Analisando parâmetros multidimensionais: como direcionar o foco na vacinação prioritária de populações-chave?

Projeto final - Disciplina: Ciência e Visualização de Dados em Saúde

Andreza Aparecida dos Santos (164213)
Leonardo Marçal (225240)
Lígia Vasconcellos (081938)
Mariana Amaral Raposo (262866)

Introdução

Coronavírus 19 (COVID-19)

- Rápida disseminação
- Crescente de casos e óbitos
- Colapso sanitário, hospitalar e econômico

COVID-19 no Brasil

- Março de 2021 - 66 mil óbitos
- Saturação dos sistemas de saúde

Plano Nacional de Vacinação contra a COVID-19 (Ministério da Saúde, 2021)

- Grupos prioritários: idosos, povos indígenas e profissionais de saúde
- Brasil: Dimensão continental X Desafios na produção de vacina

Justificativa e pergunta de pesquisa

Justificativa

Necessidade de medidas efetivas de controle da doença como forma de evitar a incidência de novos picos.

A vacinação em grupos e regiões prioritárias mais afetadas pela crise sanitária poderia auxiliar na redução da mortalidade e recuperação do sistema de saúde.

Pergunta de pesquisa

De acordo com parâmetros multidimensionais correlacionados ao COVID-19, quais regiões e públicos-alvo deveriam ser priorizados na campanha de vacinação visando minimizar o efeito da crise sanitária e econômica?

Objetivo

Analisar parâmetros multidimensionais relacionados ao COVID-19 em cada região do Brasil, buscando por relações nos dados que possam ser capazes de fornecer uma melhor análise estatística das regiões e levantar possíveis planos de vacinação que poderiam beneficiar, de maneira mais ágil, o controle da pandemia de COVID-19 no Brasil.

Metodologia

Problemática envolvida:

Compreender a potencial influência da vacinação aplicada prioritariamente a perfis em condições mais propensas a mortalidade, considerando não apenas como critério a idade do indivíduo a ser imunizado.

Modelo:

KDD – Knowledge Discovery in Databases.

O que foi feito?

Análise de dados estatística exploratória correlacionando todas as variáveis encontradas, provenientes de múltiplas origens de bases de dados.

Identificação de relações de causa e efeito e características do meio em que o indivíduo está inserido.

Ferramentas Utilizadas

As ferramentas utilizadas para exploração e manipulação dos dados foram, **python** e algumas bibliotecas consagradas para machine learning e análise de dados: **Sklearn, Tensorflow, Pandas** e etc.

Como insumo, **utilizamos múltiplas fontes públicas de informações** sobre dados de covid e informações sociodemográficas dos brasileiros.

Aprofundamento da Análise

Estudamos o ponto mais específico da análise:

Quais são os perfis de indivíduos que deveriam ser priorizados na vacinação?

- Construímos uma base cuja unidade de análise foram os indivíduos contaminados ou com suspeita de covid e todas as suas características específicas. Nesta base foi atribuído um target: morte ou não devido ao covid.

- Aplicamos múltiplas técnicas de algoritmos supervisionados, tais como regressão logística, random forest e árvore de decisão.

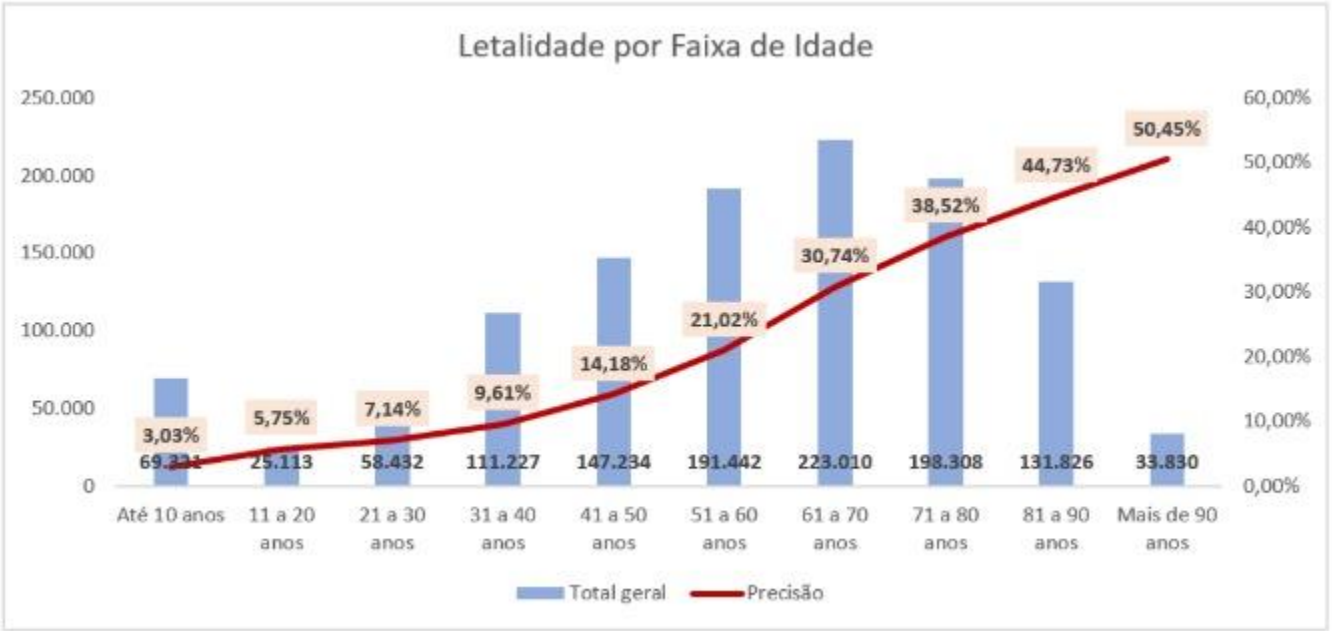
Exploração dos Dados

Após todo o trabalho de centralização, obtivemos um banco de dados de 1.189.743 linhas e 194 colunas. Foi realizada uma primeira etapa de retirada de variáveis redundantes.

Vale destacar que a variável **Evolução** está presente na base Covid Saude Gov, nossa base principal, as quantidades de cada categoria presente na base de dados e suas proporções em relação ao total estão mostradas na figura a seguir.

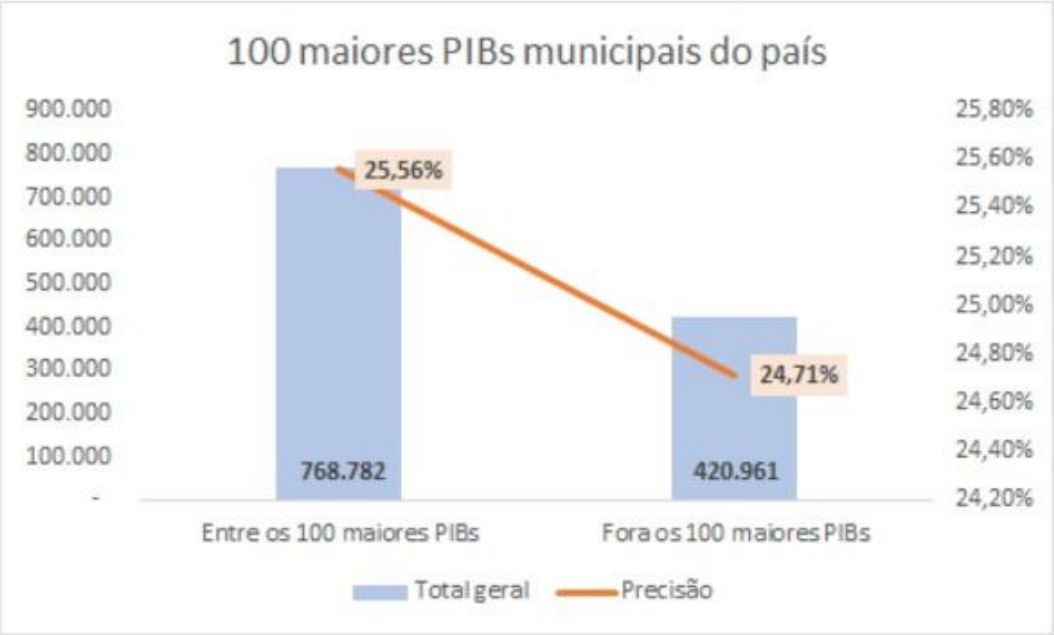
Código	Legenda	Qtd	Proporção
1	Cura	696.675	66,85%
2	Óbito	300.504	28,84%
3	Óbito por outras	13.939	1,34%
9	Ignorado	31.028	2,98%

A variável Idade é a mais relevante nos estudos, sendo o principal critério para a vacinação atual.
De fato, foi a variável com maior correlação com a resposta. À medida que a idade aumenta, a proporção de mortes por faixa de idade também aumenta consideravelmente, podendo ser analisado no gráfico abaixo.



Também foram analisados parâmetros entre comorbidades e fatores de doenças x os parâmetros econômicos

Esperava-se uma relação inversa na variável de PIB, imaginando que os municípios mais ricos conseguiriam letalidades menores devido aos recursos econômicos superiores aos demais, podendo-se observar no gráfico abaixo.



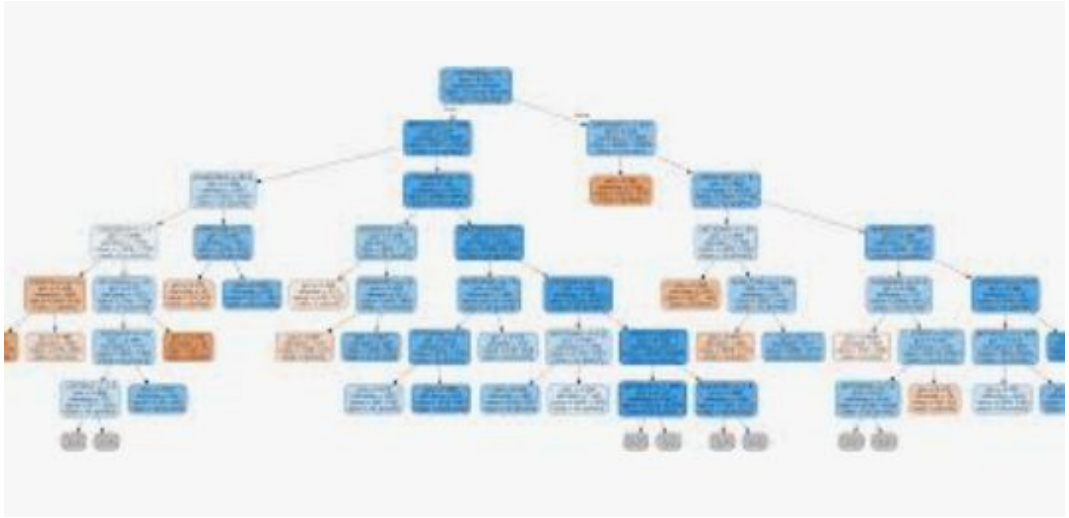
Municipios com 100 maiores PIBs Per Capta do pais	Não Óbito por COVID	Obitos por COVID	Total geral	Precisão
1	39.955	11.864	51.819	22,90%
2	849.284	288.640	1.137.924	25,37%

Análise da taxa de letalidade nos 100 maiores PIBs per capita do país

ETL e Treinamento de Modelo

- Tratamento de outliers, tratamento de valores nulos e dummyzação;
- Centralização das múltiplas bases analisadas, entendimento de chaves de cruzamento e convergência de períodos de análise;
- Estudo de estratégias para lidar com alta dimensionalidade e tamanho da base - dificuldade e custos de processamento.

A **árvore de Decisão** é um tipo de algoritmo de aprendizagem de máquina supervisionado que se baseia na ideia de divisão dos dados em grupos homogêneos, podendo ser utilizadas em um cenário de classificação ou regressão.



O algoritmo **Random Forest** cria várias árvores de decisão e as combina para obter uma predição com maior acurácia e mais estável

A **Regressão Logística** nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias.

É uma técnica recomendada para situações em que a variável dependente é de natureza dicotômica ou binária.

Resultados Obtidos

Variável	Grau de Importância
SURTO_SG	0.07498821332650338
Presença de Fator de Risco	0.04093391693710 38
61 a 70 anos	0.032335631452240 34
71 a 80 anos	0.03105370642139 66
0 a 10 anos	0.028033462014314944
51 a 60 anos	0.027085313892249264
Escolaridade “Não se aplica”	0.023477948109873883
Raça “Ignorada”	0.022186034121134034
Raça “Amarela”	0.021186836896406245
Raça Não Informada	0.020768343827725345
Sexo Feminino	0.018377662103944096
Escolaridade: “Analfabeto”	0.01623526529673434
Cardiopatia Não Informada	0.016111544157652608
medicos_100K_sus	0.016001635674307853
medicos_100K	0.015928250389449745

Peso das Variáveis

Algoritmo	Acurácia	Especificidade	Sensibilidade
Árvore de Decisão	62,75%	63,67%	66,15%
Random Forest	66,79%	66,27%	65,21%
Otimização de Hiperparametros do RF	65,68%	62,63%	70,69%
Regressão Logística	55,67%	57,26%	54,65%

Comparativo de Resultados

Árvore de Regressão

		Valor Predito	
		cura	óbito
Valor Real	cura	59.640	30.524
	óbito	36.643	53.496

Random Forest

		Valor Predito	
		cura	óbito
Valor Real	cura	58.800	31.364
	óbito	28.517	61.622

Regressão Logística

		Valor Predito	
		cura	óbito
Valor Real	cura	60.135	30.029
	óbito	49.905	40.234

RF com Otimização

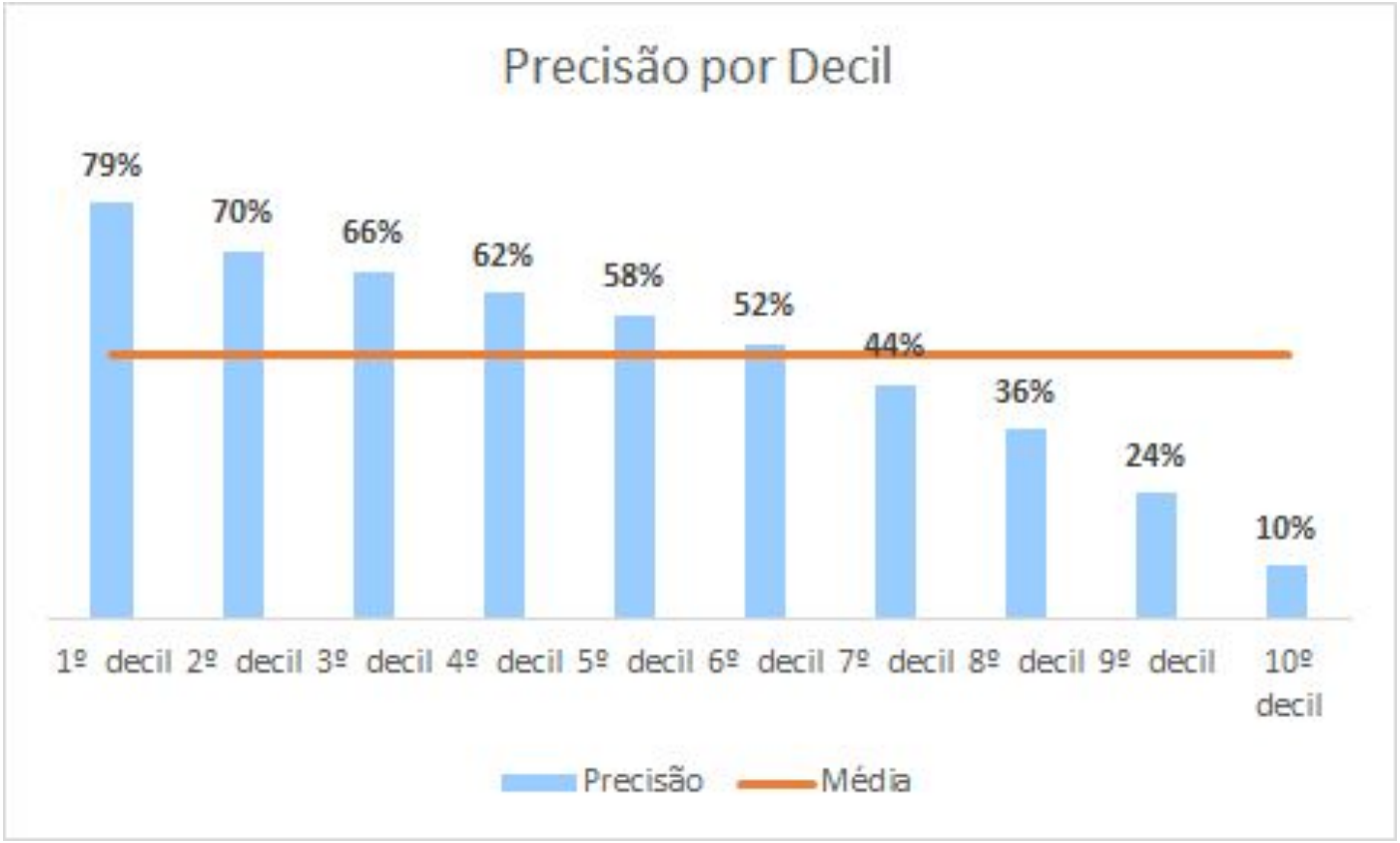
		Valor Predito	
		cura	óbito
Valor Real	cura	48.331	41.833
	óbito	20.038	70.101

Comparativo de Resultados - Matrizes de Confusão

Decis	Soma Não Mortes	Soma Mortes	Total	Precisão	Acumulado Bons	Acumulado Mals	% Acumulado Mals
1º decil	3.733	14.272	18.005	79%	3.733	14.272	16%
2º decil	5.407	12.589	17.996	70%	9.140	26.861	30%
3º decil	5.831	11.456	17.287	66%	14.970	38.317	43%
4º decil	7.160	11.661	18.821	62%	22.130	49.978	55%
5º decil	7.643	10.422	18.065	58%	29.774	60.399	67%
6º decil	8.540	9.390	17.930	52%	38.313	69.790	77%
7º decil	9.684	7.741	17.425	44%	47.997	77.531	86%
8º decil	11.994	6.728	18.722	36%	59.991	84.259	93%
9º decil	12.669	3.968	16.637	24%	72.660	88.227	98%
10º decil	17.491	1.925	19.416	10%	90.151	90.151	100%
Total Geral	90.151	90.151	180.302	50%			

Apesar de as métricas de resultado obtidas terem sido relativamente baixas, ao testarmos a ordenação da base por decil em uma amostra 50%/50%, obtivemos uma relevante diferenciação da letalidade por faixa.

Diferença de Precisão por decil da base ordenada



Diferença de Precisão por decil da base ordenada

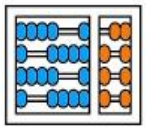
Projeto final - Disciplina: Ciência e Visualização de Dados em Saúde

Andreza Aparecida dos Santos

Leonardo Marçal

Lígia Vasconcellos

Mariana Amaral Raposo



Instituto de Computação

UNIVERSIDADE ESTADUAL DE CAMPINAS