Danillo Rodrigues Abreu

Projeto de Visualização Computacional: Análise Exploratória do Dataset - Registros dos Casos de Dengue em Recife - PE

Arapiraca – AL
2022

Danillo Rodrigues Abreu

Projeto de Visualização Computacional: Análise Exploratória do Dataset - Registros dos Casos de Dengue em Recife - PE

Projeto solicitado aos discentes do 8º período do curso Ciência da Computação, para fins avaliativos da disciplina de Visualização Computacional.

Universidade Federal de Alagoas – UFAL
Campus Arapiraca
Ciência da Computação

Orientador: Prof. Dr. Tácito Trindade de Araújo Tiburtino Neves

Arapiraca – AL
2022

Sumário

Projeto de Visualização Computacional

1 Descrição do dataset que será utilizado

O presente projeto tem como intuito fazer uma análise exploratória de um dataset que é referente aos registros dos casos de dengue contabilizados nas unidades de saúde, públicas ou particulares, da cidade de Recife - PE. Os registros foram disponibilizados publicamente pela Prefeitura de Recife e datados de 2013 até novembro de 2021. O dataset se encontra no endereço eletrônico: <http://dados.recife.pe.gov.br/it/dataset/casos-de-dengue-zika-e-chikungunya>. A análise será feita com o auxílio da linguagem de programação Python com a IDE Spyder (Python 3.9) e a biblioteca pandas.

1.1 Qual motivo para utilizar o dataset escolhido?

Para auxiliar os especialistas de saúde a encontrar uma solução para as pandemias de Dengue, se faz necessário um melhor entendimento de como a doença se comporta a partir de características associadas aos infectados, como: sintomas, idade, sexo, período dos sintomas, dentre outras. Dessa forma, com a analise do dataset disponibilizada pela prefeitura de Recife, e usando os conhecimentos de visualização computacional, é possível um melhor entendimento da doença a partir da sua análise, podendo assim extrair/filtrar algumas informações que possam auxiliar os especialistas a encontrar melhores soluções ou formas mais eficientes de prevenção da doença.

1.2 Origem do dataset

O dataset que será utilizado no projeto contém informações de registros dos casos de dengue vindos das unidades de saúde, públicas ou particulares da cidade de Recipe - PE. Os registros desses casos contém notificações de dengue com dados sobre a data de notificação, classificação, os casos confirmados, descartados ou inconclusivos, o critério de confirmação, a evolução do caso e localização, tudo por ano. No total foram 47681 instâncias de dados, distribuidas da seguinte forma:

- *2013* - total de instâncias 3229;
- *2014* - total de instâncias 1192;
- *2015* - total de instâncias 5250;
- *2016* - total de instâncias 18612;
- *2017* - total de instâncias 2454;
- *2018* - total de instâncias 2687;
- *2019* - total de instâncias 1500;
- *2020* - total de instâncias 3540;
- *2021* - total de instâncias 9217;

O conjunto de dados possui um total de 9 tabelas, que vão desde 2013 até 2021, sendo asssim, uma tabela por ano, e cada uma delas contém os mesmos 101 atributos (menos as tabelas de 2013 e 2014 que tem menos atributos), sendo eles:

- *_id*;
- *nu_notificacao*;
- *tp_notificacao*;
- *co_cid*;
- *dt_notificacao*;
- *ds_semana_notificacao*;
- *notificacao_ano*;
- *co_uf_notificacao*;
- *co_municipio_notificacao*;
- *id_regional*;
- *co_unidade_notificacao*;
- *dt_diagnostico_sintoma*;
- *ds_semana_sintoma*;
- *dt_nascimento*;
- *nu_idade*;
- *tp_sexo*;
- *tp_gastante*;
- *tp_raca_cor*;
- *tp_escolaridade*;
- *co_uf_residencia*;
- *co_municipio_residencia*;
- *co_regional_residencia*;
- *co_distrito_residencia*;
- *co_bairro_residencia*;
- *no_bairro_residencia*;
- *co_logradouro_residencia*;
- *nome_logradouro_residencia*;
- *co_geo_campo_1*;
- *co_geo_campo_2*;
- *ds_referencia_residencial*;
- *nu_cep_residencia*;
- *tp_zona_residencia*;
- *co_pais_residencia*;
- *tp_duplicidade*;
- *dt_digitacao*;
- *dt_transf_us*;
- *dt_transf_dm*;
- *dt_transf_sm*;
- *dt_transf_rm*;
- *dt_transf_rs*;
- *dt_transf_se*;
- *nu_lote_vertical*;
- *nu_lote_horizontal*;
- *tp_fluxo_retorno*;
- *st_fluxo_retorno _recebido*;
- *ds_identificador_registro*;
- *st_importado*;
- *dt_investigado*;
- *co_cbo_ocupado*;
- *dt_coleta_exame*;
- *tp_result_exame*;
- *dt_coleta_NS1*;

- *tp_result_NS1*;
- *dt_coleta_isolam*;
- *tp_result_isolam*;
- *dt_coleta_rtpcr*;
- *tp_result_rtpcr*;
- *tp_sorotipo*;
- *tp_result_histopatologia*;
- *tp_result_imonohistoquimica*;
- *tp_classificacao_final*;
- *tp_criterio_confirmado*;
- *tp_autoctone_residencia*;
- *co_uf_infeccao*;
- *co_pais_infeccao*;
- *co_municipio_infeccao*;
- *co_distrito_infeccao*;
- *co_bairro_infeccao*;
- *no_bairro_infeccao*;
- *st_doenca_trabalho*;
- *tp_evolucao_caso*;
- *dt_obito*;
- *dt_encerramento*;
- *st_ocorreu_hospital*;
- *dt_internacao*;
- *co_uf_hospital*;
- *co_municipio_hospital*;
- *co_unidade_hospital*;
- *nu_ddd_hospital*;
- *nu_telefone_hospital*
- *febre*;
- *mialgia*;
- *cefaleia*;
- *exantema*;
- *vomito*;
- *nausea*;
- *dor_costas*;
- *conjutivite*;
- *artrite*;
- *artralgia*;
- *petequia_n*
- *leucopenia*;
- *laco*;
- *dor_retro*;
- *diabetes*;
- *hematolog*
- *hepatopat*
- *renal*
- *hipertensao*
- *acido_pept*
- *auto_imune*

2 Referencial teórico

2.1 Trabalhos que já utilizaram esse dataset

O artigo "PROJETO DE ANÁLISE DE DADOS PARA IMPLANTAÇÃO DE DATA MART COMO FERRAMENTA PARA TOMADA DE DECISÃO EM COMBATE AOS VÍRUS DA DENGUE, ZIKA E CHIKUNGUNYA", publicado em 2017 na revista **InterScientia**, apresentou uma utilização das bases de dados, disponibilizadas publicamente pela Prefeitura de Recife referente as doenças epidêmicas de Dengue, Zika e Chikungunya na cidade de Recife. O objetivo deste trabalho é utilizar de ferramentas e técnicas de Business Intelligence(BI) para análise e mapeamento da bases de dados citada na saúde pública, para que através desta análise possa-se obter um sistema de apoio à tomada de decisão a favor do combate as doenças mencionadas.

O resultado desse artigo foi a criação de um *Data Mart*, em modelo estrela e usando esse *Data Mart*, fizeram registros e dashboards (painéis gráficos) com desempenho de alta qualidade e diferentes amostras matemáticas e da estatística dos casos por: bairro, ano, gerais, unidade de saúde e vários outros filtros, com os dados mecionados.

Outro trabalho que também faz uso desse dataset é o artigo "UTILIZAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DE ARBOVIROSES TRANSMITIDAS PELO AEDES AEGYPTI", publicado em 2020 na revista **Conexões - Ciência e Tecnologia**. O trabalho tem como objetivo utilizar algoritmos de aprendizagem de máquina para predizer casos das arboviroses dengue e chikungunya, transmitidas pelo mosquito Aedes aegypti, a partir de características associadas ao paciente, tais como, sintomas, idade, sexo, período dos sintomas, dentre outros. Para a predição, os dados passaram por uma etapa de pré-processamento, processamento e análise. Foram utilizados três algoritmos de aprendizagem de máquina para comparação de resultados: J48, Random Forest e Redes Neurais, com o balanceamento de dados através do SMOTE.

Nesse trabalho, conclui-se que o algoritmo Random Forest apresenta melhores resultados se comparados com o demais, alcançando 90,6443% de acurácia e 0,907 de f-measure, sendo, portanto, uma alternativa promissora para a predição de dengue e chikungunya.

2.2 Como se dará a analisada exploratória dos dados

A análise exploratória de dados (EDA) é usada para analisar e investigar conjuntos de dados e resumir suas principais características, muitas vezes usando métodos de visualização de dados. Ela permite determinar a melhor forma de controlar as fontes de dados para obter as respostas que você precisa, tornando mais fácil para os cientistas de dados descobrir padrões, detectar anomalias, testar uma hipótese ou verificar suposições.

Inicialmente vamos extrair algumas informações do dataset e tentar fazer algumas representações visuais (gráficos) que auxiliem no melhor entendimendo desse conjunto de

dados. Algumas informações que podem ser extraídas são:

- Quantidade de casos notificados por ano;
- Quantidade de casos notificados por mês do ano que mais teve casos confirmados;
- Quantidade de casos resistrados por bairro;
- Conclusão da enfermindade (cura, óbito, não informado);
- Frequência das pessoas atingidas com base na idade.

Então criaremos algumas hipóteses baseadas no dataset, e mais na frente, quando dermos início de fato a análise exploratória, com auxílio da linguagem de programação Python (usando a biblioteca pandas), poderam ser comprovadas ou descartadas. Por hora penso sobre quatro hipóteses que posteriormente poderemos testar:

1. Pacientes que já apresentavam quadro de diabentes e hipertensão antes de serem atingidos pela dengue demostraram sistomas mais graves!
2. Entre todos os atingidos pela dengue, as mulheres foram as que aprensentaram sintomas mais graves de febre, vômito, nause e exantema!
3. Água Fria é o Bairro que mais apresentou casos de dengue em 2020, o que implica que esse Bairro deveria ter um maior apoio por parte da prefeitura para evitar que o número de casos continuem crescendo nessa área.
4. A quantidade de óbitos está relacionada com a quantidade de casos atingidos no ano (quanto mais casos de dengue, mais óbitos).

De início vou aplicar algumas operações estatísticas de exploração de dados, visando vacilitar o processo de análise do dataset afim de verificar se as hipóteses estão corretas ou não. Tais operações serão:

1. Frequência - usaremos para precisar a fração de frenquência de um certo valor, como por exemplo verificar a fração de frenquência de pessoas afetas pela dengue por bairro da cidade de recife.
2. Moda - pode auxiliar na análise de dados nominais, como por exemplo qual o nome do bairro que mais apresenta casos de dengue em denterminado ano, ou também podemos usar para descobrir qual o sexo que mais foi afeto pela dengue, ou outros fatos que viermos a querer descobrir sobre aquele conjunto de dados.
3. Média - no caso de variáveis quantitativas usaremos ela para determinar a tendência central.

4. Mediana - também para variáveis quantitativas, podemos usar quando quisermos verificar o valor que se encontra na posição central.

Posterior a isso, poderemos aplicar outras técnicas da análise exploratória que vão auxiliar nos testes das hipóteses. Contudo, antes de iniciar a análise exploratória é preciso preparar o conjunto de dados de forma que seja possível a aplicação da análise sem eventuais problemas, como desnormalização e despadronização do conjunto de dados e também dados faltantes no conjunto de dados. Dessa forma, entedesse que é necessário que façamos o pré-processamento de dados no dataset antes de iniciarmos a analise exploratória.

2.3 Tarefas de preparação dos dados necessárias

Para dar início ao pré-processamento será necessário ter domínio sobre as bases de dados que seram precisas para o caso. Estas bases são sobre os casos de dengue nas unidades de saúde obtidas através do site de dados abertos da Prefeitura de Recife. As bases estão desnormalizadas e despadronizadas, com alguns valores nulos, incoerentes e incorretos. Tais dados teram de passar por etapas de transformação para serem tratados e se tornarem úteis ao uso.

Antes de iniciar o pré-processamento em si, é importante conhecer as principais estruturas dos dados: dados estruturados, dados semiestruturados e dados não estruturados.

Os dados estruturados contêm uma organização rígida e previamente planejada. Normalmente são “etiquetados” em linhas e colunas que identificam suas características a respeito de determinados assuntos. São exemplos desse tipo de dado os bancos de dados relacionais, planilhas excel e arquivos CSV.

Já os dados semiestruturados são os dados que possuem uma estrutura, mas que não está de acordo com estruturas formais dos modelos associados a bancos de dados relacionais ou outras formas de tabelas de dados. Eles possuem marcadores, como tags, para separar elementos semânticos e criar hierarquias para os registros e campos. Alguns exemplos desse modelo são os arquivos XML, JSON e HTML.

E por fim, os dados não estruturados são os dados nos quais não conseguimos identificar uma organização clara. Para gerar insights sobre estes dados é preciso realizar um intenso pré-processamento para recuperar a informação. São exemplos de dados não estruturados os documentos de texto, áudio e imagens.

O dataset que utilizaremos se encontra disposível para download em vários formatos, sendo eles: CSV, TSV, JSON e XML, dessa forma, não será necessário realizar pré-processamento referente a transformação em estrutura de dados, pois o dataset já é disponibilizado nas formas: estruturadas (CSV e TSV) e semiestruturadas (XML e JSON).

Entre os principais problemas encontrados dentro de um conjunto de dados, também conhecido como dataset, podemos elencar os atributos com valores faltantes, os outliers e as escalas diferentes para valores iguais. Uma técnica de mineração de dados usada

para resolver esses problemas é a transformação desses dados brutos em formatos úteis e eficientes (atrás de exclusão de linhas e colunas, adoção de valores para dados faltantes através da média ou moda dos demais dados e outras mais técnicas). Pode ser necessário fazer pré-processamento em qualquer uma das 3 estruturas de dados citadas.

A primeira técnica de pré-processamento que será aplicada é a limpeza, ela é utilizada para manuseio e/ou preenchimento de dados ausentes, redução de ruídos, identificação e remoção de valores aberrantes e a resolução de inconsistências. No caso do dataset que estamos utilizando, vamos usar a técnica de limpeza para remover alguns registros cujo alguns atributos tem valor nulo e também será necessário desconsiderar alguns atributos que apresentam valor nulo em todos os registros, dessa forma diminuindo a quantidade de registros/instâncias/linhas e diminuindo também a quantidade de atributos/colunas. Por exemplo, temos alguns registros cujo a data de nascimento não foi registrada e isso é um problema, pois a idade é derivado do atributo data de nascimento, caso um especialista entenda que para fazer uma melhor análise das pessoas que foram atingidas pela doença dengue é necessário separaras pela idade, teremos registros inutéis, pois não carregam esse dado, dessa forma é melhor que sejam removidos esses registros cujo a data de nascimento não consta. Outro atributo que também deve ser considerado na hora da limpeza é o tipo de febre, em alguns registros não é contabilizado qual o tipo de febre que o paciente atingiu, dessa forma, os registros que não contém esse dado são inrelevantes, logo também devem ser removidos.

Alguns outros atributos, como: mialgia, vômito, cefaleia, exantema, nausea, entre outros que também guardam informações referentes aos sintomas que os pacientes apresentaram, também são muitos importantes na hora de analisar o dataset, dessa forma também é necesserário aplicar a limpeza, removendo os registros cujo esses atributos tem valor nulo. Também podemos remover registro que não apresentam valor nulo, mas que não são de nosso interessa na análise dos dados.

É bom frisar que cada situação pode exigir uma estratégia diferente para lidar com dados faltantes, nesse projeto será usada a estratégia de apagar os registros que tem dados nulos nos quais, ao meu ver, são dados que não podem faltar na hora de analisar o dataset, porém existem outras estratégias de como tratar esses dados faltantes, como por exemplo: realizar uma média com os valores do mesmo atributo; realizar uma mediana com os valores do mesmo atributo ou preencher o atributo faltante com os valores que mais ocorrem no dataset (moda).

A segunda técnica de pré-processamento que seria usada é a normalização. Como os dados de sintomas da dengue são cruciais para entendermos como essa infermidade afeta os infermos eu pensei em normalizar os dados referentes aos sintomas em uma escala de 0 à 10 para que mais na frente pudesse fazer algumas combinações e comparações, visto que os algoritmos não trabalham muito bem com dados despradonizados ou de dimensões muito diferentes, essa padronização dos dados por meio de escala ou média seria muito

útil. Porém os dados referentes ao sintomas (febre, mialgia, vômito, cefaleia, exantema, nausea e outras mais) já se encontram normalizados dentro do dataset em uma escala de 0 à 2 (dimensões não muito distantes), dessa forma foi decidido não manipular esses dados, sendo assim encerrando a etapa de normalização.

3 Projeto

Para manipular o dataset será usado a linguagem de programação Python com a IDE Spyder (Python 3.9) e a biblioteca pandas, que vai ajudar muito na análise de dados, pois a biblioteca disponibilizar diversos recursos que vão auxiliar a análise, como:

- Carregar dados externos (arquivo CSV do dataset) para o python;
- Fazer algumas manipulações estatísticas no dataset: contagem, média, desvio padrão, mínimo, os quartz, máximo, etc.
- Seleção/filtro de dados sem a necessidade de estruturas de condição e/ou repetição;
- Maior velocidade para obter resultados de tratamento de dados.

A IDE Spyder (Python 3.9) disponibiliza diversos recursos de visualização de dados, e se usada junto com os recursos da biblioteca pandas, facilitará ainda mais esse processo de análise exploratória que vamos fazer.

3.1 Preparação dos dados

Dando início ao pré-processamento, usando os recursos da biblioteca pandas, foi importado os arquivos CSV referentes aos dados do dataset para variáveis do tipo dataFrame (estrutura de dados bidimensional com os dados alinhados de forma tabular em linhas e colunas, no nosso caso essa estrutura vai conter o conjunto de dados de nosso interesse).

Porém, logo foi percebido que os arquivos disponibilizados pela prefeitura de recife, referente aos casos de dengue dos anos 2013 e 2014, não tem atributos que guardam informações sobre os sintomas apresentados pelos pacientes. Como essas informações são importantes para a nossa abordagem do dataset, foi decidido que seram desconsideras as duas variáveis (dataFrames de 2013 e 2014), não usando mais elas nas próximas etapas do pré-processamento, ficando apenas com os anos de 2015 a 2021.

Prosseguindo com a limpeza dos dados, foram excluidas as colunas que trazem dados que não seram úteis na nossa análise. Contudo, percebi que existem bem mais colunas que trazem informações que não vão contribuir com a análise, do que as que vão contribuir. Então resolvi criar outro dataset apenas com as colunas que nos intereção, o que é bem menos trabalhoso do que excluir todas as demais que não agregam valor para a nossa análise. As colunas escolhidas para esse novo dataset são:

- *dt_notificacao*;
- *no_bairro_residencia*;
- *dt_obito*;
- *dt _encerramento*;
- *dt_nascimento*;
- *tp_sexo*;
- *febre*;
- *mialgia*;
- *cefaleia*;
- *exantema*;
- *vomito*;
- *nausea*;
- *dor_costas*;
- *conjutivite*;
- *artrite*;
- *artralgia*;
- *petequia_n*
- *leucopenia*;
- *laco*;
- *dor_retro*;
- *diabetes*;
- *hematolog*
- *hepatopat*
- *renal*
- *hipertensao*

Com isso, foram criadas 7 novas variáveis do tipo dataFrame contendo os dados dos arquivos CSV referente aos casos de dengue dos anos 2015 a 2021.

O próximo passo é excluir do nosso novo dataset as linhas que apresentam valor nulo nos atributos referentes aos sintomas. Ao tentar fazer a limpeza desse dataset, percebi que todas as linhas que apresentam valor nulo em um dos atributos dos sintomas, também apresentam valor nulo para os demais atributos dos sintomas, com isso, se eu buscar o atributo *febre* e encontrar nulo, significa que os demais atributos referentes aos sintomas também são nulos, dessa forma basta ir excluindo a linhas cujo o atributo *febre* apresenta valor nulo.

Agora faremos o mesmo processo de limpeza nas linhas que apresentam valor nulo no atributo *dt_nascimento*.

Finalizado a limpeza, nos concluimos o pré-processamento do dataset, assim gerando um novo, com apenas dados que são do nosso interesse para realizar a análise. Ao concluir esse processo, exportamos o novo dataset em 7 arquivos CSV, um para cada ano dos casos de dengue (2015 a 2021).

3.2 Análise exploratória

4 Conclusões sobre o dataset

Referências

PORTAL de Dados Abertos da Cidade do Recife. **Casos de Dengue, Zika e Chikungunya**. 2016. Disponível em: <http://dados.recife.pe.gov.br/it/dataset/casos-de-dengue-zika-e-chikungunya> Acesso em: 15 nov. 2021.

PINHO, Maria Dominguez Costa. **Dados abertos governamentais e democracia digital: o estado da arte e uma aplicação aos portais de dados abertos de seis prefeituras brasileiras**. Monografia (Graduação em Jornalismo) – Faculdade de Comunicação, Universidade Federal da Bahia, Salvador, p. 77. 2017.

FREITAS, Pedro Augusto Mendes de. **Correlação espacial entre a dengue e o saneamento na cidade do Recife. 2019. 33 f**. TCC (Curso de Engenharia Civil) - Departamento Acadêmico de Infraestrutura e Construção Civil, Instituto Federal de Ciência e Tecnologia de Pernambuco, Recife, p. 33. 2019.

CARVALHO, N.; FERREIRA, D. G.; BRITO DE ARAÚJO, M. E.; LIMA, R. R. Projeto de análise de dados para implantação de Data Mart como ferramenta para tomada de decisão em combate aos vírus da Dengue, Zika e Chikungunya. **Revista InterScientia**, v. 5, n. 2, p. 106-123, 11 dez. 2017.

SILVEIRA, Francisca Raquel de Vasconcelos; MOREIRA, Lina Yara Monteiro Rebouças. UTILIZAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DE ARBOVIROSES TRANSMITIDAS PELO AEDES AEGYPTI. **Conexões - Ciência e Tecnologia**, [S.l.], v. 14, n. 1, p. 64-71, mar. 2020. ISSN 2176-0144.

Data Geeks Blog sobre Data Science Machine Learning. **Conheça as Etapas do Pré-Processamento de dados**. 2019. Disponível em: <https://www.datageeks.com.br/pre-processamento-de-dados/> Acesso em: 17 nov. 2021