Universidade Federal do Rio de Janeiro

Disciplina: Tópicos Especiais Sist. Informação I

Alunos: ALESSANDRA PIERRO SARAIVA, FABIANE MONTEIRO CARVALHO, MASSAO OLIVEIRA IWANAGA

PROJETO DO TRABALHO DA DISCIPLINA TÓP. ESP. SIST. INF. I

1. Detalhamento e descrição do problema/questão de análise

A humanidade vem enfrentando uma pandemia sem precedentes através da doença contagiosa chamada COVID19 (Coronavirus disease 2019). No contexto brasileiro, com o início e avanço, no ano de 2021, da vacinação contra essa enfermidade, fazem-se necessárias análises para avaliar o perfil dos vacinados nas diferentes regiões do país devido às suas desigualdades. A região norte, principalmente, com aproximadamente 18,7 milhões de habitantes, foi muito afetada nos últimos meses por nova variante do coronavírus e também a região sudeste, por ser a mais populosa, com sua população em torno de 89 milhões de habitantes [3]. Este trabalho terá como foco essas duas regiões do Brasil.

Verifica-se que essas análises poderão contribuir para um estudo mais detalhado sobre a cobertura da vacinação nessas regiões com seus quantitativos e perfis, permitindo tomadas de decisões através dos dados apresentados para aprimorar o processo de vacinação nesses locais, com informações sobre os estabelecimentos de saúde, de forma a diminuir a aglomeração de pessoas no atendimento e também comparar a distribuição desses estabelecimentos nas duas regiões do país.

2. Descrição do dataset e sua utilização

Neste trabalho, serão usados os dois datasets listados a seguir, que podem ser obtidos no Portal Brasileiro de Dados Abertos, ferramenta disponibilizada pelo governo para que todos possam encontrar e utilizar os dados abertos e informações públicas[1].

2.1. Dataset da Campanha Nacional de Vacinação contra Covid-19 [2]

Este dataset possui informações sobre os vacinados (idade, data de nascimento, sexo, raça e etc.), os estabelecimentos de saúde (nome, município, estado e etc.) onde ocorreu a vacinação, grupo de atendimento ao qual o vacinado pertence e da vacina (fabricante, lote e etc.), além de outras.

Esse dataset contém um dicionário de dados, conforme exibido a seguir, contendo os 32 campos e suas descrições.

Ordem	Campo	Descrição	Categoria
1	document_id	Identificador do documento	
2	paciente_id	Identificador do vacinado	
3	paciente_idade	Idade do vacinado	
4	paciente_dataNascimento	Data de nascimento do vacinado	
5	paciente_enumSexoBiologico	Sexo do vacinado	M = Masculino, F = Feminino
6	paciente_racaCor_codigo	Código da raça/cor do vacinado	1; 2; 3; 4; 99
7	paciente_racaCor_valor	Descrição da raça/cor do vacinado	1 = Branca; 2 = Preta; 3 = Parda; 4 = Amarela; 99 = Sem informação
8	paciente_endereco_coIbgeMunicipio	Código IBGE do município de endereço do vacinado	
9	paciente_endereco_coPais	Código do país de endereço do vacinado	
10	paciente_endereco_nmMunicipio	Nome do município de endereço do vacinado	
11	paciente_endereco_nmPais	Nome do país de endereço do vacinado	
12	paciente_endereco_uf	Sigla da UF de endereço do vacinado	
13	paciente_endereco_cep	5 dígitos para anonimizado e 7 dígitos para identificado	
14	paciente_nacionalidade_enumNacionalidade	Nacionalidade do vacinado	
15	estabelecimento_valor	Código do CNES do estabelecimento que realizou a vacinação	
16	estabelecimento_razaoSocial	Nome/Razão Social do estabelecimento	
17	estabelecimento_noFantasia	Nome fantasia do estabelecimento	
18	estabelecimento_municipio_codigo	Código do município do estabelecimento	
19	estabelecimento_municipio_nome	Nome do município do estabelecimento	
20	estabelecimento_uf	Sigla da UF do estabelecimento	
21	vacina_grupo_atendimento_code	Código do grupo de atendimento ao qual pertence o vacinado	
22	vacina_grupo_atendimento_nome	Nome do grupo de atendimento ao qual pertence o vacinado	
23	vacina_categoria_code	Código da categoria	
24	vacina_categoria_nome	Descrição da categoria	
25	vacina_lote	Número do lote da vacina	
26	vacina_fabricante_nome	Nome do fabricante/fornecedor	
27	vacina_fabricante_referencia	CNPJ do fabricante/fornecedor	
28	vacina_dataAplicacao	Data de aplicação da vacina	
29	vacina_descricao_dose	Descrição da dose	
30	vacina_codigo	Código da vacina	
31	vacina_nome	Nome da vacina/produto	
32	sistema origem	Nome do sistema de origem	

Figura 1 - Dicionário de dados do dataset da Campanha Nacional de Vacinação contra Covid-19

[1] https://dados.gov.br/

Os dados em si, dão um total de 8.958.579 registros, incluído o cabeçalho, e 3.6 GB de tamanho físico, disponibilizados em formato CSV (*Comma-separeted values*).

2.2. Cadastro Nacional de Estabelecimentos de Saúde (CNES) - Ativo [1]

Este dataset possui informações sobre todos os estabelecimentos (código, nome, endereço, tipo e etc.) de saúde ativos no país, públicos ou não. Apesar de não haver disponível um dicionário de dados, criamos um, conforme exibido a seguir, a partir de informações contidas na página do dataset e no cabeçalho do arquivo.

Ordem	Campo	Descrição
1	co_cnes	Código CNES
2	co_ibge	Código IBGE
3	no_fantasia	Nome Fantasia
4	ds_tipo_unidade	Tipo de estabelecimento de saúde: Posto de Saúde, Centro de Saúde/Unidade Básica de Saúde, Policlínica, Hospital Geral,Hospital Especializado, Unidade Mista, Pronto Socorro Geral, Pronto Socorro Especializado, Consultório Isolado, UnidadeMóvel Fluvial, Clínica Especializada/Amb. Especializado, Unidade de Serviço de Apoio de Diagnose e Terapia, Unidade MóvelTerrestre, Unidade Móvel de Nível Pré-hospitalar na Área de Urgência e Emergência: Farmácia: Unidade de Vigilância em Saúde,Cooperativa, Centro de Parto Normal Isolado,Hospital /Dia-Isolado, Central de Regulação de Serviços de Saúde, LaboratórioCentral de Saúde Pública – LACEN, Secretaria de Saúde
5	tp_gestao	Tipo de gestão: Estadual (E), Municipal (M) ou Dupla (D)
6	no_logradouro	Nome do Logradouro
7	nu_endereco	Número do Endereço
8	no_bairro	Bairro
9	co_cep	CEP
10	uf	UF do estabelecimento
11	municipio	Município
12	nu_telefone	Número de Telefone

Figura 2 - Dicionário de dados do dataset do Cadastro Nacional de Estabelecimentos de Saúde (CNES) - Ativo

Os dados em si, dão um total de 354.805 registros, incluído o cabeçalho, e 56MB de tamanho físico, disponibilizados em formato CSV.

3. Objetivos

3.1. Objetivos gerais

Aplicação de técnicas de *Data Science (DS)* que subsidiem análises do problema, tais como:

- Importação, limpeza e tratamento de dados dos datasets;
- Análise exploratória dos datasets;
- Extração do conhecimento dos datasets; e
- Visualização de dados que auxiliem na análise do problema proposto.

Esse trabalho será implementado na linguagem Python e deverá prever a reutilização do código, bem como informações de proveniência sobre o ambiente computacional e os dados utilizados.

3.2. Objetivos específicos

Fornecer informações sobre os vacinados e também dos seus locais de vacinação por município e estados das regiões norte e sudeste.

Serão apresentados como resultados:

1 - O perfil dos vacinados (sexo, idade, raça, grupo) e quantitativos por estado nas duas regiões;

2 - O perfil dos vacinados (sexo, idade, raça, grupo) e quantitativos por tipo de estabelecimento de saúde por estado nas duas regiões;

3 - O quantitativo dos estabelecimentos de saúde e seus tipos nos estados nas duas regiões por total de população;

4- O percentual de vacinados que tomaram a vacina em município diferente do da sua residência mas no mesmo estado;

5- Visualização de % da população vacinada ao longo do tempo, nas duas regiões, permitindo uma análise comparativa da evolução da vacinação entre as regiões.

Os sete estados da região norte e suas respectivas siglas estão listadas a seguir:

Amazonas (AM)	Acre (AC)	Amapá (AP)
Pará (PA)	Rondônia (RO)	Roraima (RR)

Tocantins (TO)		

Os quatro estados da região sudeste e suas respectivas siglas estão listadas a seguir:

Espírito Santo (ES)	Minas Gerais (MG)
Rio de Janeiro (RJ)	São Paulo (SP)

4. Métodos de data cleaning / tratamento de dados

Este trabalho será implementado através do Google Colab. O Google Colaboratory, ou simplesmente, Colab, é um serviço de nuvem gratuito hospedado pela Google para, dentre outros fins, escrever código Python no navegador sem nenhuma configuração local necessária e com acesso gratuito a GPUs (Unidades de Processamento Gráfico)[2]. Os datasets serão baixados dos repositórios, conforme indicado no capítulo 2, e em seguida serão colocados no Google Drive, de onde serão lidos pelo *notebook* do Colab.

A biblioteca Pandas, através do comando *read_csv*, será a responsável por importar os arquivos, pois ambos os datasets estão no formato .CSV. Pandas é uma biblioteca de software escrita para a linguagem de programação Python para manipulação e análise de dados. Em particular, ele oferece estruturas de dados e operações para manipular tabelas numéricas e séries temporais[3].

Como os dois datasets possuem informações sobre todos os estados brasileiros, visando reduzir a quantidade de dados a serem analisados e trazer eventuais ganhos de performance, será feita uma seleção inicial nos dataframes para utilizar apenas das informações dos estados das regiões norte e sudeste, que serão analisadas neste trabalho.

Para os campos considerados desnecessários nos arquivos para este trabalho, os mesmos serão eliminados, pois são irrelevantes para atingir os objetivos desta pesquisa. Por exemplo: No dataset dos vacinados, optou-se por excluir os seguintes campos:

[2] https://colab.research.google.com/
[3] https://en.wikipedia.org/wiki/Pandas_(software)

paciente_endereco_cep, vacina_lote, vacina_fabricante_referencia, vacina_codigo e sistema_origem, conforme destacado na Figura 3 a seguir.

Ordem	Descrição	Categoria	Categoria
1	document_id	Identificador do documento	
2	paciente_id	Identificador do vacinado	
3	paciente_idade	Idade do vacinado	
4	paciente_dataNascimento	Data de nascimento do vacinado	
5	paciente_enumSexoBiologico	Sexo do vacinado	M: masculino; F: feminino
6	paciente_racaCor_codigo	Código da raça/cor do vacinado	1; 2; 3; 4; 99
7	paciente_racaCor_valor	Descrição da raça/cor do vacinado	1: branca; 2: preta; 3: parda; 4: amarela; 99: sem informação
8	paciente_endereco_colbgeMunicipio	Código IBGE do município de endereço do vacinado	
9	paciente_endereco_coPais	Código do país de endereço do vacinado	
10	paciente_endereco_nmMunicipio	Nome do município de endereço do vacinado	
11	paciente_endereco_nmPais	Nome do país de endereço do vacinado	
12	paciente_endereco_uf	Sigla da UF de endereço do vacinado	
13	paciente_endereco_cep	5 dígitos para anonimizado e 7 dígitos para identificado	
14	paciente_nacionalidade_enumNacionalidade	Nacionalidade do vacinado	
15	estabelecimento_valor	Código do CNES do estabelecimento que realizou a vacinação	
16	estabelecimento_razaoSocial	Nome/Razão Social do estabelecimento	
17	estalecimento_noFantasia	Nome fantasia do estabelecimento	
18	estabelecimento_municipio_codigo	Código do município do estabelecimento	
19	estabelecimento_municipio_nome	Nome do município do estabelecimento	
20	estabelecimento_uf	Sigla da UF do estabelecimento	
21	vacina_grupoAtendimento_codigo	Código do grupo de atendimento ao qual pertence o vacinado	
22	vacina_grupoAtendimento_nome	Nome do grupo de atendimento ao qual pertence o vacinado	
23	vacina_categoria_codigo	Código da categoria	
24	vacina_categoria_nome	Descrição da categoria	
25	vacina_lote	Número do lote da vacina	
26	vacina_fabricante_nome	Nome do fabricante/fornecedor	
27	vacina_fabricante_referencia	CNPJ do fabricante/fornecedor	
28	vacina_dataAplicacao	Data de aplicação da vacina	
29	vacina_descricao_dose	Descrição da dose	
30	vacina_codigo	Código da vacina	
31	vacina_nome	Nome da vacina/produto	
32	sistema_origem	Nome do sistema de origem	

Figura 3 - Colunas (em vermelho) excluídas do dataset de Vacinação

No dataset CNES, os campos eliminados serão: no_logradouro, nu_endereco, no_bairro, co_cep, nu_telefone, conforme destacado na Figura 4 a seguir.

Ordem	Campo	Descrição
1	co_cnes	Código CNES
2	co_ibge	Código IBGE
3	no_fantasia	Nome Fantasia
4	ds_tipo_unidade	Tipo de estabelecimento de saúde: Posto de Saúde, Centro de Saúde/Unidade Básica de Saúde, Policlínica, Hospital Geral,Hospital Especializado, Unidade Mista, Pronto Socorro Geral, Pronto Socorro Especializado, Consultório Isolado, UnidadeMóvel Fluvial, Clínica Especializada/Amb. Especializado, Unidade de Serviço de Apoio de Diagnose e Terapia, Unidade MóvelTerrestre, Unidade Móvel de Nível Pré-hospitalar na Área de Urgência e Emergência: Farmácia: Unidade de Vigilância em Saúde,Cooperativa, Centro de Parto Normal Isolado,Hospital /Dia-Isolado, Central de Regulação de Serviços de Saúde, LaboratórioCentral de Saúde Pública – LACEN, Secretaria de Saúde
5	tp_gestao	Tipo de gestão: Estadual (E), Municipal (M) ou Dupla (D)
6	no_logradouro	Nome do Logradouro
7	nu_endereco	Número do Endereço
8	no_bairro	Bairro
9	co_cep	CEP
10	uf	UF do estabelecimento
11	municipio	Município
12	nu_telefone	Número de Telefone

Figura 4 - Colunas (em vermelho) excluídas do dataset do CNES

Será feita uma análise dos campos que são relevantes para o estudo. Para esses campos:

- Análise inicial dos dados para conhecer preliminarmente seu conteúdo e características: tipos dos dados, identificação de dados quantitativos e qualitativos, geração de estatísticas básicas (valores mínimos e máximos, média, mediana);
- Identificação de valores nulos e/ou faltantes e definição de estratégia de tratamento (atribuição de valor ou eliminação da linha). Utilizando o Pandas, será usada uma máscara ([A-Z],[A-Z]) para identificar se os campos de UF, seja do estabelecimento, seja do vacinado estão no formato correto e, caso não estejam, a linha será eliminada, pois como o trabalho se propõe a analisar dados de uma região específica, havendo dúvidas se o dado pertence ou não ao conjunto, é melhor que se desconsidere para evitar ruídos no resultado. O mesmo será feito para sexo ([M|F]) e raça ([1|2|3|4|99]). No caso de raça, caso o valor presente seja diferente de "1-branca", "2-preta", "3-parda", "4-amarela", "99-sem informação", será definido o valor para o campo como sendo "99-sem informação" (valor sentinela); e
- Identificação de registros duplicados e eliminação de um deles (caso os registros sejam exatamente iguais em todos os campos).

Será feito um *join* entre os datasets para que seja possível verificar o perfil dos vacinados por tipo de estabelecimento (hospital, centro de saúde, posto de saúde e etc.).

5. Proposta de modelo de extração de conhecimento e visualização dos dados

Nesta etapa, será utilizada a biblioteca *matplotlib* para que seja possível construir os gráficos definidos no item 3.2 anterior, utilizando os seus principais comandos (plot, show, grid e etc.).

6. Coleta de metadados da proveniência dos experimentos

Nesta etapa, serão coletadas as informações dos datasets que indicarão a origem dos dados, ou seja, de onde foram baixados, de que forma foram tratados e as transformações a que foram submetidos para análises futuras e posterior reuso.

Será usada a biblioteca PROV, uma implementação em Python do Modelo de Dados de Proveniência da W3C[4]. Ela suporta importação/exportação dos formatos PROV-O (RDF), PROV-XML e PROV-JSON.

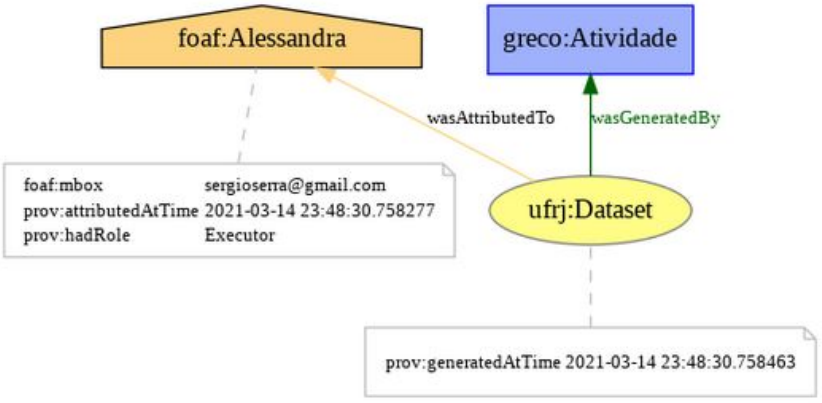

Figura 5 - Exemplo de grafo de proveniência gerado

Como exemplificado na Figura 5, serão coletadas as seguintes informações de proveniência:

- O agente executor da atividade;
- O nome da atividade executada;
- O dataset consultado/consumido/analisado;

[4] World Wide Web Consortium (W3C) é uma comunidade internacional que desenvolve padrões abertos com o intuito de garantir o crescimento da Web a longo prazo. Link: https://www.w3.org/.

- Informações adicionais do agente executor:
  - E-mail;
  - Data e hora (timestamp) da atribuição do perfil ao agente;
  - Perfil do agente;
- Data e hora (timestamp) da geração dos dados de proveniência.

7. Projeto de reprodutibilidade do experimento

Um experimento reprodutível é aquele em que se é possível recriar o mesmo resultado usando código e dados originais, mesmo sendo feito por equipes, sistemas de medição e localidades diferentes[4]. Há de se atentar para não se confundir com o conceito de replicabilidade, que é a capacidade de recriar o mesmo resultado, usando, porém, novos dados, a partir do mesmo plano de experimento. Pode ocorrer do programa não ser reprodutível devido a diferentes razões, como: versão antiga do compilador, métodos ou funções descontinuados pela linguagem, erro de programação, etc.

Com o objetivo de se reproduzir o mesmo resultado deste trabalho, é necessário que se use o mesmo código, ambiente computacional e datasets usados no experimento. Como existe a possibilidade dos datasets serem modificados/atualizados em seus repositórios originários de armazenamento, serão disponibilizados publicamente para download, as versões dos datasets usadas na experimentação, assim como o código fonte do programa em linguagem Python.

Especificamente em relação a experimentos computacionais, garantir a reprodutibilidade por um longo período é um desafio, dada a evolução da tecnologia e a eventual descontinuidade de recursos utilizados. Nesse caso, busca-se descrever o ambiente computacional com o máximo de detalhe para que seja facilmente identificável qualquer ajuste necessário quanto não for possível utilizar exatamente o mesmo ambiente computacional.

Dessa forma, as seguintes medidas serão tomadas em relação ao código do experimento visando a reprodutibilidade do mesmo:

1. Identificação em detalhes do ambiente computacional
2. Inclusão de comentários descrevendo as ações realizadas
3. Utilização de funções ou recursos da versão mais atual do Python
4. Tornar o código auto verificável.

8. Referências

[1] Ministério da Saúde. **Cadastro Nacional de Estabelecimentos de Saúde (CNES)** 2018. Disponível em: <https://dados.gov.br/dataset/cnes_ativo> Acesso em: 05 de março de 2021.

[2] OpenDataSus. **Campanha Nacional de Vacinação contra Covid-19** 2021. Disponível em: <https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao> Acesso em: 05 de março de 2021.

[3] SIDRA - **Sistema IBGE de Recuperação Automática. População Residente Estimada** - Ano 2020. Disponível em: <https://sidra.ibge.gov.br>. Acesso em: 05 de março de 2021.

[4] Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions Fabien C. Y. Benureau and Nicolas P. Rougier