

# Usando a estatística para prever resultado de teste de COVID-19 no app Dados do Bem

Paula Maçaira

Semana da Estatística 2022 - UFJF

## Quem sou eu

- ▶ Bacharelado em Estatística (ENCE, 2013)
- ▶ Mestrado em Eng. Elétrica (PUC-Rio, 2015)
- ▶ Doutorado em Eng. de Produção (PUC-Rio, 2018)
- ▶ Pós-doutorado em Eng. de Produção (PUC-Rio, 2019)
- ▶ Professora Adjunta desde 2019 (DEI, PUC-Rio)

We're data-driven!



## Me encontre em

- ▶ [@paula\\_macaira](https://twitter.com/paula_macaira)
- ▶ [github.com/paulamacaira](https://github.com/paulamacaira)
- ▶ [sites.google.com/view/paulamacaira](https://sites.google.com/view/paulamacaira)
- ▶ [paulamacaira@puc-rio.br](mailto:paulamacaira@puc-rio.br)



# Grupos que faço parte



Forecasting and Resource Optimization Group



Let's Go

**Citation:** Dantas LF, Peres IT, Bastos LSL, Marchesi JF, de Souza GFG, Gelli JGM, et al. (2021) App-based symptom tracking to optimize SARS-CoV-2 testing strategy using machine learning. PLoS ONE 16(3): e0248920. <https://doi.org/10.1371/journal.pone.0248920>

**Published:** March 25, 2021

**Copyright:** © 2021 Dantas et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The data underlying the results presented in the study are available from [https://github.com/noispuc/Dantas\\_eta\\_PLOSOne\\_App-based-symptom](https://github.com/noispuc/Dantas_eta_PLOSOne_App-based-symptom).

---

#### RESEARCH ARTICLE

## App-based symptom tracking to optimize SARS-CoV-2 testing strategy using machine learning

Leila F. Dantas  <sup>1\*</sup>, Igor T. Peres  <sup>1\*</sup>, Leonardo S. L. Bastos  <sup>1\*</sup>, Janaina F. Marchesi  <sup>2†</sup>, Guilherme F. G. de Souza <sup>1‡</sup>, João Gabriel M. Gelli <sup>1‡</sup>, Fernanda A. Baião <sup>1‡</sup>, Paula Maçaira  <sup>1‡</sup>, Silvio Hamacher <sup>1‡</sup>, Fernando A. Bozza  <sup>3,4‡\*</sup>

**1** Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil, **2** Instituto Tecgraf, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil,

**3** National Institute of Infectious Diseases Evandro Chagas (INI), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, RJ, Brazil, **4** D'Or Institute for Research and Education (IDOR), Rio de Janeiro, RJ, Brazil

● These authors contributed equally to this work.

† These authors also contributed equally to this work.

\* [bozza.fernando@gmail.com](mailto:bozza.fernando@gmail.com), [fernando.bozza@ini.fiocruz.br](mailto:fernando.bozza@ini.fiocruz.br)

# A pandemia de COVID-19 e a importância dos testes

- ▶ A pandemia de COVID-19 requereu extensos programas de testes para entender a transmissão, diagnosticar e isolar os casos positivos
- ▶ Dada a alta mortalidade e a ausência de um tratamento específico ou de uma vacina confiável, grandes programas de testes foram parte essencial do controle da epidemia
- ▶ A frequência dos testes, no entanto, é muito heterogênea entre os países, no Brasil, as taxas de testagem foram uma das mais baixas<sup>1</sup> no mundo, tornando os sistemas de triagem essenciais para priorizar a testagem

---

<sup>1</sup>120.548 testes por um milhão de habitantes, em 02 de dezembro de 2020

## Objetivo do estudo

Sabendo então da importância da triagem para priorização da testagem para COVID-19, o presente estudo usou a **combinação de sintomas e técnicas de aprendizado de máquina para desenvolver um modelo preditivo** que identifique pessoas e áreas com maior risco de infecção por SARS-CoV-2

## Fonte dos dados

- ▶ Este estudo utilizou dados coletados de indivíduos cadastrados no aplicativo “Dados do Bem”<sup>2</sup>
- ▶ Por meio de uma breve pesquisa, o aplicativo coleta:
  - ▶ Dados georreferenciados dos usuários inscritos
  - ▶ Características demográficas e ocupacionais
  - ▶ Sintomas autorreferidos
  - ▶ Se o participante é profissional de saúde
  - ▶ Se o participante esteve em contato com uma pessoa infectada por SARS-CoV-2

---

<sup>2</sup>lançado no Brasil em 28 de abril de 2020

# Interface do aplicativo

<p>Register to start a self assessment</p> <p>Name Birth date Gender Social security number Email Phone number Zipcode</p> <p><b>NEXT</b></p>	<p>Question 01</p> <p>Have you had fever in the last 7 days (Temperature 38 or higher)?</p> <p><b>YES</b> <b>NO</b></p>	<p>Question 02</p> <p>Check the symptoms you are experiencing right now:</p> <p><input checked="" type="checkbox"/> Cough <input type="checkbox"/> Nausea/Vomiting <input type="checkbox"/> Sore throat <input type="checkbox"/> Diarrhea <input type="checkbox"/> Runny nose/Nasal congestion <input type="checkbox"/> Loss of smell/taste <input type="checkbox"/> Shortness of breath or difficulty breathing <input type="checkbox"/> Muscle or body pain/myalgia</p> <p><b>NEXT</b></p>	<p>Question 03</p> <p>Risk factor: Do you have a chronic disease?</p> <p><input checked="" type="checkbox"/> Heart disease <input type="checkbox"/> Pulmonary <input type="checkbox"/> Diabetes <input type="checkbox"/> Obesity <input checked="" type="checkbox"/> Cirrhosis <input checked="" type="checkbox"/> Neoplasia <input checked="" type="checkbox"/> Kidney disease <input checked="" type="checkbox"/> Hepatopathy <input checked="" type="checkbox"/> Transplanted patient <input checked="" type="checkbox"/> Regularly take immunosuppressant medications</p> <p><b>NEXT</b></p>	<p>Question 04</p> <p>Do you have a suspected or confirmed case of COVID-19 in your home?</p> <p><b>YES</b> <b>NO</b></p>	<p>Question 05</p> <p>Are you a healthcare worker?</p> <p><b>YES</b> <b>NO</b></p>
---	---	--	--	---	--

## Como era a priorização...

- ▶ Antes do estudo, o aplicativo combinava as informações pesquisadas e selecionava indivíduos para teste por meio de alguns critérios de seleção, como por exemplo: **aqueles que haviam sido indicados por um participante previamente testado positivamente tinha a maior prioridade para serem testados, seguidos pelos profissionais de saúde**

## Como o estudo foi desenhado

- ▶ Foram incluídos participantes cadastrados no aplicativo desde sua data de lançamento (28 de abril) até 16 de julho de 2020
- ▶ Para treinar o modelo, foram selecionados os participantes que responderam ao questionário, fizeram o teste do anticorpo e obteviram um resultado conclusivo (positivo ou negativo)
- ▶ Para a identificação das áreas de risco, foram incluídos também os participantes que não haviam sido testados, aplicando o modelo para estimar os resultados de seus testes

## Variáveis e objeto de estudo

- ▶ Como o objetivo do estudo foi identificar manifestações clínicas e fatores individuais associados a testes com resultados positivos, as variáveis coletadas e analisadas foram:
  - ▶ dados demográficos dos participantes (idade, sexo)
  - ▶ sintomas (perda de olfato ou anosmia, febre, mialgia, tosse, náusea, falta de ar, diarreia, coriza e dor de garganta)
  - ▶ se o usuário mora junto com alguém com infecção confirmada por SARS-CoV-2
  - ▶ resultado do teste de COVID

## Desenho do estudo

1. Análise estatística das base de dados
  2. Associação entre sintomas (individuais) e resultado do teste (**ajustado por idade e gênero**) através de regressão logística - isto é, 11 modelos, um para cada sintoma - obtendo o Odds Ratio (OR) de cada sintoma
  3. Aplicação de **cinco técnicas<sup>3</sup> diferentes de aprendizado de máquina** construir um modelo de previsão que determine, através da combinação de sintomas, a probabilidade de um participante estar infectado por SARS-CoV-2
- Para tratar o desequilíbrio da variável de resposta (apenas 11,8% são testes positivos) foram avaliadas quatro técnicas diferentes de balanceamento de dados<sup>4</sup>

---

<sup>3</sup>regressão logística (LR) stepwise, Naïve Bayes (NB), Random Forest (RF), Decision Tree usando C5.0 (DT) e eXtreme gradient Boosting

<sup>4</sup>Downsampling, Upsampling, Synthetic Minority Oversampling Technique (SMOTE) e Random Over-Sampling Exemplos (ROSE)

## Desenho do estudo

4. Dados foram divididos em conjunto de treinamento (80%) e um conjunto de teste (20%), mantendo a mesma proporção de classes majoritárias e minoritárias entre as subamostras
5. Após obter os melhores hiperparâmetros para cada combinação de modelo e técnica de balanceamento no grupo de treinamento, foi utilizado o Coeficiente de Correlação de Matthew (MCC) para avaliar os resultados nos conjuntos de teste - além do valor de MCC, consideramos a interpretabilidade do modelo para a escolha do modelo final

## Desenho do estudo

6. Por fim, foi avaliada a distribuição dos riscos de infecção por SARS-CoV-2 na área geográfica do Rio de Janeiro modelada como um mapa de grade (cada grade é uma área quadrada de 400m x 400m)
7. Juntamente com os participantes com resultados de testes confirmados, o modelo escolhido foi aplicado à amostra de participantes que ainda não haviam sido testados no período deste estudo para obter o resultado do teste estimado
8. A proporção<sup>5</sup> de infecções estimadas por SARS-CoV-2 para cada grade foi calculada como:

$$\text{Grade de Risco} = \frac{\text{quantidade de usuários positivos}}{\text{todos os usuários}}$$

---

<sup>5</sup>Para evitar interpretações erradas de proporções em grades com dados escassos, foram consideradas apenas grades com pelo menos 10 participantes

## Desenho do estudo

9. Após avaliar a distribuição dos riscos da rede entre todas as redes foram criados cinco grupos de risco usando a média e o desvio padrão (DP) como limites:
  - ▶ “muito baixo” ( $<$  média- $1,5 \cdot DP$ )
  - ▶ “baixo” (de média- $1,5SD$  até média- $0,5SD$ )
  - ▶ “médio” (de média- $0,5SD$  até média+ $0,5SD$ )
  - ▶ “alto” (de média+ $0,5SD$  até média+ $1,5SD$ )
  - ▶ “risco muito alto” ( $>$ média+ $1,5 \cdot DP$ )
10. A partir dessa classificação, construímos um mapa de risco para o Rio de Janeiro.

## Validação externa

- ▶ O modelo final foi incorporado ao aplicativo em 17 de julho de 2020 e para verificar os ganhos com esse modelo proposto, foi realizada uma validação com os dados do Rio de Janeiro
- ▶ Comparamos a proporção de resultados positivos antes da implementação do modelo no aplicativo (usando dados de 15 de junho de 2020 a 16 de julho de 2020) e após sua implementação (usando dados de 01 de agosto de 2020 a 01 de setembro de 2020)<sup>6</sup>
- ▶ Foi avaliado se a média das proporções de resultados positivos antes da implementação do modelo pode ser considerada menor do que a proporção média de resultados positivos após sua implementação através do teste de Wilcoxon

---

<sup>6</sup>O intervalo de duas semanas entre a incorporação do modelo no aplicativo e a validação foi necessário, pois ainda havia testes agendados de acordo com a política de priorização anterior

## Disponibilidade dos dados

Os dados e o código utilizados no estudo foram disponibilizados em  
[https://github.com/noispuc/Dantas\\_etal\\_PLOSOne\\_App-based-symptom](https://github.com/noispuc/Dantas_etal_PLOSOne_App-based-symptom)

# Resultados - Análise descritiva da base de dados

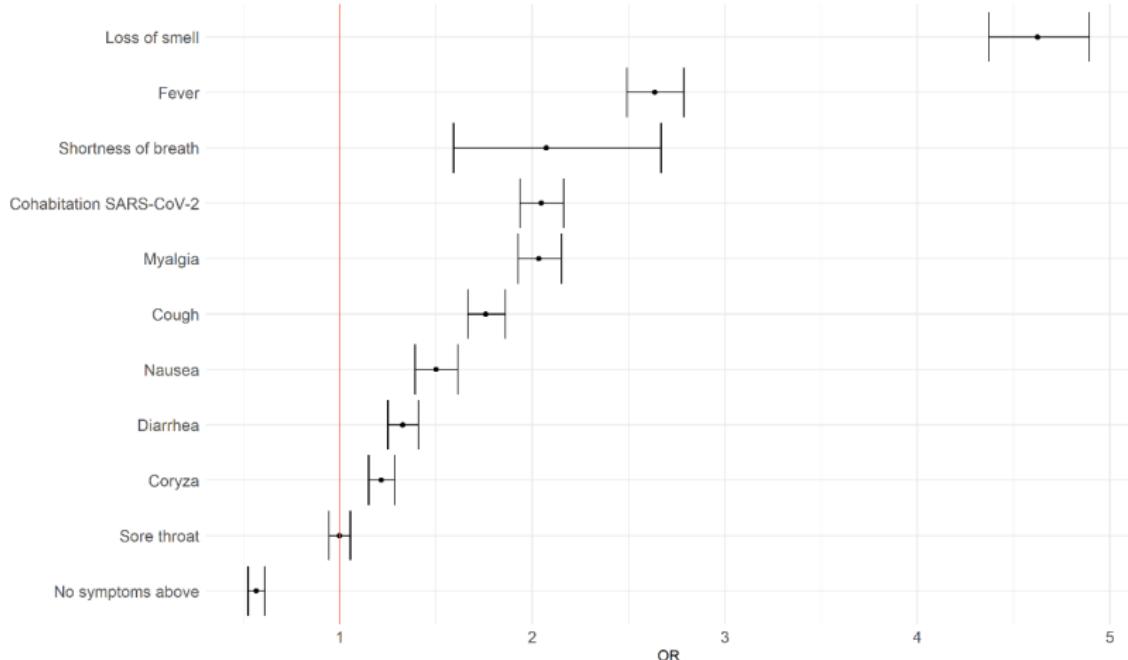
**Table 1. Characteristics and symptoms of the study population tested for SARS-CoV-2 infection.**

Results are displayed in median (interquartile range, IQR) for continuous variables and percentage values for categorical variables.

	Total	Positive test	Negative test
<b>Participants, n (%)</b>	49,721	5,888 (11.8)	43,833 (88.2)
<b>Characteristics</b>			
Female, n (%)	30,769 (61.9)	3,641 (61.8)	27,128 (61.9)
Age (years), median [IQR]	41 [33-51]	43 [34-53]	40 [33-51]
Cohabitation - lives with a SARS-CoV-2 infected person, n (%)	20,944 (42.1)	3,398 (57.7)	17,546 (40.0)
Health professional, n (%)	27,737 (55.8)	3,099 (52.6)	24,638 (56.2)
<b>Self-reported symptoms, n (%)</b>			
Coryza	25,973 (52.2)	3,315 (56.3)	22,658 (51.7)
Cough	23,430 (47.1)	3,507 (59.6)	19,923 (45.5)
Myalgia	20,858 (42.0)	3,380 (57.4)	17,478 (39.9)
Sore throat	20,794 (41.8)	2,459 (41.8)	18,335 (41.8)
Fever	13,042 (26.2)	2,640 (44.8)	10,402 (23.7)
Diarrhea	12,573 (25.3)	1,778 (30.2)	10,795 (24.6)
Loss of smell	11,835 (23.8)	3,112 (52.9)	8,723 (19.9)
Nausea	6,461 (13.0)	1,025 (17.4)	5,436 (12.4)
Shortness of breath	354 (0.7)	74 (1.3)	280 (0.6)
No symptoms above	10,865 (21.9)	844 (14.3)	10,021 (22.9)

De 28 de abril de 2020 a 16 de julho de 2020, **337.435 indivíduos** registraram seus sintomas por meio do aplicativo Dados do Bem

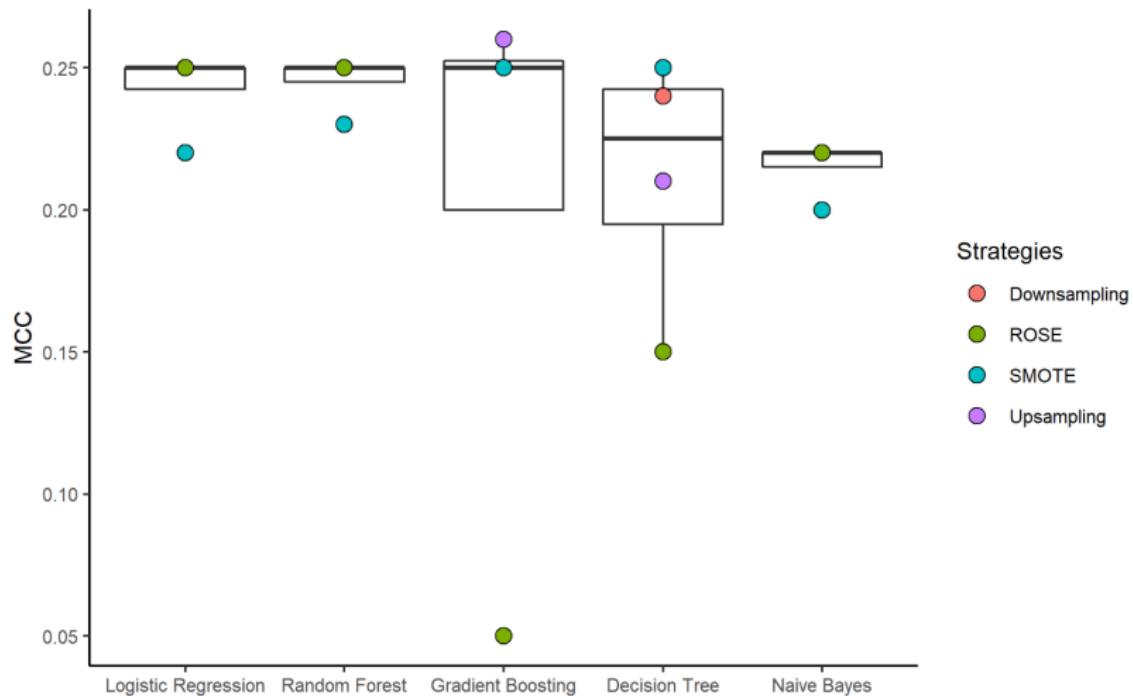
# Resultados - Odds Ratio dos sintomas



## Resultados - Combinação de sintomas e modelos

- ▶ Para desenvolver um modelo para prever participantes positivos com base no conjunto de dados disponível, foram executadas 25 combinações diferentes de técnicas de aprendizado de máquina e estratégias de amostragem.
- ▶ Foram avaliadas comparativamente o desempenho dos modelos no conjunto de teste de acordo com as métricas de Sensibilidade, Especificidade, Predictive Positive Value (PPV), Negative Predictive Value (NPV), F1-Score e MCC

# Resultados - Combinação de sintomas e modelos



## Resultados - Decisão do modelo final

- ▶ O modelo final escolhido foi o método de regressão logística combinado com a estratégia de balanceamento de upsampling

$$\text{Probability of testing positive} = \frac{e^{\textit{prediction}}}{1 + e^{\textit{prediction}}} \quad , \text{where}$$

$$\begin{aligned}\textit{prediction} = & -1.078 + (1.309 * \textit{loss of smell}) + (0.481 * \textit{fever}) + (0.407 * \textit{COVID at home}) \\ & + (0.338 * \textit{shortness of breath}) + (0.237 * \textit{myalgia}) + (0.153 * \textit{cough}) \\ & + (0.035 * \textit{nausea}) + (0.033 * \textit{gender[male]}) + (0.008 * \textit{age}) \\ & - (0.441 * \textit{sore throat}) - (0.227 * \textit{coryza}) - (0.045 * \textit{diarrhea})\end{aligned}$$

## Resultados - Performance no grupo de teste

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	706	471	Sensitivity 0.60
	Negative	2164	6602	Specificity 0.75
		PPV 0.25	NPV 0.93	Accuracy 0.73

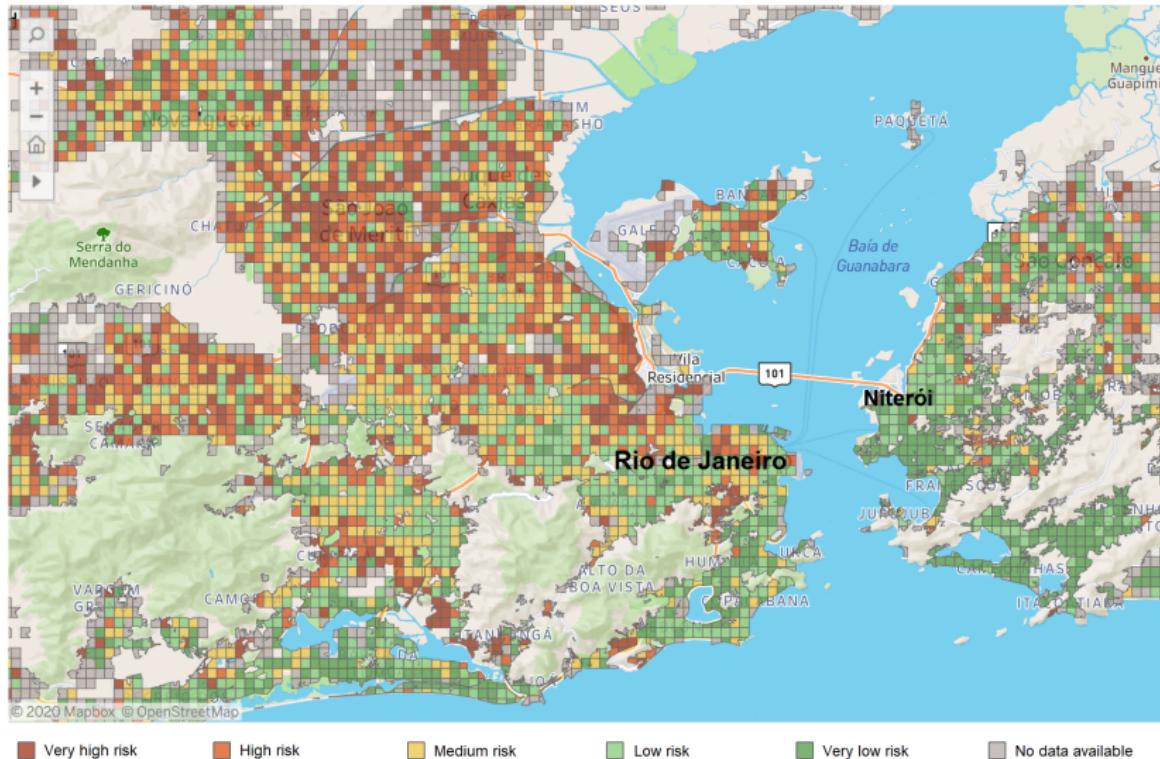
O modelo prediz corretamente quase todos os testes negativos com apenas 7% de usuários falso-negativos entre os previstos como negativos

# Resultados - Performance no grupo de teste

**Table 2. Characteristics and symptoms of false-negative and false-positive users predicted from our model.**

	<b>False-negative</b>	<b>False-positive</b>
<b>Participants, n (%)</b>	471	2,164
<b>Characteristics</b>		
Female, n (%)	258 (54.8)	1,356 (62.7)
Age (years), median [IQR]	42 [34-51]	42 [32-52]
Cohabitation - lives with a SARS-CoV-2 infected person, n (%)	179 (38.0)	1,464 (67.7)
<b>Self-reported symptoms, n (%)</b>		
Loss of smell	4 (0.8)	1,694 (78.3)
Fever	88 (18.7)	1,288 (59.5)
Myalgia	156 (33.1)	1,473 (68.1)
Cough	197 (41.8)	1,511 (69.8)
Nausea	43 (9.1)	429 (198)
Sore throat	173 (36.7)	1,062 (49.1)
Coryza	200 (42.5)	1,321 (61.0)
Diarrhea	88 (18.7)	731 (33.8)
Shortness of breath	2 (0.4)	32 (1.5)

# Resultados - Grade de risco



Aplicamos o modelo preditivo aos 287.714 indivíduos que se registraram no aplicativo (não testados) e 99.431 (34,5%) foram classificados como positivos

## Resultados - Validação externa

- ▶ O aplicativo “Dados do Bem” incorporou o modelo final em 17 de julho de 2020, utilizando-o para priorizar usuários para testes em alguns estados brasileiros
- ▶ A validação externa com dados do Rio de Janeiro compreendeu 57.762 testes de 01 de agosto a 01 de setembro, resultando em **18,1%** de resultados positivos (10.466/57.762)
- ▶ Se considerarmos os dados de 15 de junho a 16 de julho (antes da implantação do modelo), observamos apenas **14,9%** de positividade (5.296/35.626), indicando que o modelo incorporado aumentou a proporção de testes positivos
- ▶ Os resultados do teste de hipóteses mostraram diferença estatisticamente significativa entre a proporção de resultados positivos antes e após a implementação do modelo ( $p$ -valor < 0,001 com nível de confiança de 95%)

## Conclusões

- ▶ O trabalho usou dados sobre sintomas individuais e dados demográficos obtidos de um sistema baseado em aplicativo para prever indivíduos com maior probabilidade de serem infectados por SARS-CoV-2
- ▶ Foi desenvolvido um modelo de triagem visando priorizar os usuários para testes e após a aplicação do modelo, dos 57.762 usuários selecionados, 18,1% foram testados positivos
- ▶ Essa taxa de positividade foi mais significativa do que a observada sem modelo (14,9%), o que indica que o modelo contribuiu para melhorar a estratégia de teste e selecionar os usuários com maior probabilidade de serem positivos no cenário atual
- ▶ Além disso, foi desenvolvido um mapa de risco derivado do modelo, que pode ajudar os tomadores de decisão a localizar regiões com maior risco de testes positivos, permitindo melhores políticas de testagem e controle de doenças

# Inscrições abertas para Mestrado e Doutorado na PUC-Rio

## PROCESSO SELETIVO 2023.1

DEI  
DEPARTAMENTO  
DE ENGENHARIA  
INDUSTRIAL



MESTRADO & DOUTORADO  
ENG. DE PRODUÇÃO

INSCRIÇÕES

ABERTAS!

ATÉ

13/11



Mais informações em:  
<http://www.ind.puc-rio.br/ensino/pos-graduacao-academica/inicio/>

Para nos conhecer melhor acesse:  
<https://vimeo.com/537856245>

{OPERAÇÕES E NEGÓCIOS}  
EM ENGENHARIA

{PESQUISA OPERACIONAL}