

COVID-19: Influência de exames na precisão e recall de modelos preditivos

Jairo Freitas
Christian Espinoza



Autores



Jairo da Silva Freitas Júnior



Analista de Dados



Ciência da Computação

Christian Espinoza



Estagiário em Ciência de Dados



Ciência da Computação

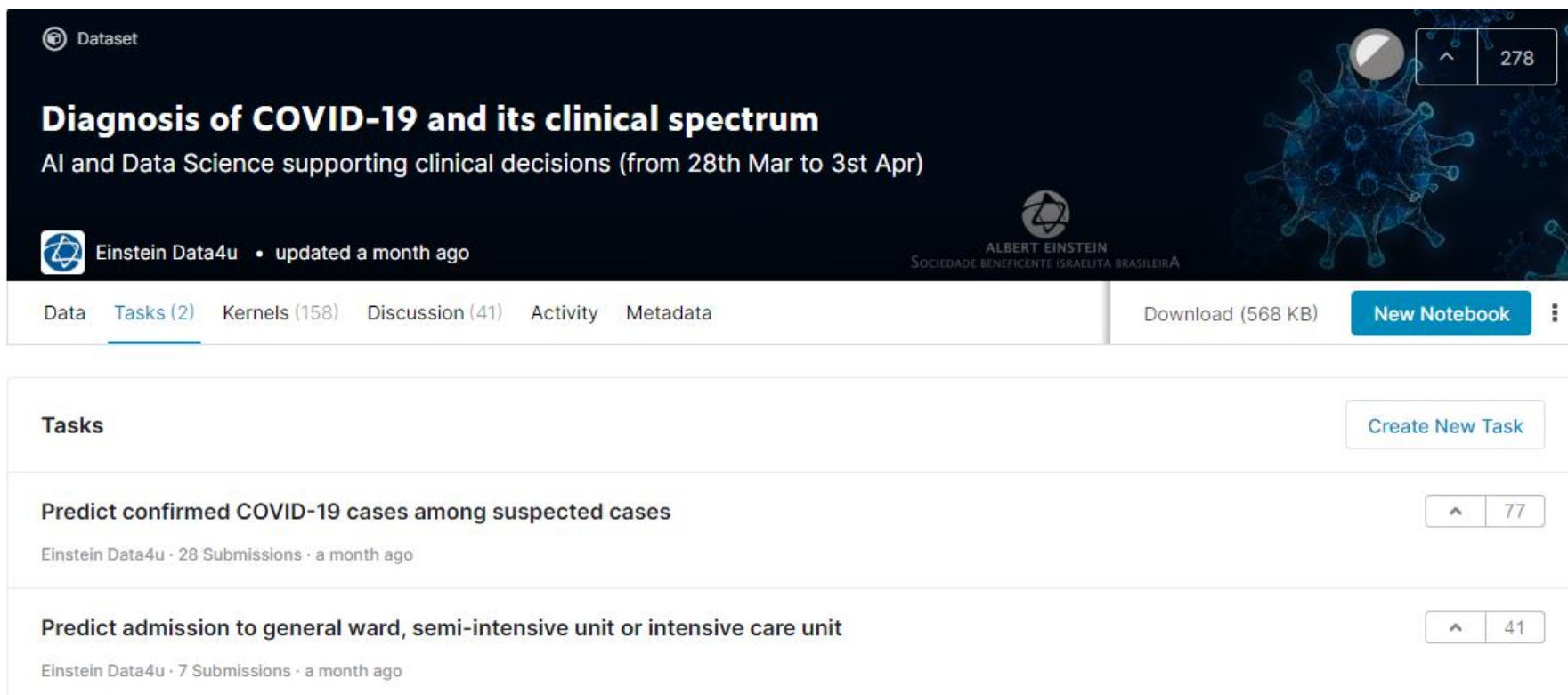


Sumário

- **O desafio e os dados disponibilizados**
- **Análise Exploratória dos Dados**
- **Resumo de todos os modelos testados**
- **Modelo 1 (“Fique em Casa”): melhor recall**
- **Modelo 2 (“Business As Usual”): melhor precisão**
- **Backtest em outras infecções respiratórias**
- **Disclaimers**



O desafio e os dados disponibilizados



The screenshot shows the Kaggle dataset page for "Diagnosis of COVID-19 and its clinical spectrum". The page header includes the dataset title, a subtitle "AI and Data Science supporting clinical decisions (from 28th Mar to 3st Apr)", and the creator "Einstein Data4u" with a note "updated a month ago". The Einstein logo and "SOCIIDADE BENEFICENTE ISRAELITA BRASILEIRA" are also visible. A navigation bar at the top right shows "Download (568 KB)" and a "New Notebook" button. Below the navigation bar, the "Tasks" section is active, displaying two tasks: "Predict confirmed COVID-19 cases among suspected cases" (77 submissions) and "Predict admission to general ward, semi-intensive unit or intensive care unit" (41 submissions). A "Create New Task" button is located in the top right of the tasks section.

Dataset

Diagnosis of COVID-19 and its clinical spectrum

AI and Data Science supporting clinical decisions (from 28th Mar to 3st Apr)

Einstein Data4u • updated a month ago

Download (568 KB) New Notebook

Tasks

Create New Task

Predict confirmed COVID-19 cases among suspected cases 77

Einstein Data4u • 28 Submissions • a month ago

Predict admission to general ward, semi-intensive unit or intensive care unit 41

Einstein Data4u • 7 Submissions • a month ago

O dataset contém dados anonimizados de **5.644 pacientes** que foram atendidos no Hospital Israelita Albert Einstein, em São Paulo, que tiveram amostras coletadas para **105 testes laboratoriais** durante a visita hospitalar. **88% dos valores do dataset estão faltantes** (missing values).

Duas variáveis resposta foram incluídas: resultado **SARS-CoV-2-RT-PCR** e **ala de admissão hospitalar**. As variáveis clínicas foram padronizadas para média zero e desvio padrão unitário.

<https://www.kaggle.com/einsteindata4u/covid19>

Testes pouco frequentes apresentam vazamento de informação das variáveis resposta pois são realizados em quadros clínicos graves

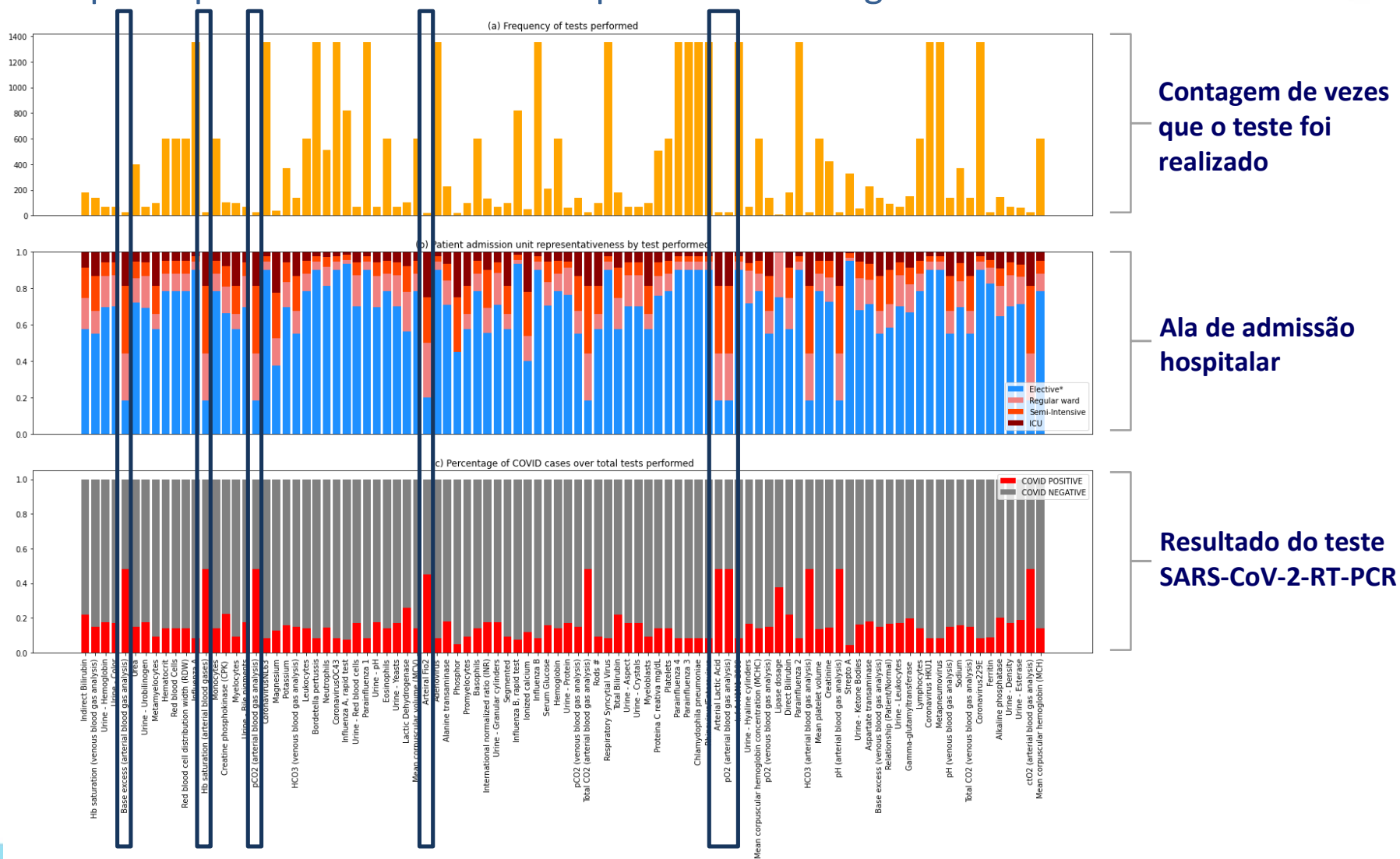


Figura 1 – Análise das variáveis em relação a frequência, ala de admissão hospital e SARS-CoV-2-RT-PCR

Alguns exames laboratoriais parecem coocorrer na amostra

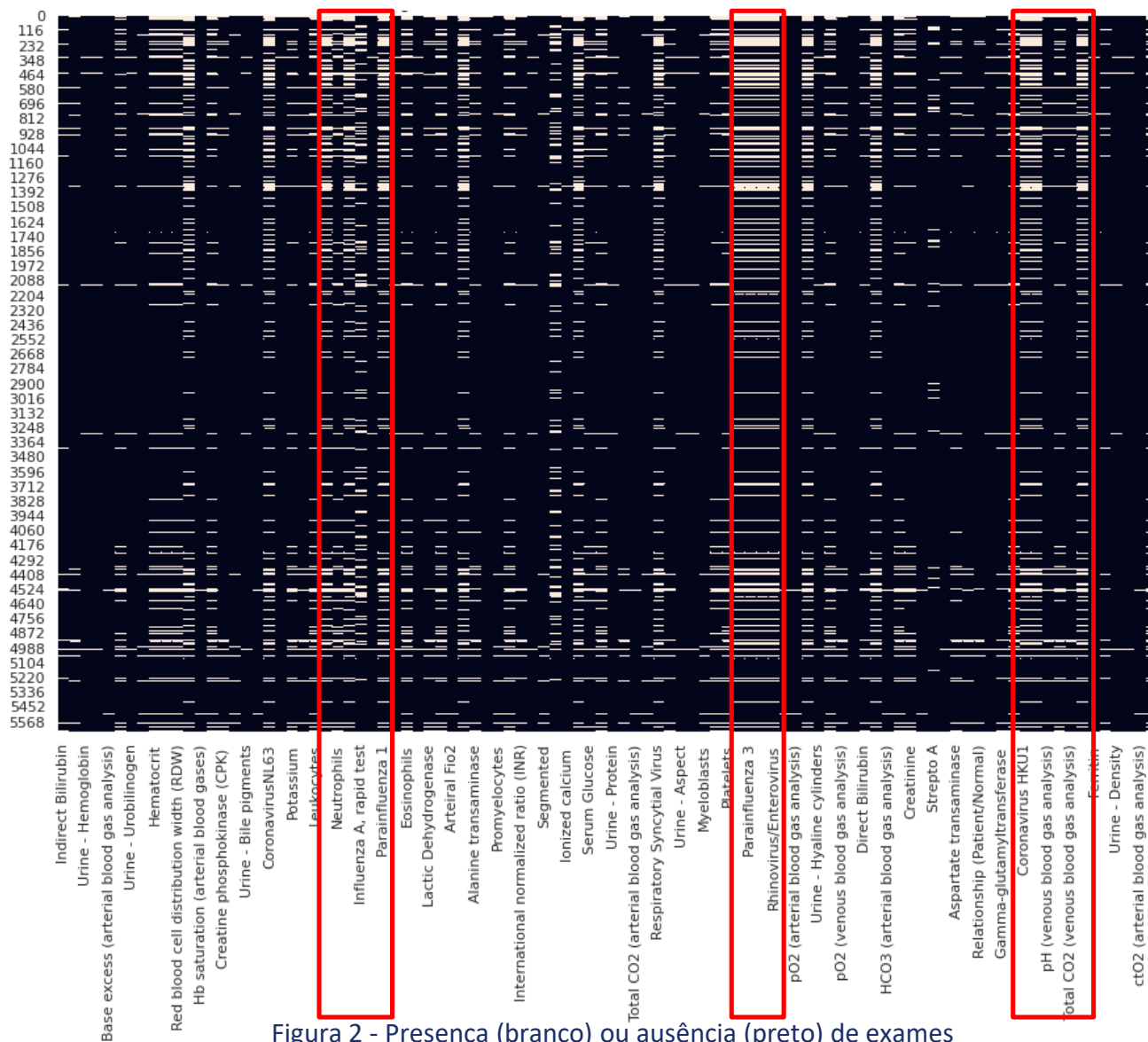


Figura 2 - Presença (branco) ou ausência (preto) de exames
Versão sem ordenação inspirada no kernel de Nasser Boan

Aplicando PCA sobre a presença (ou ausência) de exames descobrimos 6 grandes grupos de exames que costumam coocorrer

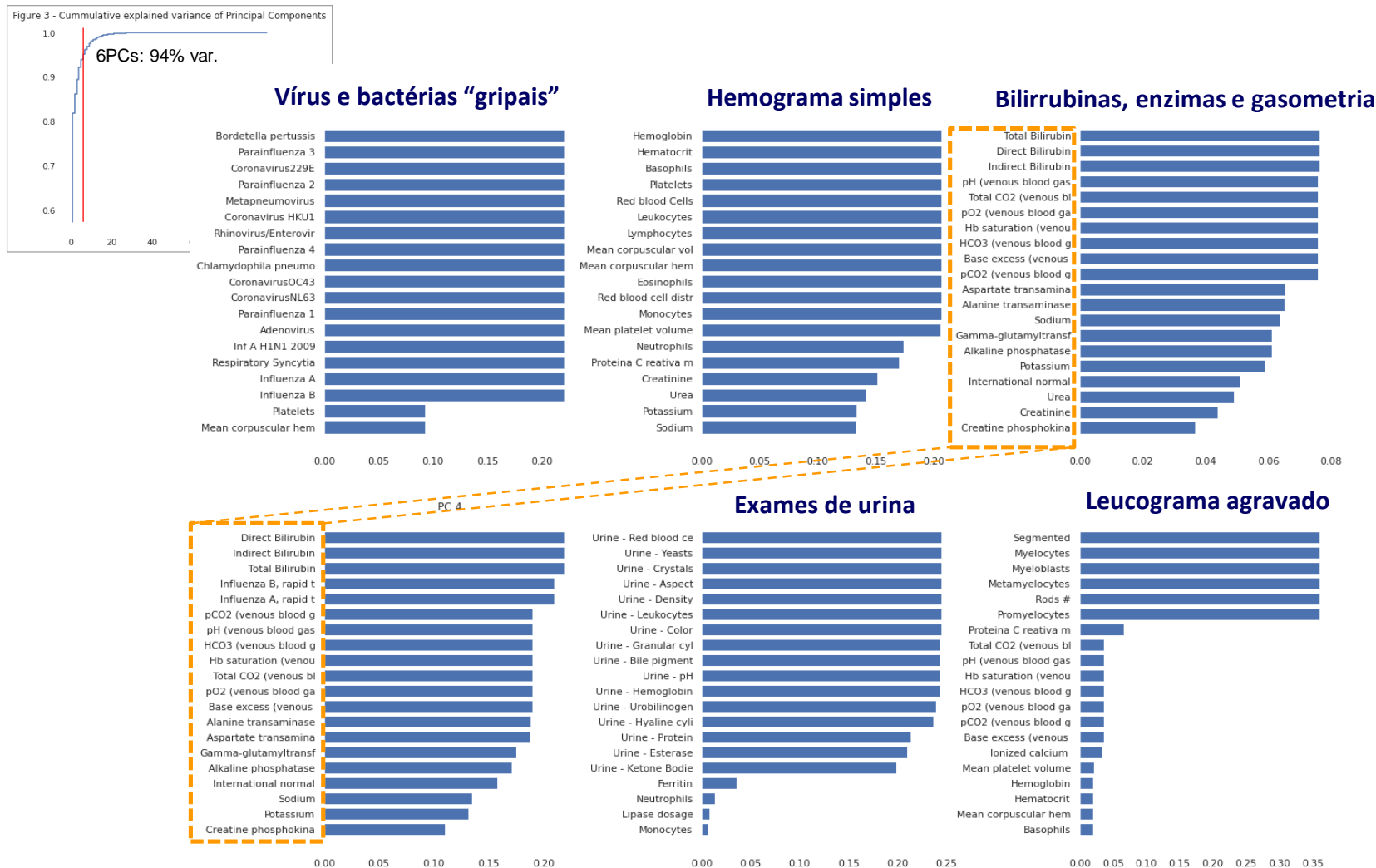


Figura 4 – Composição de 6 componentes principais

Ao reordenarmos o dataset pelos grupos de exames, vemos a coocorrência dos testes. 62.8% dos pacientes realizaram apenas a testagem SARS-CoV-2-RT-PCR

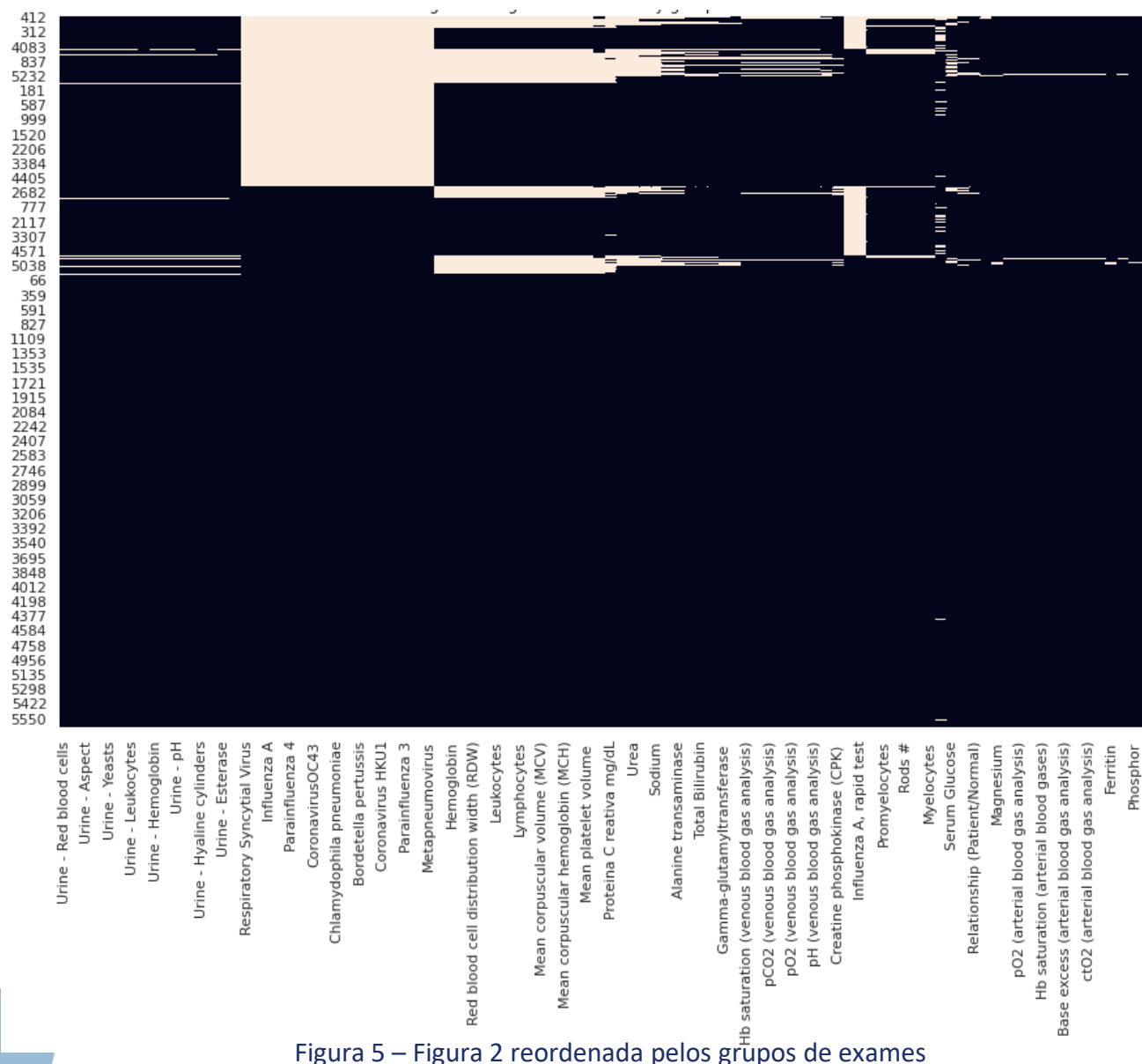


Figura 5 – Figura 2 reordenada pelos grupos de exames

A coocorrência entre os grupos de exames aumenta com a complexidade da ala de admissão do paciente.

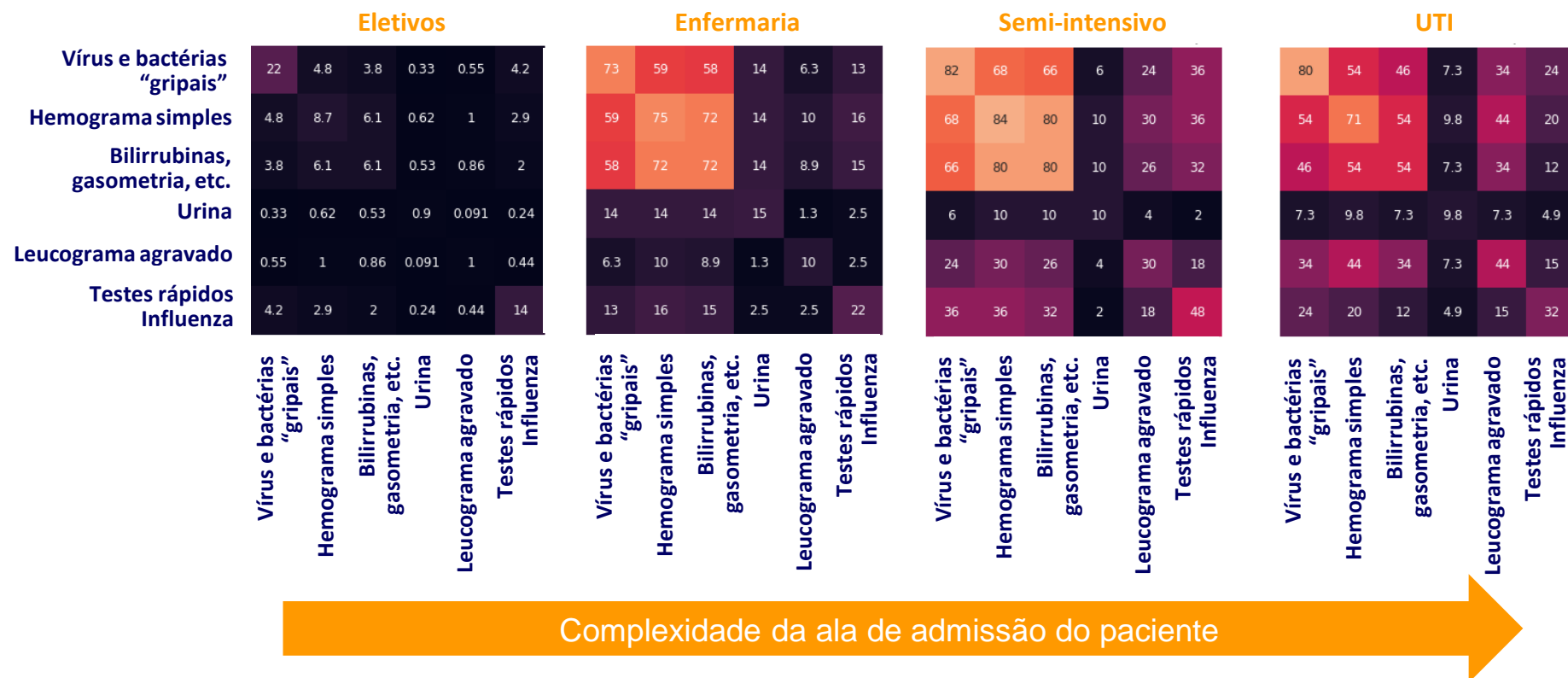


Figura 6 – Coocorrência de grupos de exames por ala de admissão hospitalar

Para reduzir o viés de gravidade do quadro clínico incorporado na realização de mais de um grupo de exames, **nossos modelos foram treinados apenas em variáveis do mesmo grupo de exames.**

Detecção de SARS-CoV-2: Amostras de hemograma simples mostraram-se as mais eficazes na tarefa preditiva

Grupo de exames	Modelo	F1-Score Macro avg	Recall Macro avg	Precisão Macro avg	Acurácia
Hemograma simples	Baseline	46%	50%	43%	87%
	SVM	66%	68%	65%	83%
	Gradient Boost	67%	66%	67%	85%
	Random Forest	67%	68%	66%	83%
	Ada Boost	66%	73%	64%	79%
Vírus e bactérias “gripais”	Baseline	48%	50%	46%	92%
	SVM	48%	50%	46%	92%
	Gradient Boost	48%	50%	46%	92%
	Random Forest	60%	66%	59%	82%
	Ada Boost	51%	74%	58%	63%
Testes rápidos Influenza	Baseline	48%	50%	46%	92%
	SVM	47%	60%	53%	61%
	Gradient Boost	48%	50%	46%	92%
	Random Forest	49%	57%	52%	70%
	Ada Boost	48%	63%	54%	62%
Urina	Amostragem insuficiente				
Leucograma agravado	Amostragem insuficiente				
Bilirrubinas, gasometria, etc.	Amostragem insuficiente				

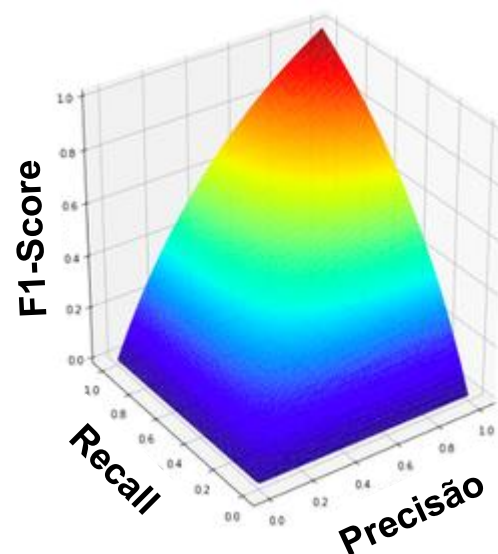
Hiperparâmetros otimizados

C, kernel, peso das classes

#Estimadores, profundidade máxima, taxa de aprendizado

#Estimadores, profundidade máxima, # máximo de variáveis, peso das classes

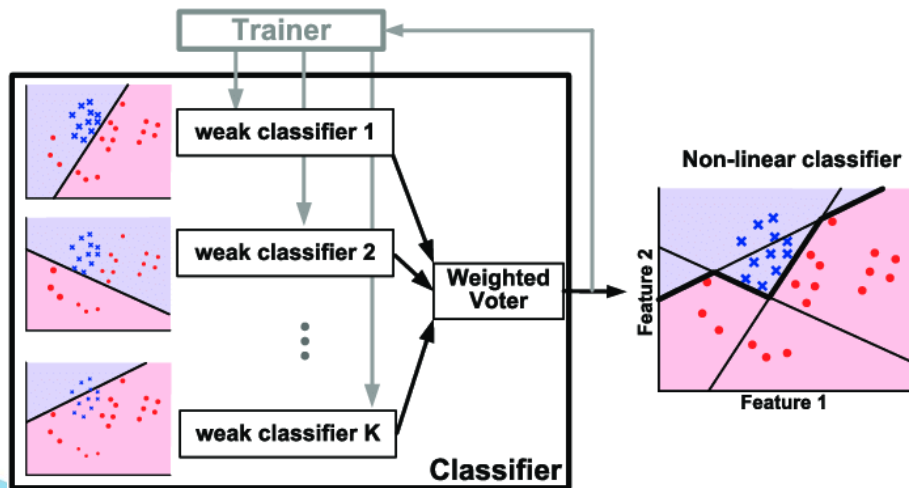
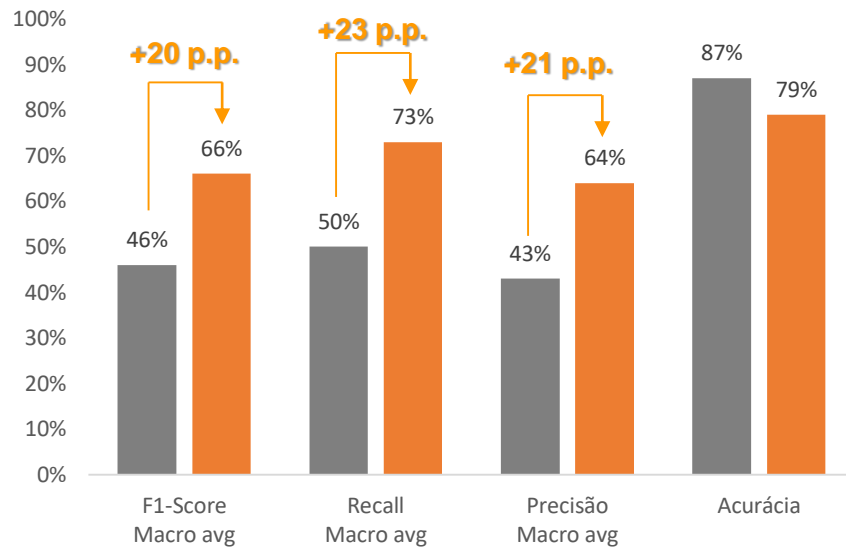
estimadores, profundidade máxima, peso das classes, taxa de aprendizado



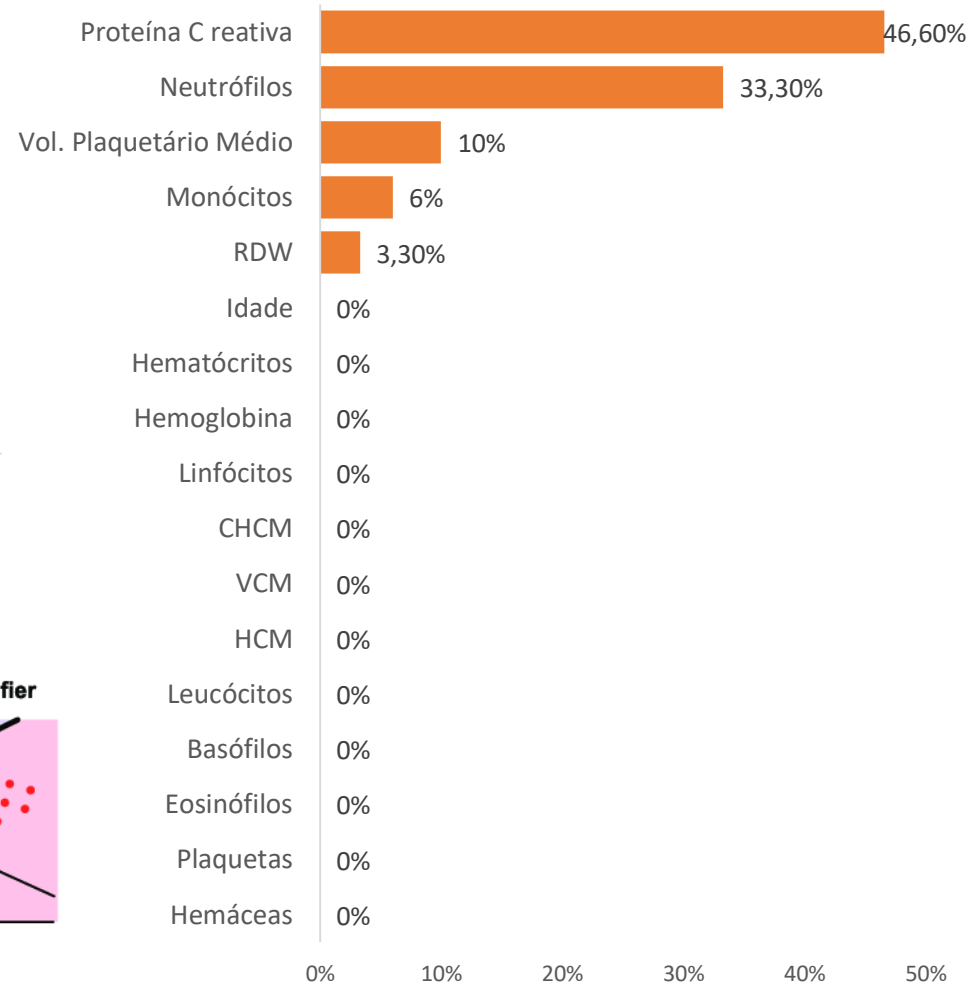
Conjunto de treino e teste estratificado por **SARS-CoV-2-RT-PCR** e **ala de admissão**.

Modelo “Fique em Casa” (Ada Boost)

Indicadores de performance

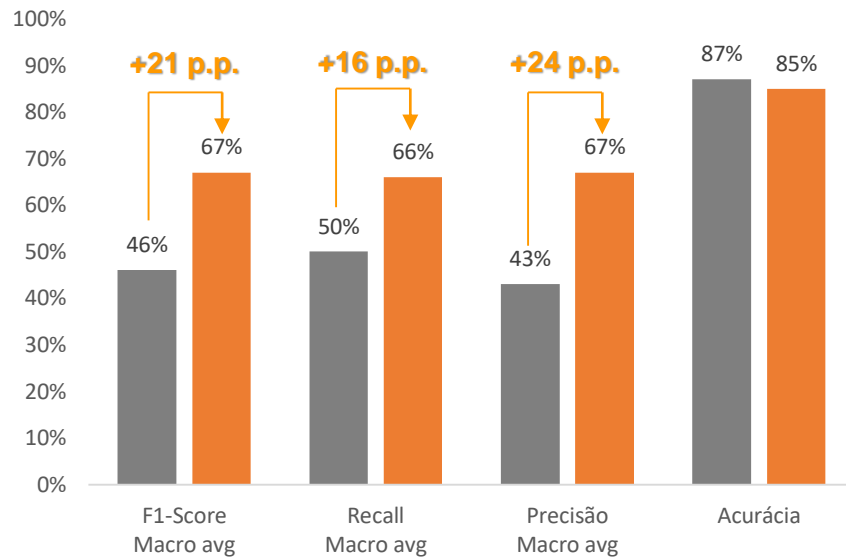


Importância das variáveis

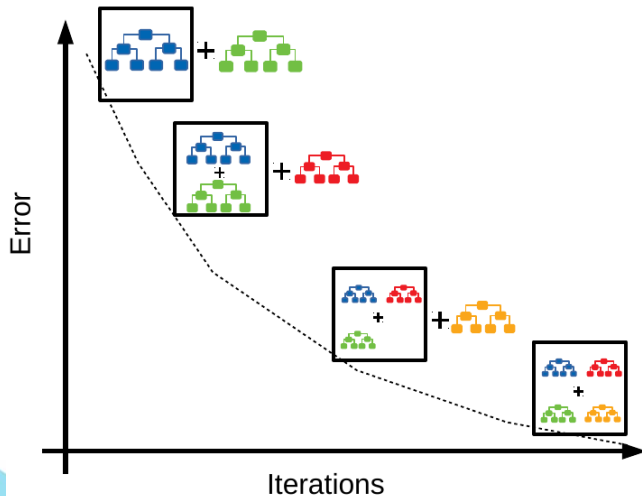
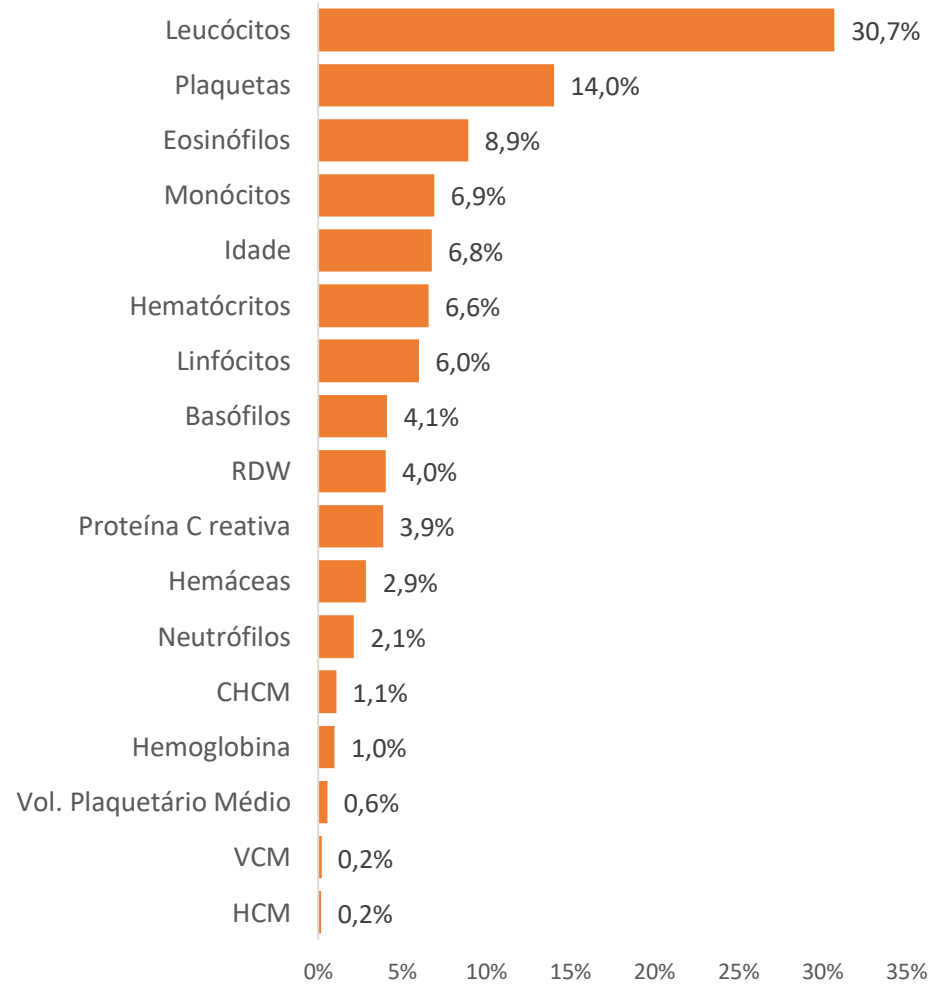


Modelo “Business As Usual” (Gradient Boosting)

Indicadores de performance



Importância das variáveis



Nossa interpretação do modelo (hipóteses)

O Gradient Boosting possui precisão maior pois dá mais importância para sinais de **severidade da infecção**, especialmente leucócitos.

O AdaBoost favorece o recall porque é mais sensível a **componentes do hemograma que apresentam-se precocemente nos processos infecciosos**, especialmente a Proteína C Reativa.

Figura 7 – Cinética dos processos infecciosos com relação a glóbulos brancos

(a) General [6]

(b) C Reactive Protein [5]

Sequence of Events - Infection

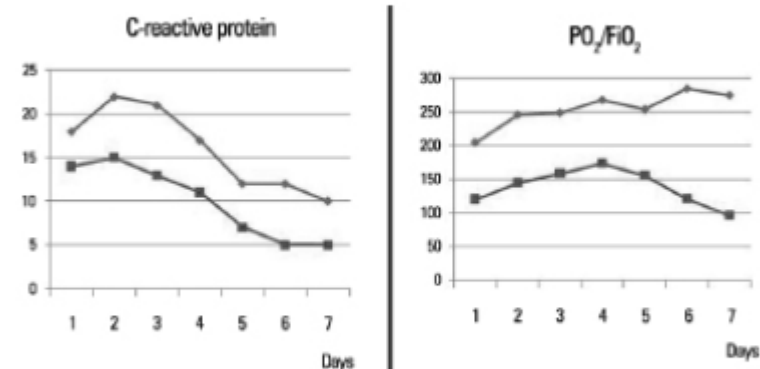
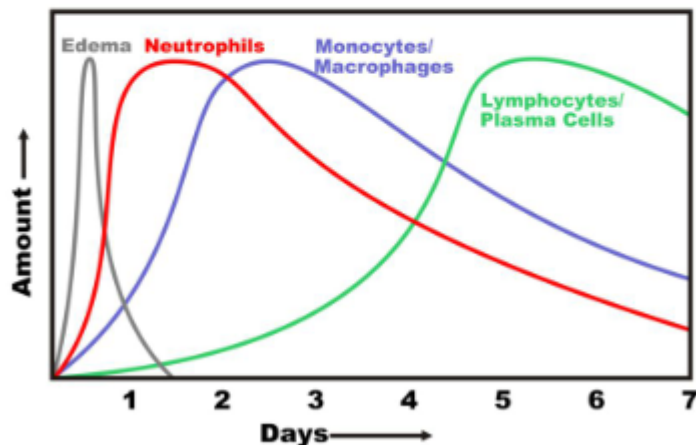


Figure 1 - Mean values of C-reactive protein (mg/dL) and arterial oxygen pressure/fraction of inspired oxygen (PO_2/FiO_2) during the first week of evolution for the H1N1 (■) and community-acquired pneumonia (◆) groups.

[5] Nardocci Paula, Gullo Caio Eduardo, Lobo Suzana Margareth. Severe virus influenza A H1N1 related pneumonia and community-acquired pneumonia: differences in the evolution. Rev. bras. ter. intensiva [Internet]. 2013 June; 25(2): 123-129. Available from: <https://bit.ly/2W8Qe2M>

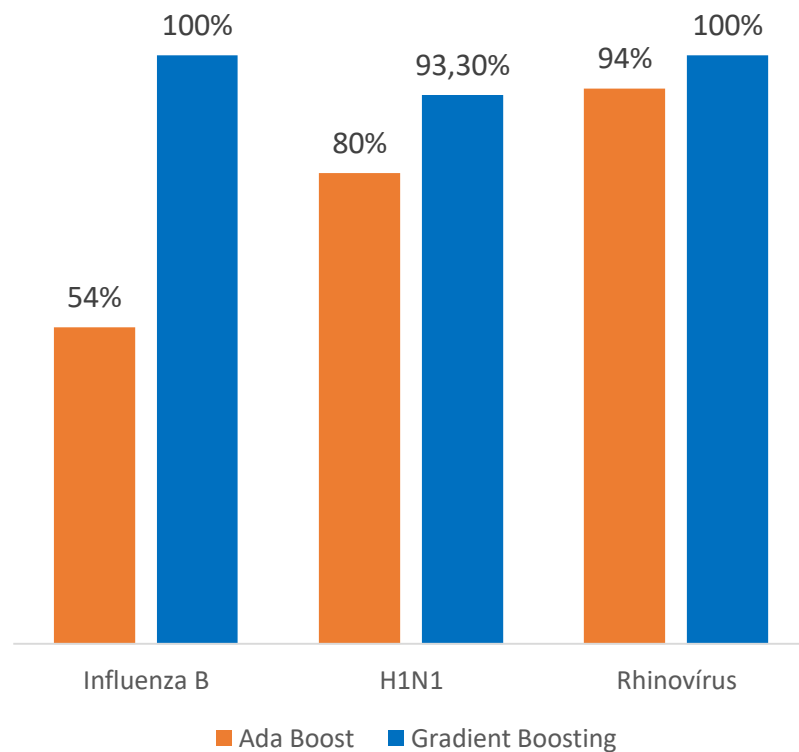
[6] J. Matthew Velkey. Cell Injury, Death, Inflammation, and Repair. Lecture notes. Duke University. Available at: <https://slideplayer.com/slide/4382692/>

Backtest em outras infecções respiratórias: Nossos modelos diferenciaram SARS-CoV-2 de Influenza B, H1N1 e Rhinovírus

Figura 8 – Pacientes infectados com outras doenças respiratórias (SARS-CoV-2 Neg)

Influenza B	24	0	0	1	0	1	0	0	0	0	0	0	5	0	0	0	0
Respiratory Syncytial Virus	0	11	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
Influenza A	0	0	4	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Metapneumovirus	1	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
Parainfluenza 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Inf A H1N1 2009	1	0	1	0	0	21	0	0	0	0	0	0	0	0	1	0	0
Bordetella pertussis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Chlamydia pneumoniae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coronavirus229E	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
Parainfluenza 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Parainfluenza 3	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0
CoronavirusNL63	0	0	0	0	0	0	0	0	0	0	1	11	0	0	0	0	0
Rhinovirus/Enterovirus	5	1	0	0	0	0	0	0	0	0	0	0	98	1	0	1	0
CoronavirusOC43	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0
Coronavirus HKU1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	5	0	0
Adenovirus	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
Parainfluenza 4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	Influenza B	Respiratory Syncytial Virus	Influenza A	Metapneumovirus	Parainfluenza 1	Inf A H1N1 2009	Bordetella pertussis	Chlamydia pneumoniae	Coronavirus229E	Parainfluenza 2	Parainfluenza 3	CoronavirusNL63	Rhinovirus/Enterovirus	CoronavirusOC43	Coronavirus HKU1	Adenovirus	Parainfluenza 4

Acurácia dos modelos em amostras de Influenza B, H1N1 e Rhinovírus



Disclaimers

- Os modelos não levam em consideração **aspectos demográficos** como gênero e etnia, assim como **comorbidades dos pacientes**.
- A **amostra disponível pode não ser representativa da população brasileira**, principalmente devido às variações de saúde decorrentes de particularidades regionais e desigualdades socioeconômicas.
- Os modelos apresentados não foram criticados por especialistas médicos.
- Os modelos não foram testados para **variações de protocolos de coleta e processamento de amostras** que podem apresentar-se ao escalá-lo para nível nacional.

