Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Henrique Naoto Kato

Análise Quantitativa de Febre Amarela no Brasil

Curitiba
2023

Henrique Naoto Kato

Análise Quantitativa de Febre Amarela no Brasil

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. André R. A. Grégio

Curitiba
2023

Análise Quantitativa de Febre Amarela no Brasil

Quantitative Analysis of Yellow Fever in Brazil

Henrique Naoto Kato[1], André R. A. Grégio[2]

[1]Aluno do programa de Especialização em Data Science & Big Data*
[2]Professor do Departamento de Informática, Universidade Federal do Federal, Rua Cel. Francisco Heráclito dos Santos, 100 – Centro Politécnico, Jardim das Américas, 81531-980, Curitiba, Paraná, Brasil †

Neste projeto é feita uma análise quantitativa das infecções causadas por febre amarela que ocorreram no Brasil entre 1995 e 2021. Apresenta-se o método de análise utilizado, a origem de obtenção do *dataset*, os processos aplicados para geração de gráficos e informações, bem como os resultados obtidos.
Palavras-chave: Febre Amarela, Árvore de Decisão, Ciência de Dados

In this project, a quantitative analysis of infections caused by yellow fever that occurred in Brazil between 1995 and 2021 is conducted. In the project, it presents the analysis method utilized, the location where the dataset was taken from, the process that was applied to generate the graphs and the information, and the results gathered.
Keywords: Yellow Fever, Decision Tree, Data Science

1. Introdução

A febre amarela é uma doença cujo vetor de disseminação ocorre através das picadas dos mosquitos e é comum na América Latina. Assim, diversos países exigem a vacinação contra a febre amarela para permitir a entrada de turistas em seus territórios.[4] [3]. Apesar de não ser transmitida de pessoa para pessoa, ela é uma doença grave que apresenta alguns sintomas inciais como calafrios, dor de cabeça, náuseas, vômitos, entre outros. Em torno de 20 a 50% das pessoas que desenvolvem a doença podem morrer. [4]

Este estudo está sendo realizado para entender melhor os casos que ocorreram no país durante 25 anos.

2. Base de Dados

A base que foi utilizado está disponível no site do SUS [1], no qual tem:

- 2758 linhas
- 14 colunas
- dados desde 1995 até 2021.

Como cada linha é um paciente infectado pela doença, temos então 2758 casos de febre amarela. A tabela 1 mostra parte da base de dados.

Variáveis	Descrição
ID	ID do paciente
MACRORREG_LPI	Macro região
COD_UF_LPI	Código da macro região
UF_LPI	Código da unidade federada do local provável
COD_MUN_LPI	Código do município do local provável
MUN_LPI	Município
SEXO	Sexo do paciente
IDADE	Idade do paciente
DT_IS	Data do início da data dos sintomas
SE_IS	Semama do início da data dos sintomas
MES_IS	Mês do início da data dos sintomas
ANO_IS	Ano do início da data dos sintomas
MONITORAMENTO_IS	Ano de monitoramento
OBITO	Se houve óbito do paciente

Tabela 1: Nomes de colunas da tabela e descrição desses campos

Na figura 1 mostra os dados iniciais da tabela

*kato.henrique@gmail.com
†gregio@inf.ufpr.br

ID	MACRORREG_LPI	COD_UF_LPI	UF_LPI	COD_MUN_LPI	MUN_LPI
1	N	14	RR	140005	Alto Alegre
2	N	14	RR	140045	Pacaraima
3	NE	21	MA	210060	Amarante do Maranhão
4	NE	21	MA	210060	Amarante do Maranhão
5	N	15	PA	150270	Conceição do Araguaia
6	N	13	AM	130410	Tapauá

Figura 1: Exemplos de amostra dos dados correspondentes às 6 primeiras linhas e 7 colunas do dataset

2.1. Pré processamento

Antes de realizar alguma análise nos dados, foi feito limpeza de modo a evitar erros, portanto foram retirados todas as linhas que na coluna idade estavam vazias e na de óbitos foram eliminadas todas com "IGN", no qual significa ignorados. No final sobrou 2669 linhas, ou seja, 2669 pacientes infectados.

3. Método

Para realizar a análise foi utilizada a linguagem R e as bibliotecas utilizadas estão na tabela 2. Nela apresenta um breve resumo sobre o que cada pacote é utilizada [2].

Biblioteca	Descrição	Versão
tidyverse	Pacote com o foco em Data Science	2.0.0
ggplot2	Biblioteca para gerar gráficos	3.4.2
dplyr	Foco em função para manipular os dados	1.1.1
geobr	Pacote que disponibiliza informações sobre o mapa do Brasil	1.7.0
rpart	Biblioteca de particionamento recursivo e árvore de regressão	4.1.19
rpart.plot	Biblioteca para plotar um modelo rpart	3.1.1

Tabela 2: Nomes das bibliotecas utilizadas, mostrando uma breve descrição e a versão utilizada

Figura 2: Imagens que representam cada biblioteca utilizada

3.1. Índice de Gini

O índice de Gini é mais conhecido como uma medida de desigualdade, no qual foi desenvolvido pelo italiano Corrado Gini. Este índice é geralmente utilizado para calcular a desigualdade de distribuição de renda, pobreza, desenvolvimento econômico, entre outros. O resultado disto é sempre um número entre 0 e 1. O número 0 indica igualdade perfeita, com todas as classes com a mesma proporção,. Já quando o valor indica 1 a desigualdade seria a máxima. [9]

Em uma aplicação em árvore de decisão, em vez do índice ser utilizado para calcular a desigualdade, ele é utilizado para medir a pureza dos nós (no tópico de árvore de decisão é aprofundado um pouco mais), criando assim uma árvore.

Já a sua fórmula é apresentada na fórmula abaixo [8].

$$Indice\ de\ Gini = 1 - \sum (p_i)^2 \quad (1)$$

3.2. Árvore de decisão

A árvore de decisão é um tipo de algoritmo de aprendizado de máquina que pode ser usado para resolver problemas de regressão e classificação, podendo ter uma predição categórica ou numérica.

Existem alguns tipos de árvore de decisão como: ID3, C4.5 e `Classification and regression tree` (CART) . A árvore de decisão utilizado pela biblioteca rpart é o de Leo Breiman, ou seja, que seria o CART [8].

Para criar a árvore de decisão, é começado pelo nó, que contém todos os dados, depois dividimos o nó em subconjuntos e cada subconjunto é separado em outros subconjuntos. Esse processo é repetido até chegar na profundidade máxima da árvore [10].

O método CART utiliza o índice de Gini para identificar o atributo ideal para separar os atributos. A árvore busca encontrar a divisão que minimize a impureza de Gini nos subconjuntos [11].

Um exemplo simples é apresentado abaixo, em que temos uma tabela que mostra a data, o panorama do dia, a temperatura, umidade, o vento e se a pessoa vai jogar ou não. Com base na tabela 3, é possível criar a árvore de decisão da figura 3 [7].

4. Resultados e Discussões

Após a realização da limpeza dos dados, é iniciado uma análise para entender a quantidade de pessoas infectadas, conforme pode-se observar na figura 4. Nela, é

Dia	Panorama	Temperatura	Umidade	Vento	Jogar
1	Sol	Quente	Alta	Fraco	Não
2	Nublado	Quente	Alta	Fraco	Sim
3	Sol	Moderado	Normal	Forte	Sim
4	Nublado	Moderado	Alto	Forte	Sim
5	Chuva	Moderado	Alto	Forte	Não
6	Chuva	Frio	Normal	Forte	Não
7	Chuva	Moderado	alto	Fraco	Sim
8	Sol	Quente	Alta	Forte	Não
9	Nublado	Quente	Normal	Fraco	Sim
10	Chuva	Moderado	Alto	Forte	Não

Tabela 3: Dataset utilizada de exemplo, mostrando se o dia está apto para um jogar, mostrando os atributos que são avaliados

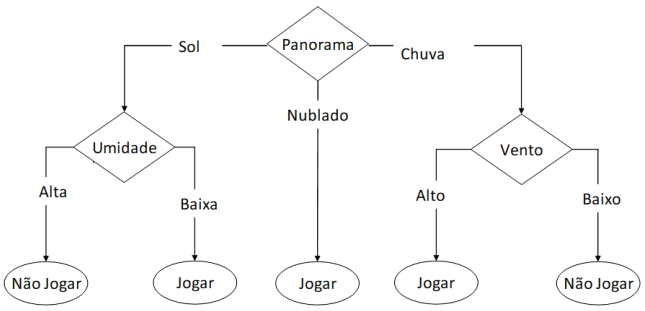

Figura 3: Exemplo de uma árvore de decisão utilizando os dados da tabela 3

verificado o número de pessoas infectadas por ano e a quantidade de óbitos que ocorreram durante o período. Durante os 25 anos dos dados analisados o pico ocorreu no ano de 2018 com 1293 infecções e 446 óbitos.

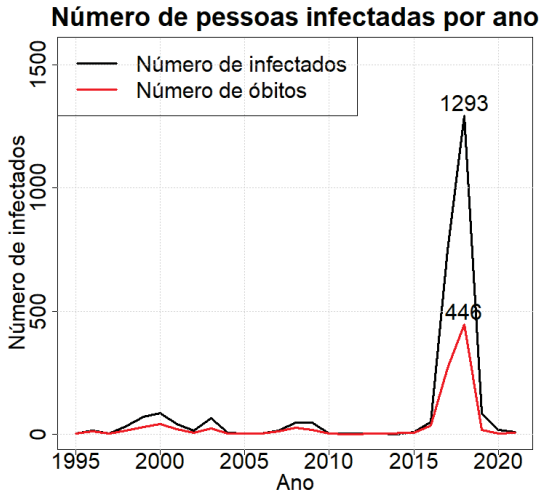

Figura 4: Gráfico de pessoas infectadas no ano

Já na Figura 5, mostra que o período que as pessoas mais se infectam com a doença é no período de calor, sendo entre o mês de dezembro e maio, sendo janeiro o mais propício.

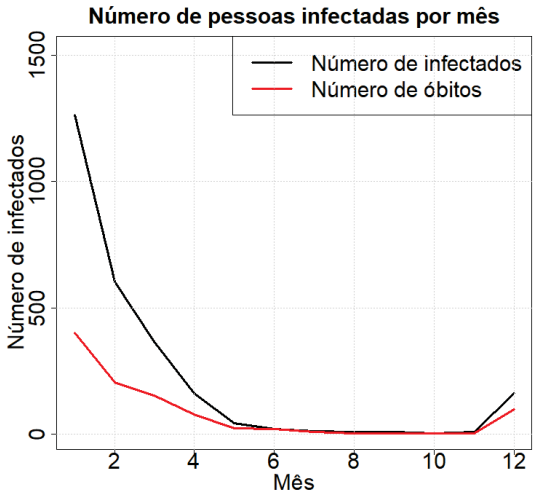

Figura 5: Gráfico de pessoas infectadas no mês

Na sequência é feito um gráfico para verificar qual o sexo mais infectado (figura 6), verificando que o sexo masculino acabou sendo a maioria da amostra, com 83% de infectados.

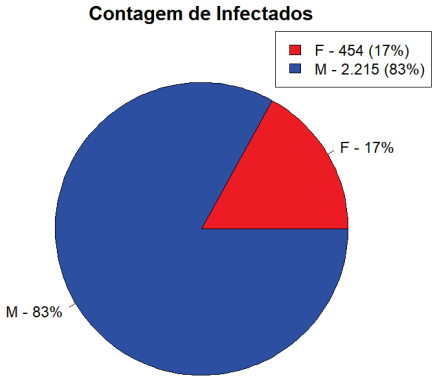

Figura 6: Infecções separado pelo sexo

Para aprofundar a análise, foi feito a proporção de óbito e pode-se verificar que a taxa de óbito foi maior nos homens, representando 15% a mais que as mulheres.

A faixa etária mais afetada foi entre 40 a 50 anos de idade. Este mesmo padrão é notado quando classificado por sexo.

Já analisando por região, descobrimos que o Sudeste é onde ocorreu as maiores taxas, sendo que os estados de São Paulo e Belo Horizonte tiveram as maiores índices.

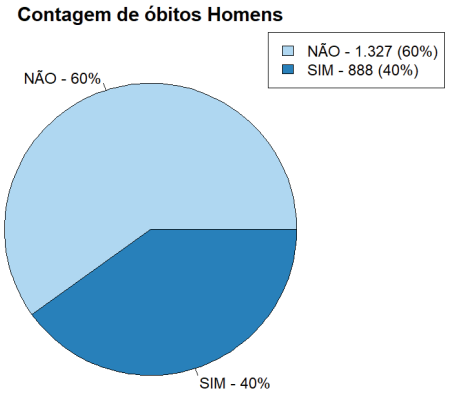

Figura 7: Gráfico de pessoas infectadas no mês - Homem

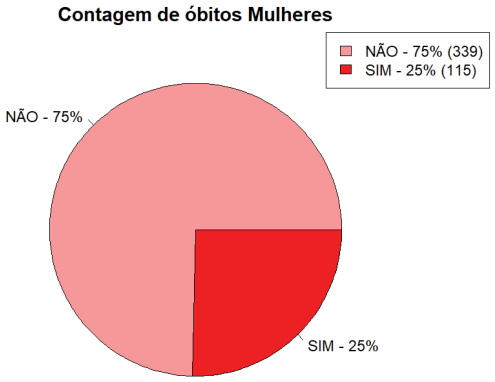

Figura 8: Gráfico de pessoas infectadas no mês - Mulher

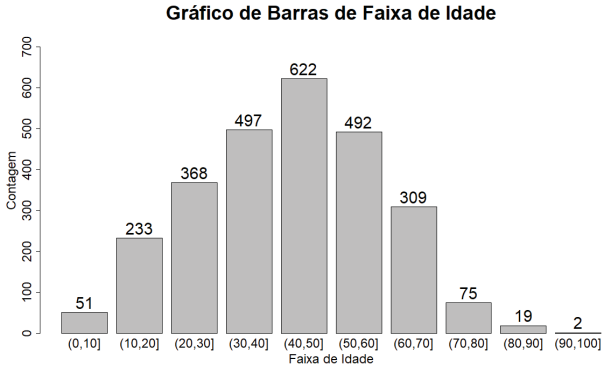

Figura 9: Faixa de Infectados por idade

Antes de realizar a árvore de decisão, foi feito uma análise aprofundando os pontos de pico, no qual seriam os anos de 2017 e 2018. Inicialmente, foi pensado em utilizar apenas os dados destes anos, no qual teria 2040 infecções. Porém, verificando a tabela 4 por região, é notado que a maioria das infecções são no

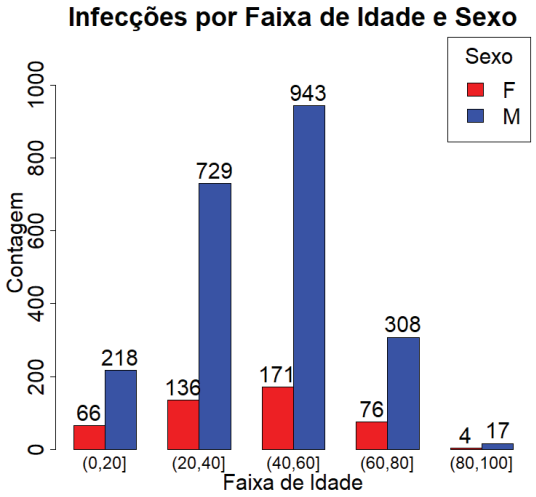

Figura 10: Faixa de Infectados por faixa de idade e sexo

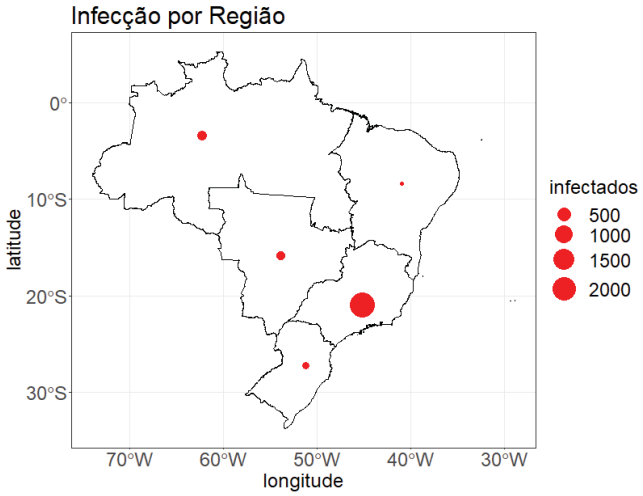

Figura 11: Infecção por região no Brasil

Sudeste. Como a infecção foi em apenas uma região, o resultado ficaria muito tendencioso para esta área.

Região	Quantidade Infectados
Norte	7
Nordeste	0
Centro Oeste	4
Sudeste	2029
Sul	0

Tabela 4: Tabela com os dados consolidados de infecções por região

Em seguinte foi pensado em utilizar os dados que não fossem o pico, porém iria ficar com apenas apenas 629 linhas, ou seja 629 pacientes. Após realizar estas análises decidimos utilizar todos os dados.

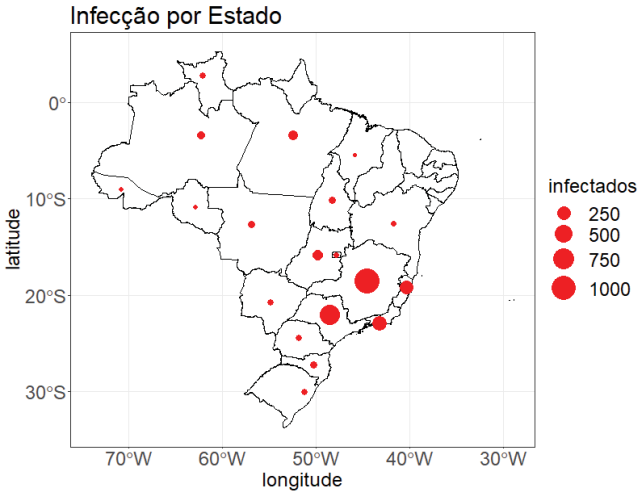

Figura 12: Gráfico das Infecção por estado no Brasil

4.1. Árvore de Decisão

Obtendo estas informações, foi aplicada uma árvore de decisão tradicional, sem *fine tuning*. A divisão do *dataset* foi de 80% dos dados para a realização dos treinamentos e 20% para os testes. Na Figura 5, observa-se o resultado e também as probabilidades de uma pessoa infectada ter óbito ou não, dependendo de algumas características.

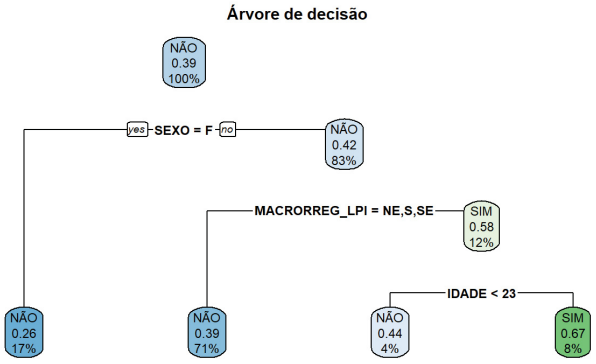

Figura 13: Caso onde todas as medidas de distâncias são exatas

Utilizando este modelo de realização de engenharia de características ou modificação de configurações, obteve-se um resultado de 68,16% de acurácia.

5. Conclusão

Com base nos resultados preliminares alcançados pela análise dos dados presentes no *dataset* coletado, pôde-se concluir que foi possível entender melhor as infecções que ocorreram no Brasil e também desenvolver uma árvore de decisão que, se ajustada e analisada mais aprofundadamente, pode auxiliar na previsão de óbitos em casos por região, idade e sexo.

Através da árvore de decisão, pode-se observar que as pessoas que tiveram o maior risco de óbito foram os homens, que se infectaram na região Norte ou Centro-Oeste, e que possuíam a idade maior que 23 anos. Caso a pessoa adentrasse dentro desses critérios, eles tinham um 67% chance de óbito. O grupo que teve a maior representatividade foram os homens que nasceram no Nordeste, Sudeste e Sudoeste, sendo 71% dos nossos dados, com 39% chance de falecer. Um dado interessante mostrado nos resultados foi que apenas o fato de ser mulher, você já tem grandes chances de sobreviver. Um ponto no qual seria interessante aprofundar, para verificar quais outros fatores poderiam estar faltando ou se realmente é uma informação significativa.

Ainda se vê uma necessidade de melhorar os resultados, pois a acurácia de 68% não é satisfatória para um sistema de suporte à decisão. Porém, foi possível compreender o ferramental utilizado e começar um entendimento sobre quais são as pessoas com mais fatores de risco.

Um sistema futuro pode refinar o modelo e os dados, visando dar suporte a um profissional de saúde a fim de priorizar pacientes com febre amarela baseado em suas características.

Agradecimentos

Agradeço pelo meu orientador, o professor André Grégio pela paciência e atenção, podendo assim realizar e finalizar este trabalho. Obrigado à minha família pelo apoio e pelo suporte durante este período. E por final agradeço também o Departamento de Informática, que possibilitou a realização desta pós-graduação.

Referências

[1] Ministério da Saúde: *Febre Amarela em humanos e primatas não-humanos - 1994 a 2021* https://opendatasus.saude.gov.br/dataset/febre-amarela-em-humanos-e-primatas-nao-humanos.

[2] CRAN *The Comprehensive R Archive Network* https://cran.r-project.org/.

[3] Ministério da Saúde: *Tirar o Certificado Internacional de Vacinação* https://www.gov.br/pt-br/servicos/obter-o-certificado-internacional-de-vacinacao-e-profilaxia.

[4] Ministério da Saúde: *Febre Amarela* https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/f/febre-amarela.
[5] CRAN *Package 'rpart'* https://cran.r-project.org/web/packages/rpart/rpart.pdf.
[6] CRAN *An Introduction to Recursive Partitioning* https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.
[7] CRAN *Decision Tree Alhorithm - Complete Guide* https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/.
[8] IBM *What is a Decision Tree* https://www.ibm.com/topics/decision-trees.
[9] USP *Entendendo o índice de Gini* https://edisciplinas.usp.br/pluginfile.php/7767443/mod_resource/content/0/Desigauldade_Gini.pdf.
[10] T. DANIYA1, M. GEETHA, E K. SURESH KUMA *Classification and Regression trees with Gini Index* https://www.researchgate.net/profile/Suresh-Kumar-K-Dr/publication/344385674_Classification_and_regression_trees_with_gini_index/links/5f780b4d92851c14bca9e8a5/Classification-and-regression-trees-with-gini-index.pdf.
[11] S. Dan *CART: Classification and Regression trees* https://www.researchgate.net/profile/Dan-Steinberg/publication/265031802_Chapter_10_CART_Classification_and_Regression_Trees/links/567dcf8408ae051f9ae493fe/Chapter-10-CART-Classification-and-Regression-Trees.pdf