REPOSITÓRIO COVID-19 Data Sharing/BR

INSTRUÇÕES PARA CRIAÇÃO DE BASE DE DADOS RELACIONAL E UPLOAD DE ARQUIVOS

Danilo Carlotti, João Eduardo Ferreira, Fátima L. S. Nunes

1. Introdução

Este conjunto de documentos disponibiliza um modelo relacional e programas que implementam uma base de dados e visualizações iniciais considerando os dados disponibilizados pelo repositório COVID-19 Data Sharing/BR, disponível em: https://repositoriodatasharingfapesp.uspdigital.usp.br/.

Este conjunto é composto pelos seguintes arquivos:

- analises_descritivas.py – programa que gera análises descritivas (ver seção 6 deste documento);
- connection_mysql.py – programa de conexão com o banco de dados (ver seção 5 deste documento);
- create_db.sql – *script* de texto com as informações necessárias para criação da base de dados no formato para fazer o *upload* das informações (ver seção 4 deste documento);
- data_dictionary.pdf – dicionário de dados com modelo das tabelas na língua inglesa (ver seção 5 deste documento);
- dicionário_de_dados.pdf – dicionário com modelo das tabelas (ver seção 5 deste documento);
- instruções.pdf – este documento em língua portuguesa;
- instructions.pdf – este documento em língua inglesa;
- map_server_bokeh_infections_per_city.py – programa para geração do mapa com número de infectados por cidades;
- model_db.mwb – arquivo do mysql workbench com o modelo da base de dados (ver seção 2 deste documento);
- parse_exams.py – programa de interpretação dos resultados e dos valores de referência;
- standard_reference_value.py – programa de apoio utilizado pelo processamento dos dados (ver seção 5 deste documento);
- upload_dados.py – programa utilizado para upload dos dados na base de dados criada pelo script mencionado (ver seção 5 deste documento);
- validator.py – programa validador dos dados para conferir se os dados foram corretamente anonimizados e se estão com encoding e outras informações necessárias de acordo com o dicionário dos dados.

Os arquivos com dados anonimizados de pacientes diagnosticados com COVID-19 encontram-se disponíveis em https://metabuscador.uspdigital.usp.br/handle/doc/12344.

Para fazer a instalação do servidor e executar os programas para upload de dados na base de dados, siga as instruções apresentadas na seção 3.

2. Modelo de dados

A Figura 1 apresenta o diagrama Entidade Relacionamento da base de dados, contendo as entidades que compõem a base de dados e seus relacionamentos, conforme dicionário de dados disponível no repositório.

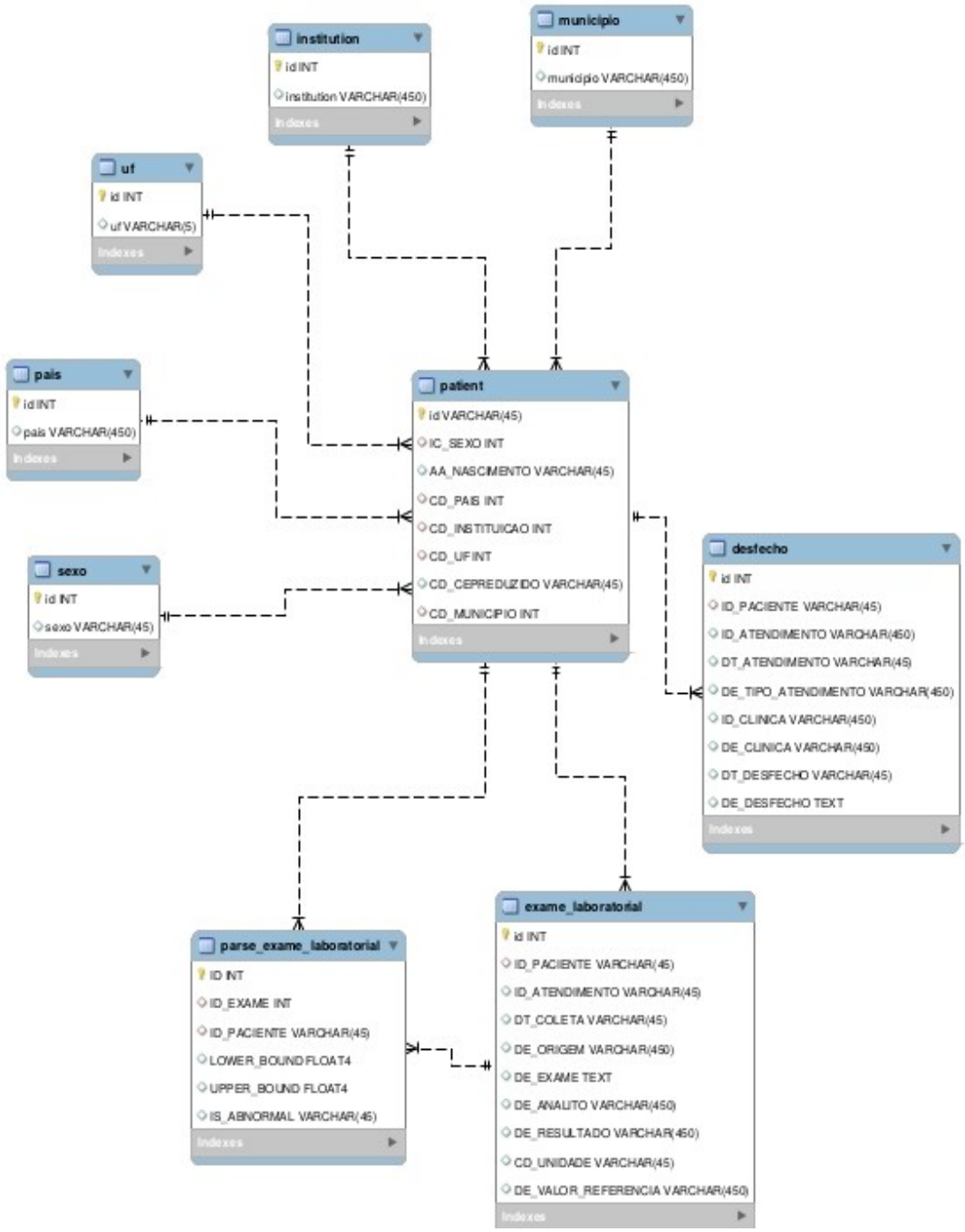

Figura 1 - Diagrama Entidade-Relacionamento da base de dados

3. Instalação do servidor

A base de dados foi projetada para um servidor SQL, que pode ser instalado na versão gratuita disponível do MySQL (https://dev.mysql.com/downloads/mysql/).

Um tutorial para executar a instalação está disponível em: https://dicasdeprogramacao.com.br/como-instalar-o-mysql-no-windows/.

Caso o usuário deseje uma interface gráfica por meio da qual trabalhar com o servidor há alternativas gratuitas como SQL Workbench cujo *download* pode ser feito em: https://dev.mysql.com/doc/workbench/en/wb-installing-windows.html.

4. *Script* para criação da base de dados

O *script* para criação da base de dados com todas suas tabelas encontra-se no arquivo "*create_db.sql*". Por meio de interface gráfica ou copiando o texto do *script* no terminal é possível restaurar a base.

5. *Upload* e processamento dos dados

O processamento dos dados é feito por um programa Python. Os requisitos para executar os programas são:

- python versão 3.5 ou superior
- pandas
- pymysql
- geopandas
- matplotlib
- bokeh

A instalação de python (versão Windows) pode ser feita pelo link: https://python.org.br/instalacao-windows/.

A instalação das bibliotecas pode ser feita conforme tutorial em: https://computadorcomwindows.com/2018/01/19/tutorial-como-instalar-uma-biblioteca-python-no-computador/.

O arquivo que contém as credenciais de conexão ao banco de dados é o "*connection_mysql.py*". O endereço do servidor padrão no arquivo, o usuário e senha, respectivamente, são: 'localhost', 'root', 'root'. Esses parâmetros de acesso poderão ser oportunamente alterados de acordo com o interesse do leitor.

A inserção dos dados pressupõe que eles estejam adequados ao dicionário de dados disponibilizado no arquivo "dicionário_de_dados.pdf".

Para inserir os dados de pacientes é necessário executar a seguinte instrução em linha de comando correspondente à chamada para o arquivo python e três argumentos:

python3 upload_dados.py "[CAMINHO DO ARQUIVO]" [TIPO DO ARQUIVO(exame,paciente ou desfecho)] [SIGLA OU NOME DA INSTITUIÇÃO]

Exemplo:

python3 upload_dados.py "hsl_small_dataset_fapesp/small/COVID_PATIENT.csv" paciente HSL

Após o *upload* de todos os arquivos é necessário executar, sem argumentos, o programa "parse_exams.py".

A interpretação dos resultados (a partir de um processo de *parser*) dos exames indica, na tabela *parse_exame_laboratorial*, se o resultado do exame foi anormal ou não.

6. Visualizações preliminares

Há dois programas que geram visualizações dos dados.

- "*analises_descritivas.py*": gera um arquivo no formato html com gráficos no formato pizza, a saber: cidades dos pacientes, desfechos, dos sexos dos pacientes e 15 exames mais anormais nos pacientes. É possível também gerar um histograma com a idade dos pacientes. Há uma função neste arquivo para geração de um arquivo em excel com os casos por municípios. Este arquivo é necessário para gerar a visualização do arquivo "*map_server_bokeh_infections_per_city.py*".
- "*map_server_bokeh_infections_per_city.py*": gera um arquivo html com a biblitoeca *bokeh* para visualização do número de infectados por cidade. Os nomes das cidades, devem estar grafados corretamente colocados para serem equiparados aos nomes constantes no arquivo shapefile disponibilizado pelo IBGE com os municípios de São Paulo (*35MUE250GC_SIR.shp*).

Alguns exemplos de visualização considerando os dados disponíveis em 02.06.2020 são apresentados nas figuras de 2 a 7.

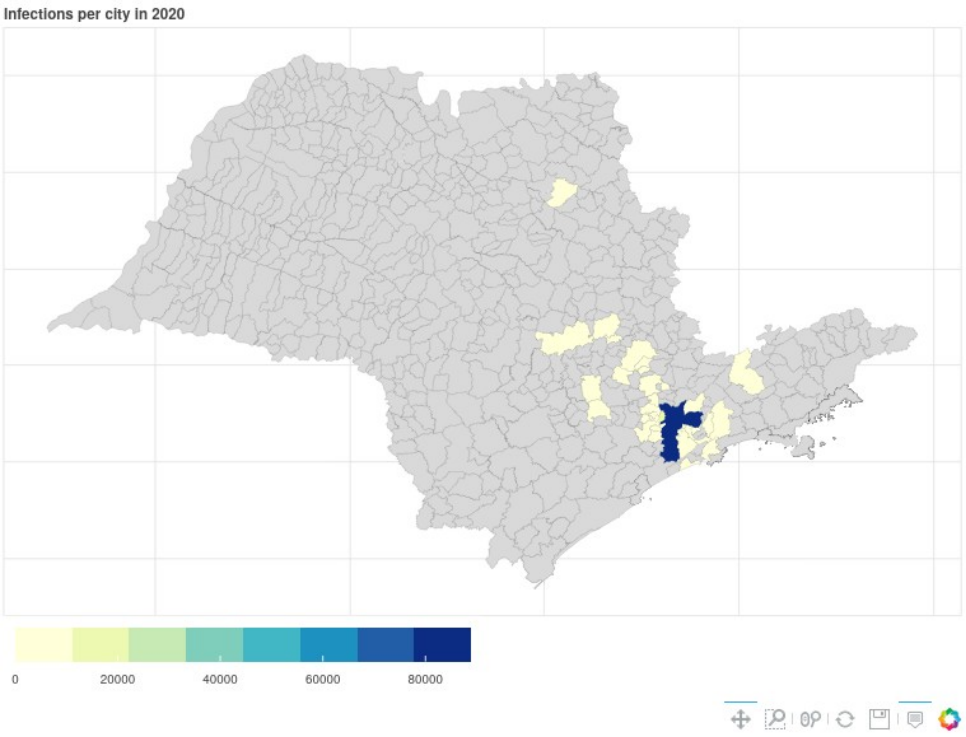

Figura 2 – Mapa no formato html das infecções por municípios do Estado de São Paulo

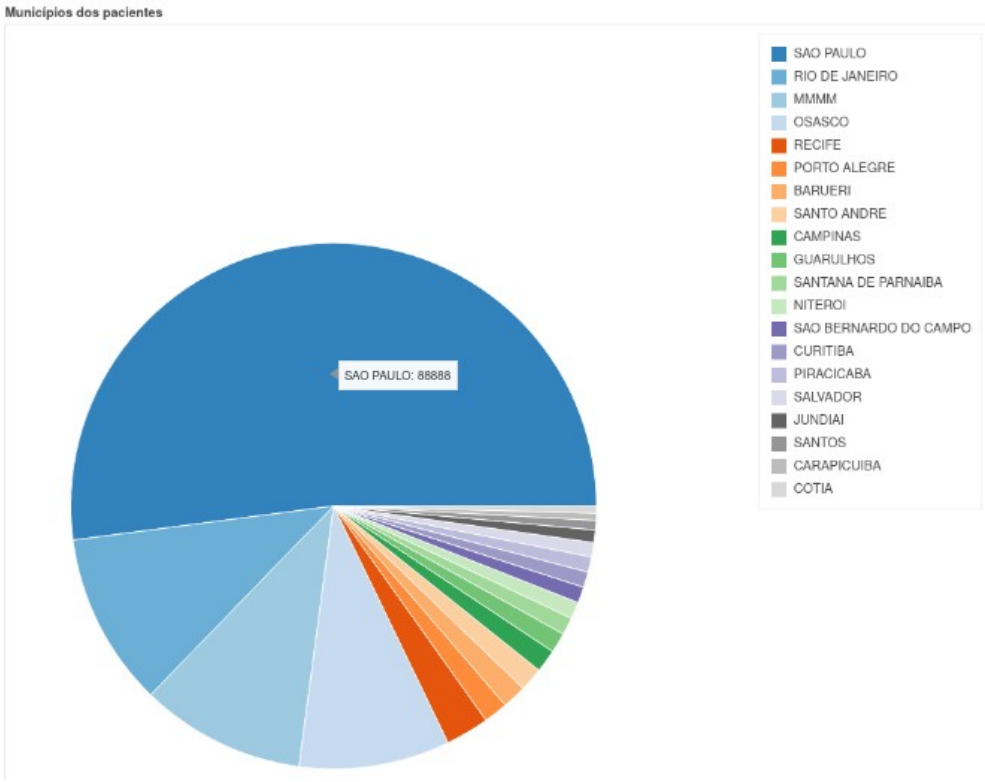

Figura 3 - Gráfico de pizza com infecções por municípios em 2020

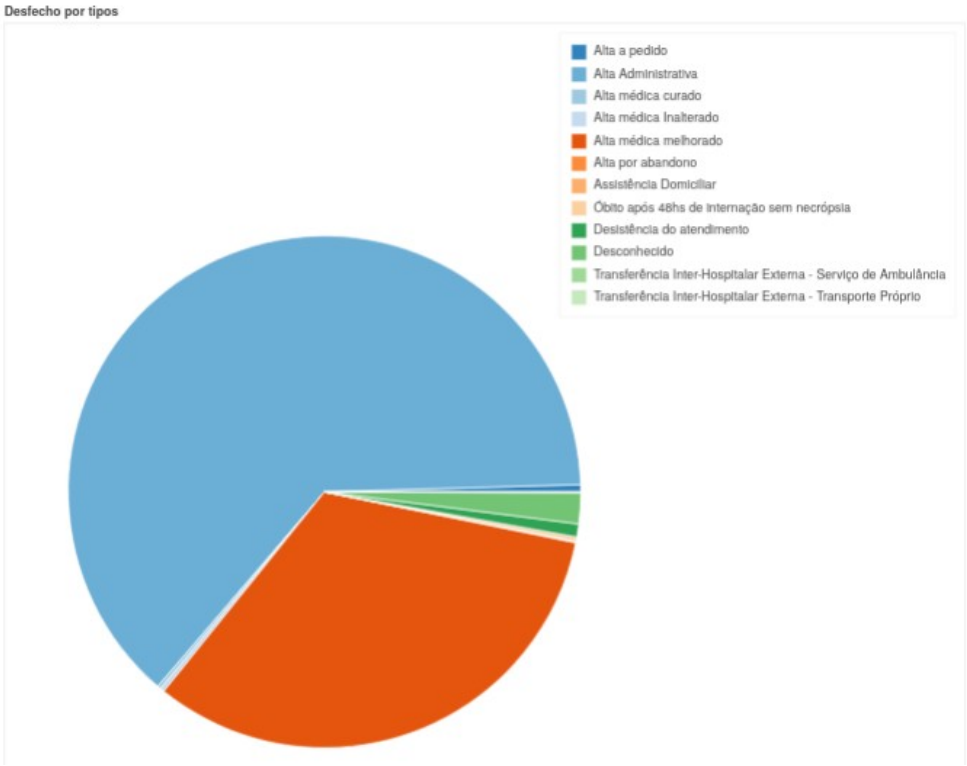

Figura 4 - Gráfico de pizza de desfechos (por conta de problema no encoding do arquivo, os nomes dos desfechos foram alterados manualmente para esta apresentação)

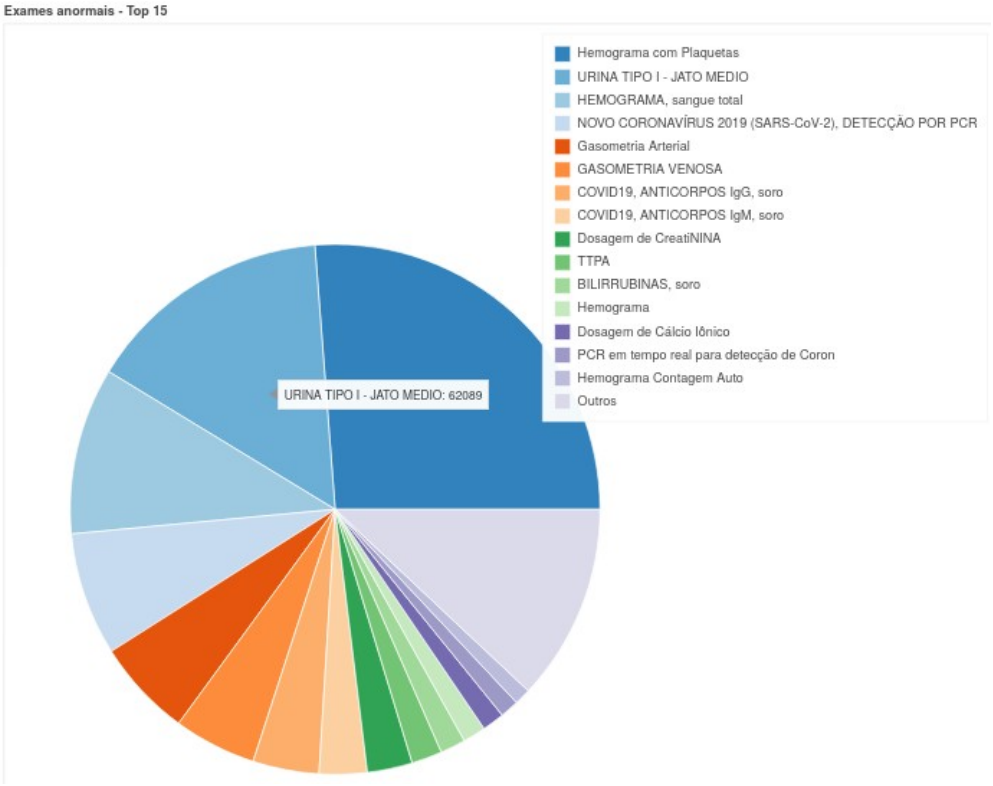

Figura 5 - Gráfico de pizza de exames anormais.

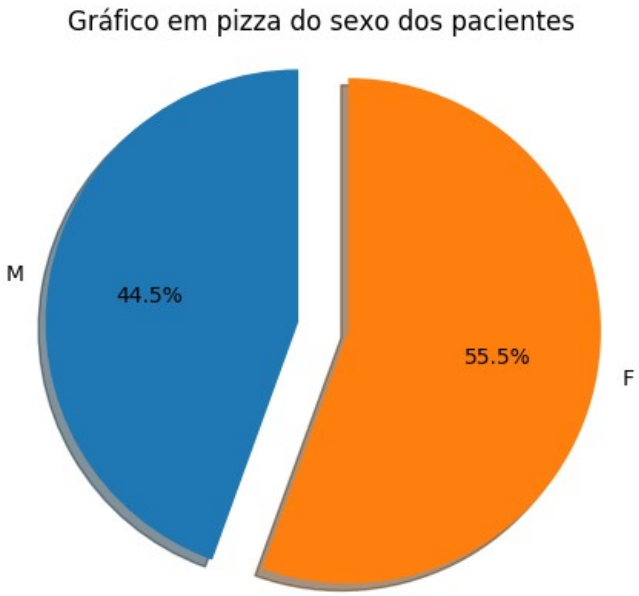

Figura 6 - Gráfico de pizza de sexo dos pacientes

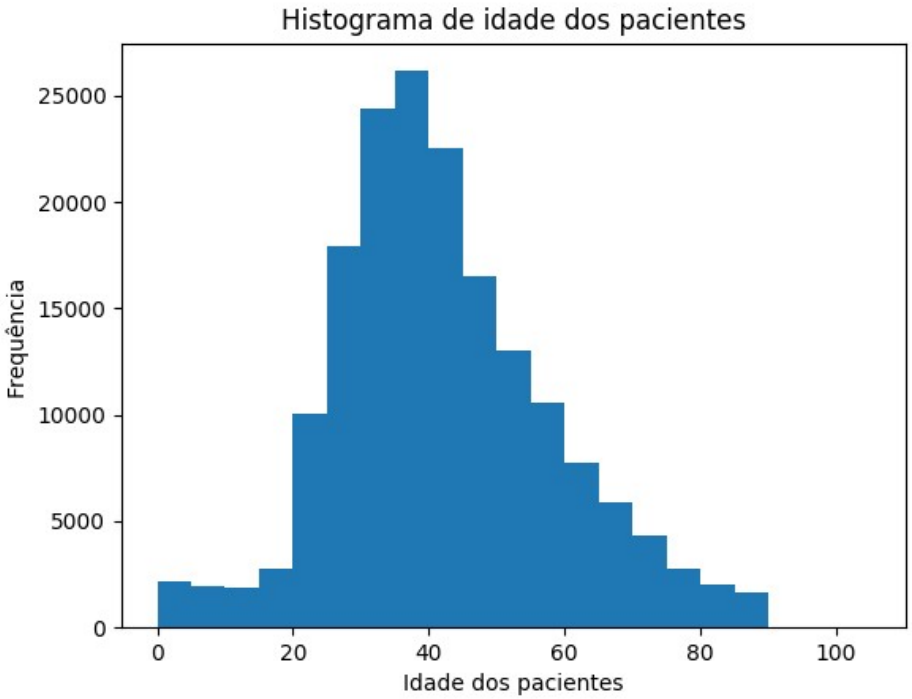

Figura 7 - Histograma da idade dos pacientes