Dengue por Município

A variável resposta são as notificações de dengue. Não há necessidade de mantermos os dados analíticos para isso, foi realizada a agregação da base por município e data.

Passo 2

Relacionamento dos casos de dengue com municípios

Cruzamento da base de dengue com a base do IBGE de municípios. O objetivo é trazer para a base de dengue o nome do município.
Apenas cruzamentos bem sucedidos foram trazidos (inner join)

Passo 3

Agregação da Base de Clima

Selecionados apenas registros onde foi possível descobrir o município através de dados de latitude e longitude.
Agregação de dados por dia e por município (média diária)

Passo 4

Cruzamento de Dados de Clima e Dengue

Selecionados apenas registros onde foi possível cruzar dados de clima e dengue (inner join). As chaves usadas para o cruzamento foram o nome do município e a data.

3.ii Filtros Realizados nas Bases

Validação de Nome de Município

Como toda a estruturação das bases se dava em torno de um cruzamento de dados usando o nome do município que foi descoberto usando latitude e longitude, por segurança, foi feita uma validação da cidade.

A ideia é garantir que a cidade extraída da latitude e longitude realmente está correta.

Município e CEP foram gerados através da análise de latitude e longitude (biblioteca geopy)

Validação entre CEP e Município usando outra biblioteca (PyCEP_Correios).

Base Filtrada apenas com municípios validados (CEP e Município são coerentes entre si)

Metodologia de análise de dados

Planejamento | **Análise Exploratória** | Modelagem Estatística | Algoritmos de Inteligência Artificial | Conclusões

Definição do problema

- Objetivos
- Contextualização do problema
- Base de Dados

Variáveis Qualitativas

- Tabela de Frequência

Variáveis Quantitativas

- Medidas de Posição
- Medidas de Dispersão
- Boxplot

Estatística Tradicional

- Regressão Linear
- Árvore de Decisão

Inteligência Artificial e Otimização

- Random Forest
- XGBoost
- LGBM
- Cross Validation

4. Análise Exploratória de Dados

Devido ao grande volume de dados e principalmente devido ao fato de uma grande parte não ser relevante para o objetivo do trabalho, foi feita a opção de realizar a análise exploratória da base já agregada, e não das bases de dados individuais.

Variáveis Qualitativas:

Estado, Cidade, CEP, ID_MUNICIP

Tabela de Frequência

Coluna	Freq
estado	25
cidade	106
cep	129
ID_MUNICIP	106

Obs: Variáveis repetidas removidas da tabela de frequência

Analisando as variáveis de frequência é possível perceber que:

- Temos quase todos os estados representados na base de dados
- Temos mais CEPs que cidades. Como o CEP foi pego da latitude e longitude, é possível perceber que temos mais de uma estação por cidade.
- No caso do ID_MUNICIP a relação esperada é de 1:1 já que cada cidade possui um ID único no IBGE

4. Análise Exploratória de Dados

Variáveis Quantitativas:

- Lat
- Long
- PRECIPITAÇÃO TOTAL, HORÁRIO (mm)
- PRESSAO ATMOSFERICA AO NIVEL DA ESTAÇÃO, HORÁRIA (mB)
- PRESSÃO ATMOSFÉRICA MAX.NA HORA ANT. (AUT) (mB)
- PRESSÃO ATMOSFÉRICA MIN. NA HORA ANT. (AUT) (mB)
- RADIAÇÃO GLOBAL (W/m²)
- TEMPERATURA DO AR - BULBO SECO, HORÁRIA (°C)
- TEMPERATURA DO PONTO DE ORVALHO (°C)
- TEMPERATURA MÁXIMA NA HORA ANT. (AUT) (°C)
- TEMPERATURA MÍNIMA NA HORA ANT. (AUT) (°C)
- TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (°C)
- TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (°C)
- UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)
- UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)
- UMIDADE RELATIVA DO AR, HORÁRIA (%)
- VENTO, DIREÇÃO HORÁRIA (gr) (° (gr))
- VENTO, RAJADA MÁXIMA (m/s)
- VENTO, VELOCIDADE HORÁRIA (m/s)
- NU_NOTIFIC

4. Análise Exploratória de Dados

Medidas de Posição e Dispersão:

	lat	long	PRECIPITAÇÃO TOTAL, HORÁRIO (mm)	PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB)	PRESSÃO ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB)	PRESSÃO ATMOSFERICA MIN. NA HORA ANT. (AUT) (mB)	RADIACAO GLOBAL (W/m²)	TEMPERATURA DO AR - BULBO SECO, HORARIA (°C)	TEMPERATURA DO PONTO DE ORVALHO (°C)	TEMPERATURA MÁXIMA NA HORA ANT. (AUT) (°C)
count	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00
mean	-16,49	-46,15	0,16	961,06	961,22	960,72	1.271,62	23,85	17,81	24,48
std	7,53	6,45	0,41	75,20	75,31	75,20	228,25	3,29	3,88	3,31
min	-32,08	-68,17	0,00	0,40	0,20	0,17	0,00	5,25	0,09	5,66
25%	-22,36	-49,22	0,00	931,77	932,00	931,50	1.271,62	21,99	15,93	22,62
50%	-19,00	-46,62	0,00	962,01	962,21	961,66	1.271,62	23,88	17,93	24,53
75%	-10,18	-41,98	0,16	1.006,76	1.007,01	1.006,48	1.271,62	26,25	20,68	26,88
max	2,82	-34,82	10,28	1.031,47	1.031,72	1.031,32	4.640,17	37,40	30,70	37,50

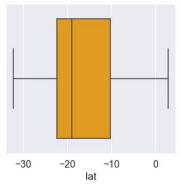
	TEMPERATURA MÍNIMA NA HORA ANT. (AUT) (°C)	TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (°C)	TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (°C)	UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)	UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)	UMIDADE RELATIVA DO AR, HORARIA (%)	VENTO, DIREÇÃO HORARIA (gr) (° (gr))	VENTO, RAJADA MAXIMA (m/s)	VENTO, VELOCIDADE HORARIA (m/s)	NU_NOTIFIC
count	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00	121.117,00
mean	23,25	18,34	17,31	74,49	68,94	71,77	150,13	4,92	1,93	26,62
std	3,27	3,85	3,91	11,65	12,11	11,90	52,61	1,79	1,07	119,76
min	4,73	0,03	0,04	11,90	10,00	10,00	6,00	0,00	0,00	1,00
25%	21,40	16,47	15,40	69,54	63,29	66,46	113,92	3,77	1,24	2,00
50%	23,26	18,47	17,41	75,04	69,04	72,08	150,13	4,92	1,86	5,00
75%	25,65	21,19	20,20	82,46	77,08	79,83	179,92	5,83	2,38	16,00
max	35,60	33,10	30,10	100,00	100,00	100,00	360,00	19,44	11,63	4.017,00

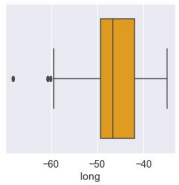
Olhando para as medidas de posição e dispersão, os principais pontos notáveis são a variação da temperatura de bulbo seco que é baixa apesar de termos grande diferença entre regiões e estações (temperatura média acima de 20°C com desvio padrão de apenas 3,29°C. Nota-se pelos quartis a concentração nos valores 21 e 26°C
Além disso é possível perceber também uma umidade relativa bem concentrada em valores entre 66 e 79%.

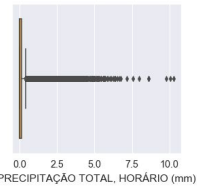
Já para o número de notificações de dengue, é curioso notar um alto desvio padrão apesar de uma média baixa de casos. A maior parte dos casos se concentra em valores baixos porém o valor máximo demonstra que existem situações muito atípicas na base. Isso demonstra realidades muito distintas entre as regiões contidas na base.

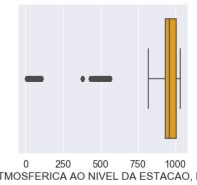
4. Análise Exploratória de Dados

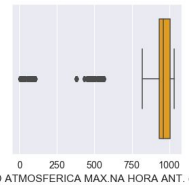
Boxplot

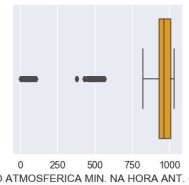


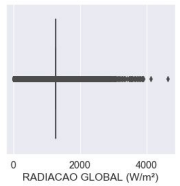


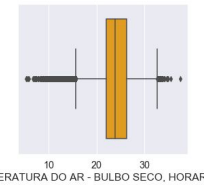


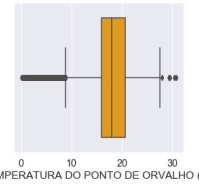


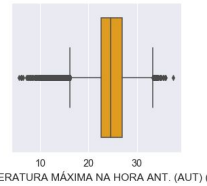


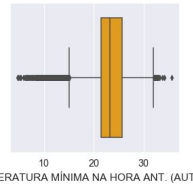


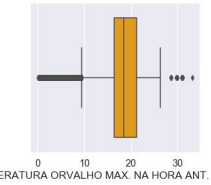


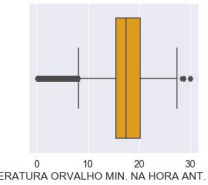


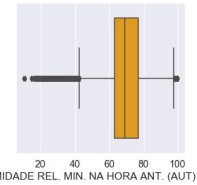


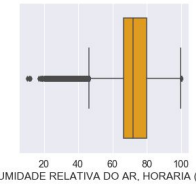


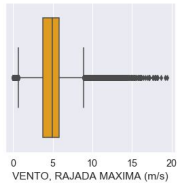


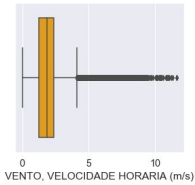


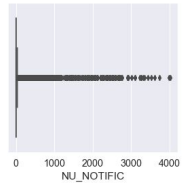












Pela análise de Boxplot é possível perceber que todas as variáveis lidam com uma quantidade significativa de outliers.

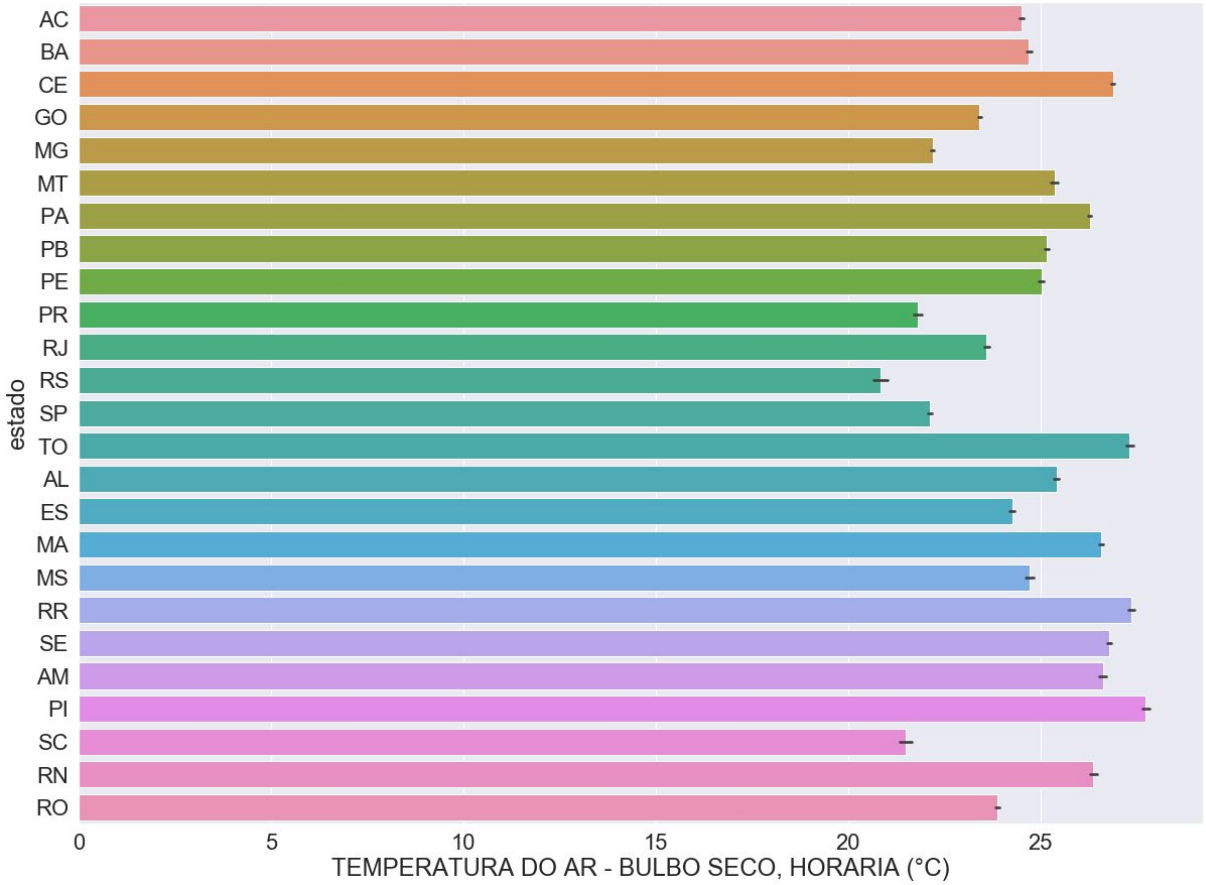
Será necessário entender o quão relevante isso será para a modelagem. Como estão sendo analisados dados do Brasil inteiro, é possível entender que existam diferenças climáticas significativas entre as cidades analisadas.

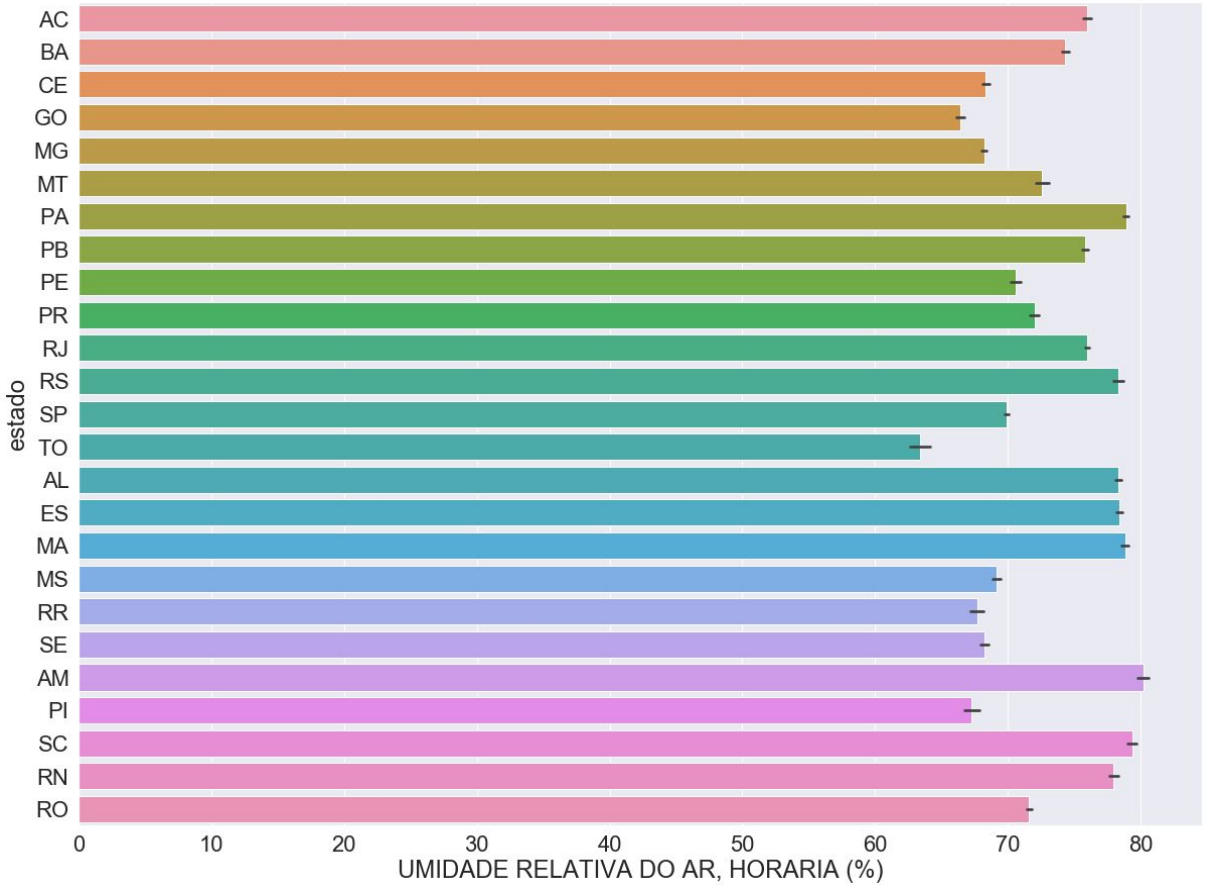
Em alguns casos como a variável de precipitação, os valores são extremamente concentrados em valores baixos, mostrando que nas cidades analisadas a precipitação não foi significativa no período analisado.

O mesmo ocorre com a variável resposta de notificações, que também é baixo para a maioria dos casos mas ao mesmo tempo possui cidades que notificam muito, como os grandes centros (SP e RJ), o que explica os outliers.

4. Análise Exploratória de Dados

Temperatura e Umidade Relativa do Ar por Estado

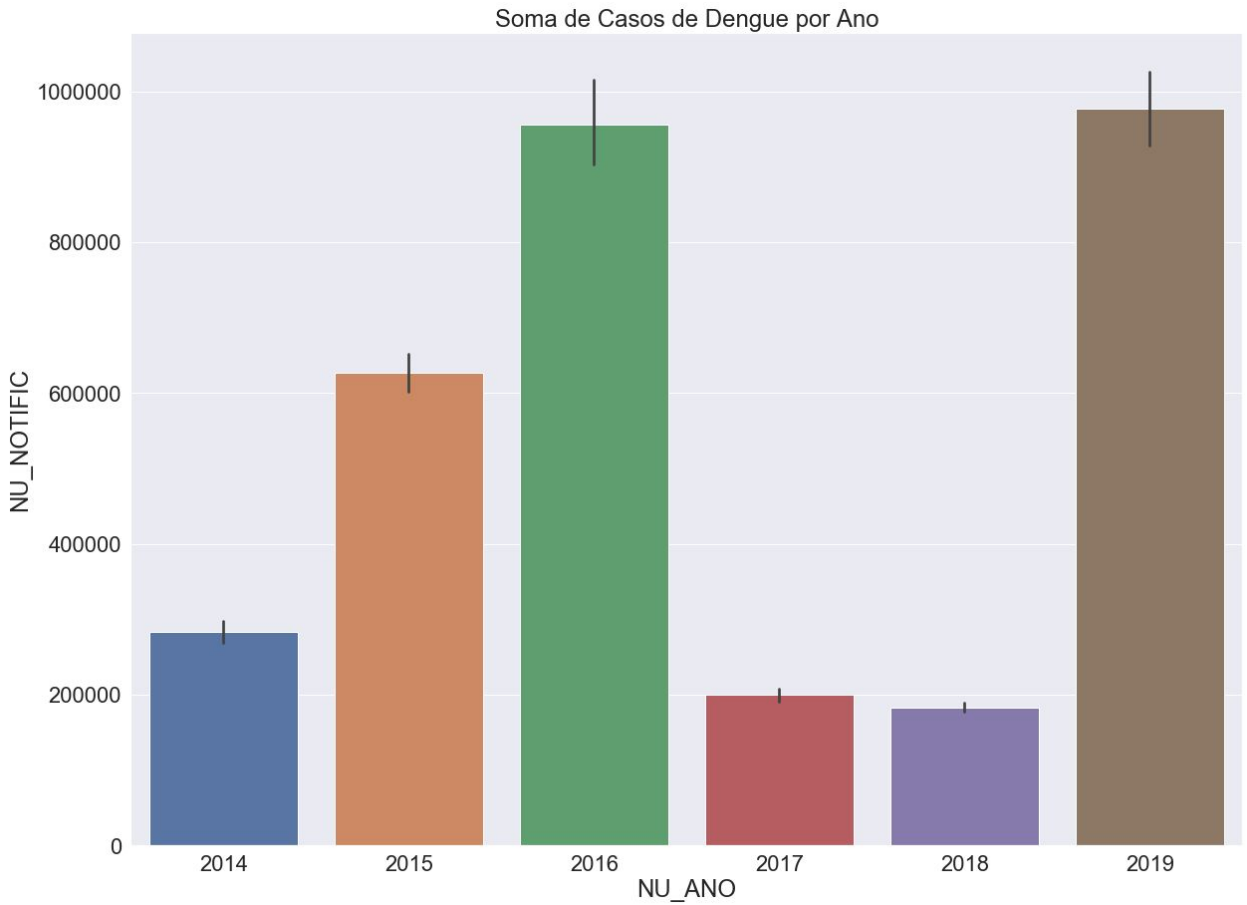


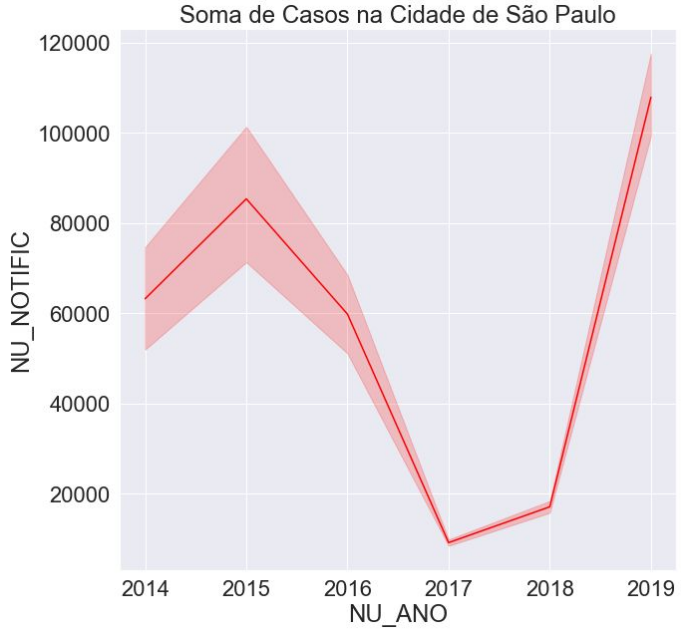


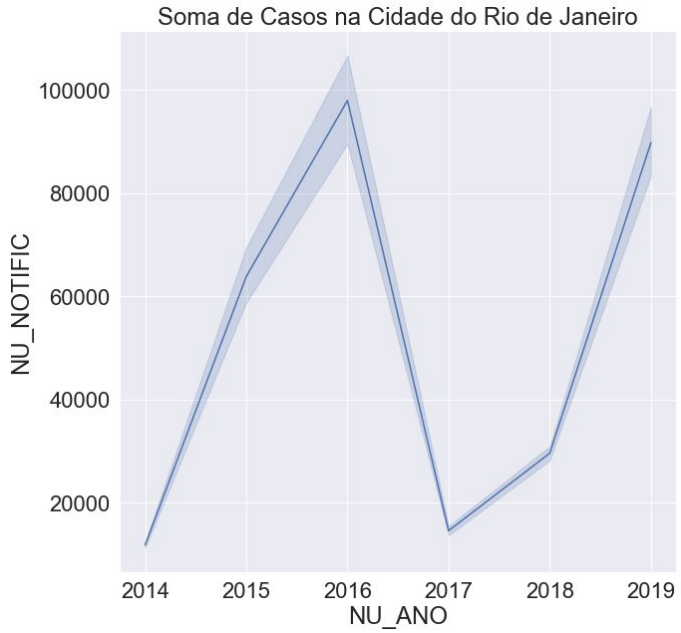
É possível perceber que em geral os estados com maior temperatura tem maior umidade relativa, como norte e nordeste. O que é esperado, mas é interessante perceber estados como SC que possuem um nível de precipitação bem alto apesar do clima menos quente.
Analisando as cidades de São Paulo e Rio de Janeiro, conhecidas pelo alto volume de notificações possuem valores intermediários de Temperatura e Umidade, o que mostra que apenas estas variáveis não explicam tão bem a variável resposta

4. Análise Exploratória de Dados

Número de Casos de Dengue notificados







Olhando para o número de casos de dengue é possível perceber um pico em 2016, e os valores subindo novamente chegando a patamares similares em 2019.

Como são duas cidades geralmente exploradas em noticiários quanto ao combate à dengue, olhamos para São Paulo e Rio de Janeiro no número de notificações.
Olhando para SP e RJ, temos uma realidade um pouco mais preocupante no caso de São Paulo, onde já temos valores atuais com topos históricos.

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

Metodologia de análise de dados

Planejamento | Análise Exploratória | **Modelagem Estatística** | Algoritmos de Inteligência Artificial | Conclusões

Definição do problema

- Objetivos
- Contextualização do problema
- Base de Dados

Variáveis Qualitativas

- Tabela de Frequência

Variáveis Quantitativas

- Medidas de Posição
- Medidas de Dispersão
- Boxplot

Estatística Tradicional

- Regressão Linear
- Árvore de Decisão

Inteligência Artificial e Otimização

- Random Forest
- XGBoost
- LGBM
- Cross Validation

5. Modelagem com Estatística Tradicional

Preparação da base para modelagem

Passo 1

Remoção de variáveis redundantes

Cidade x ID_MUNICIP por exemplo, Estado x Nome_UF. Removemos variáveis em excesso que tinham significado igual.
Foram mantidas as variáveis: Data, Latitude, Longitude, Variáveis Climáticas e Número de Notificações

Passo 2

Criação de Novas Variáveis

Como a dengue é uma doença que tem um ciclo: Com condições favoráveis de clima o mosquito se desenvolve e pode transmitir a doença que posteriormente será notificada, as variáveis climáticas foram derivadas em novas variáveis com média de 7 e de 30 dias da mesma métrica.

Passo 3

Remoção da Data

A princípio julgou-se que a data não seria uma variável interessante para influenciar a análise, já que estações do ano poderiam ser percebidas pelo clima não havendo necessidade de termos a data como influência.

Passo 4

Padronização dos Valores

Como a base possui muitos valores de escalas muito diferentes (exemplo da temperatura e pressão atmosférica), a base foi toda padronizada antes da modelagem para evitar desvios derivados da escala das variáveis.

5. Modelagem com Estatística Tradicional

Como estamos lidando com um problema de regressão, serão aplicadas duas técnicas a princípio:

1. Regressão Linear
2. Árvore de Decisão

Para avaliar os modelos foram usadas 3 métricas:

- R^2
- MAE
- RMSE

LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

5. Modelagem com Estatística Tradicional

Os resultados com cada uma das técnicas podem ser vistos abaixo:

Treino

Modelo	Regressão Linear
R^2	0,04
MAE	0,30
RMSE	0,98

Modelo	Árvore de Decisão
R^2	1,00
MAE	8,00E-18
RMSE	2,00E-17

Teste

Modelo	Regressão Linear
R^2	0,04
MAE	0,30
RMSE	0,98

Modelo	Árvore de Decisão
R^2	0,47
MAE	0,14
RMSE	0,73

É possível notar que o erro absoluto para a regressão linear é maior que a média de casos calculada.

Apesar de resultados melhores, o R^2 ajustado para a árvore de decisão é muito abaixo do esperado para uma previsão satisfatória

A princípio percebe-se que o modelo de regressão linear não está bem ajustado ao problema. Tanto o treino quanto o teste obtiveram resultados nada satisfatórios.

Apesar de resultado melhor, o modelo de árvore de decisão ainda pode ser melhorado, olhando a diferença entre treino e teste é possível verificar que o modelo sofreu com overfitting.

Nesta primeira tentativa não foi usada nenhuma técnica de validação cruzada e ajuste de hiperparâmetros dos modelos. Essa abordagem será usada posteriormente após comparação com técnicas de Inteligência Artificial.

Metodologia de análise de dados

Planejamento | Análise Exploratória | Modelagem Estatística | **Algoritmos de Inteligência Artificial** | Conclusões

Definição do problema

- Objetivos
- Contextualização do problema
- Base de Dados

Variáveis Qualitativas

- Tabela de Frequência

Variáveis Quantitativas

- Medidas de Posição
- Medidas de Dispersão
- Boxplot

Estatística Tradicional

- Regressão Linear
- Árvore de Decisão

Inteligência Artificial e Otimização

- Random Forest
- XGBoost
- LGBM
- Cross Validation

5. Modelagem com Inteligência Artificial

Para continuação do problema, foram aplicadas algumas técnicas de inteligência artificial para problemas de regressão:

1. Random Forest
2. XGBoost
3. LightGBM

Para avaliar os modelos foram mantidas as 3 métricas anteriores:

- R^2
- MAE
- RMSE

5. Modelagem com Inteligência Artificial

Os resultados com cada uma das técnicas podem ser vistos abaixo:

Treino

Modelo	Random Forest
R^2	0,94
MAE	0,04
RMSE	0,22

Modelo	XGBoost
R^2	0,61
MAE	0,19
RMSE	0,61

Modelo	LightGBM
R^2	0,86
MAE	0,13
RMSE	0,36

Teste

Modelo	Random Forest
R^2	0,71
MAE	0,11
RMSE	0,53

Modelo	XGBoost
R^2	0,58
MAE	0,19
RMSE	0,64

Modelo	LightGBM
R^2	0,69
MAE	0,15
RMSE	0,55

Com os modelos de inteligência artificial foi possível perceber uma melhora nos resultados, já não tivemos o mesmo problema de overfitting que foi encontrado com a Árvore de Decisão.

Ainda sim, é possível melhorar os resultados dos modelos, dado que ainda não foram usadas técnicas de otimização como o cross-validation por exemplo.

5. Escolha dos modelos com melhor desempenho

Com os resultados coletados, foram escolhidos 2 modelos para serem levados a uma etapa posterior de otimização de resultados

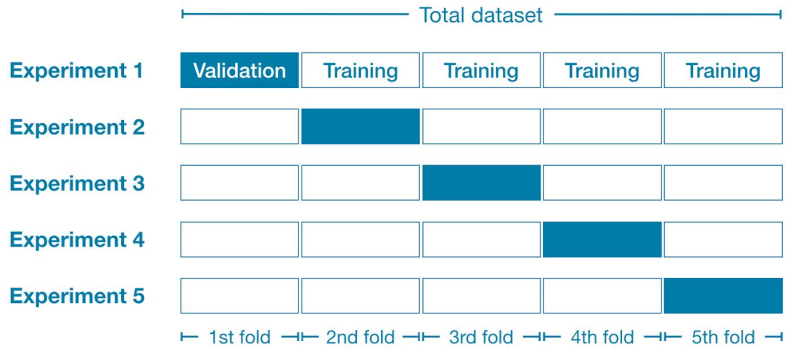
	1	2	3	4	5
	Regressão Linear	Árvore de Decisão	Random Forest	XGBoost	LightGBM
R^2	0,04	0,47	0,71	0,58	0,69
MAE	0,30	0,14	0,11	0,19	0,15
RMSE	0,98	0,73	0,53	0,64	0,55

LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

5. Cross Validation

Com os 2 modelos escolhidos (Random Forest e LGBM) foi usada a técnica de cross validation para otimização dos resultados

Foi usado com 5 folds e os parâmetros observados podem ser vistos abaixo.



Random Forest:

- max_depth [30,50,100]
- n_estimators [50,100,150]

R^2

0,71 → **0,73**

antes **depois**

LGBM:

- max_depth [30,50,100]
- boosting_type [gbdt, dart, goss]
- n_estimators [50,100,150]

R^2

0,69 → **0,70**

antes **depois**

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

6. Conclusões

Após análise dos dados e aplicação das técnicas de estatística tradicional foi possível perceber:

Foi possível perceber que os modelos de inteligência artificial tiveram desempenho superior aos de estatística tradicional no problema em questão.

Dentro de todos os modelos testados, o Random Forest e o LGBM tiveram o melhor desempenho, com um R2 ~ 0,7 e sem problemas significativos de overfitting ou underfitting.

Como a quantidade de dados não é tão grande após todos os cruzamentos de informação, não foi considerado para análise nenhum modelo de deep learning, mesmo assim os modelos usados foram capazes de resolver o problema de forma satisfatória.

Caso seja necessário aumentar ainda mais a precisão, uma boa estratégia pode ser encontrar mais variáveis que expliquem o problema, como dados socioeconômicos por exemplo.