Projeto xo-mosquito

Silva, Eduardo
els6@cin.ufpe.br

Martins, Thiago
tasm2@cin.ufpe.br

Almeida, Wellington
wba@cin.ufpe.br

1 de Julho, 2018

Este relatório tem como objetivo descrever as atividades desenvolvidas durante o projeto da disciplina *IF706 - Introdução a Ciência dos Dados*, os resultados e as dificuldades que foram encontradas. Os assuntos aqui descritos se complementam com a apresentação e os *notebooks* que se encontram neste Repositório do Github.

1 Proposta

Levando em consideração o contexto das doenças tropicais negligenciadas, em especial Dengue, Zika e Chikungunya, e seu impacto social e, segundo IANPHI (2016), econmico na américa latina, especialmente no Brasil, o projeto possui como proposta:

- Analisar a relaçã o entre o perfil social dos indivíduos afetados e as incidências de doenças transmitidas pelo mosquito *Aedes Aegypti*;
- Investigar a influência da localidade física e climáticos em que se encontram tais indivíduos;
- Encontrar padrões nos casos relatados para assim sugerir abordagens de vigilância que possam ser implementados por governos.

O nosso objetivo foi analisar essas hipóteses e tentar validá-las ou não em relação a Recife, Pernambuco, mas se não fosse possível, devido a falta de dados, então o escopo seria analisar globalmente.

2 Metodologia

A metodologia para desenvolvimento do projeto foi baseda nos pilares de Coleta de Dados, Pré-processamento dos dados, Análise Exploratória de Dados, Modelos de Aprendizado e Visualização de Dados Explanatória. As seguintes seções pretendem nortear as atividades desenvolvidas em cada uma delas.

3 Coleta de Dados

Primeiramente, precisávamos de dados relacionados às doenças. Para isso recorremos à base da dados abertos da cidade do Recife[1], onde se encontravam dados relacionados as doenças que estávamos analisando. Com importantes *features* relacionadas as pessoas, as doenças adquiridas, sintomas, sexo, idade, raça e gênero, como também ao local onde elas residem.

Mas também era necessário analisar os locais onde as pessoas residiam em relação a outras variáveis para saber se a influência da localidade física e climática eram válidas. Para isso coletamos os dados também da plataforma da cidade do recife, mas agora relacionada a Área Urbana[2]. Dataset fornecido pela Secretaria de Infraestrutura e Serviços Urbanos de Recife, este conjunto descreve dados sobre a área espacial da cidade do Recife, incluindo dados sobre divisões do espaço físico da cidade, como: bairros, distritos, setores, quadras, faces de quadra, RPAs (Regiões Político Administrativas), micro-regiões, praças, parques, áreas verdes, recursos hídricos, lotes e edificações.

Como tinhámos o que era necessário para a análise dos dados em relação a cidade do Recife, o nosso objetivo inicial, então descartamos a opção de analisar as doenças globalmente.

4 Análise dos dados

Começamos a analisar os dados para ver se eles realmente poderiam ser úteis e quais deles que mais valorizariam o desenvolvimento do projeto.

Após uma limpeza de dados invalidados e organização dos dados válidos, pôde se perceber quais eram os passíveis de serem analisados.

Em relação aos dados dos dataframes das doenças, variáveis relacionadas ao sexo, raça e cor, escolaridade, bairro da residência, e o tipo de dengue que a pessoa adquiriu, foram os que mais se sobressairam, tanto para uma abordagem de categorização dos tipos de dengues da Região Metropolitana do Recife, quanto para o desenvolvimento de um modelo logístico para analisar a dependência de características físicas do local. Análise exploratória sobre os dados é demonstrada nas figuras 1 e 2.

Em relação ao dataframe dos bairros, as características de Iluminação, pavimentação, coleta, arborização, rede de água, limpeza urbana, rede de esgoto, rede telefônica, sarjetas, galerias pluviais e rede elétrica, foram de extrema importância para a conexão entre os dois dataframes, além da caracterização física do local.

5 Visualização dos Dados

A análise exploratória dos dados foi feita majoritariamente por dados explícitos, mas para realmente entendermos qual o impacto das análise utilizamos a biblioteca *folium* para plotar os dados nos bairros.

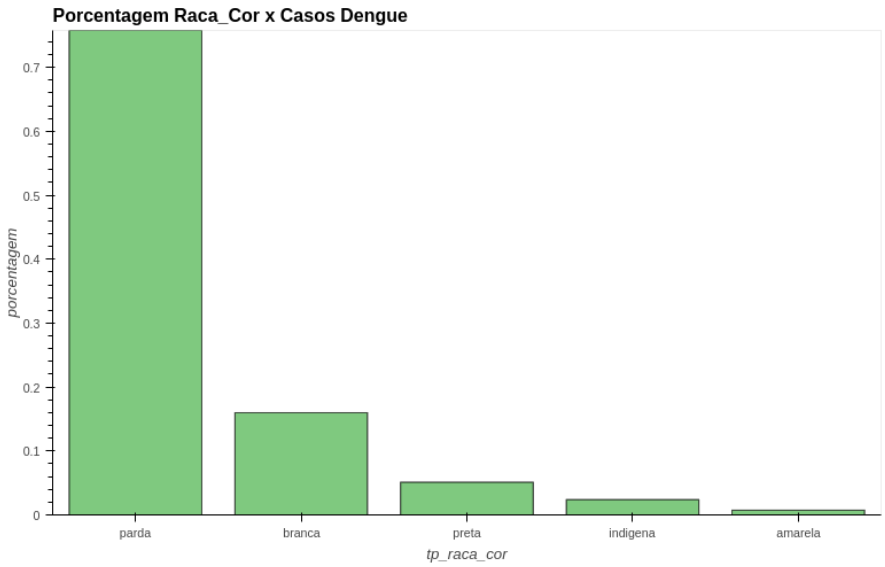

Figure 1: Porcentagem casos de dengue vs Racça/Cor, 2016, Recife.

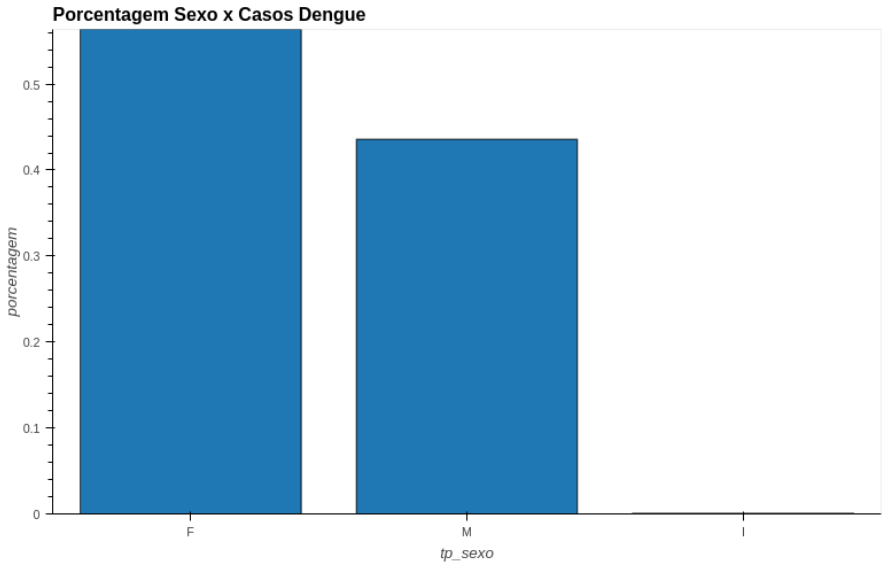

Figure 2: Porcentagem casos de dengue vs Sexo, 2016, Recife.

A figura 3 demonstra a distribuição da quantidade de casos de dengue que ocorreram em Recife no ano de 2016.

Além de termos desenvolvido um sistema de camadas, onde pode-se escolher

qual o tipo de dengue que se quer analisar. Como mostra a figura 4.

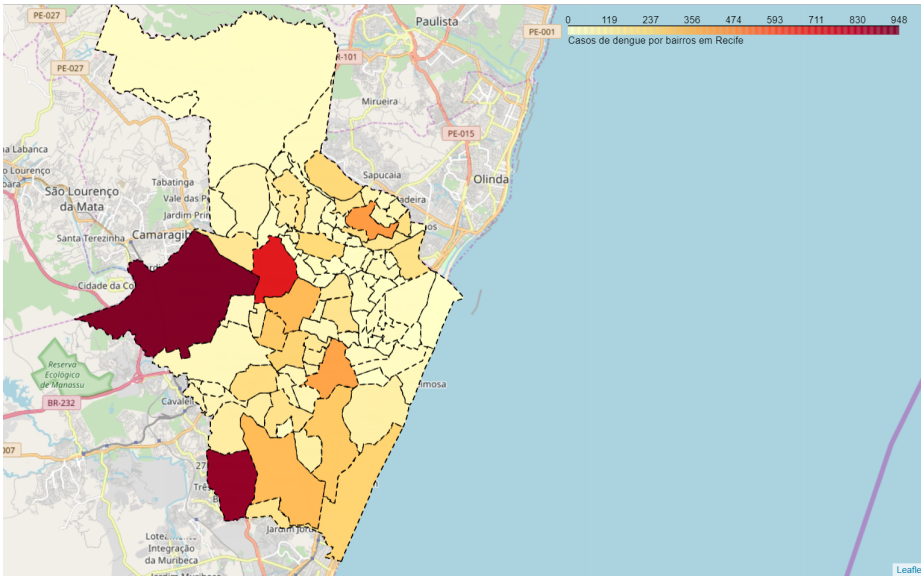

Figure 3: Quantidade de casos de dengue em geral na cidade de Recife, 2016.

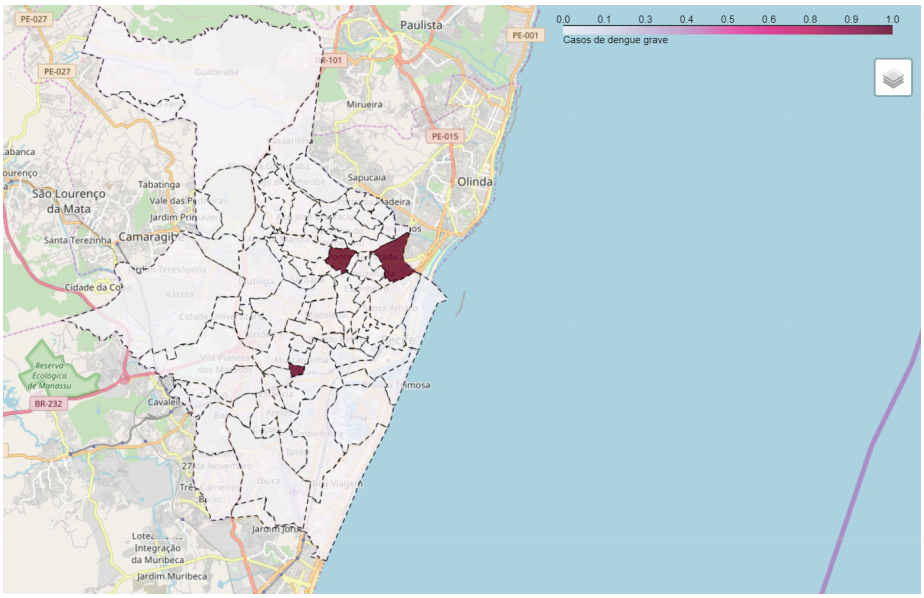

Figure 4: Quantidade de casos de dengue caso grave na cidade de Recife, 2016.

Mediana	2
Quantidade	6682
Média	4.046992
Desvio Padrão	5.434665
Valor mínimo	1
1^o quartil	1
2^o quartil	2
3^o quartil	4
Valor máximo	29

Table 1: Número de casos de dengue por logradouro.

6 Modelo de Aprendizado

Associou-se a cada logradouro o número de casos de dengue ocorridos. Utilizou-se Regressão Logística para treinar um modelo que classifique se um logradouro possui grande risco de futuros casos de dengue ou baixo risco e, assim, identificar através do p-value quais variáveis independentes que são cruciais para determinar se um logradouro possui o risco citado.

No projeto considerou-se como limiar de que um logradouro está em risco de casos de dengue ou não, o valor da mediana, 2 casos por logradouro, ou seja: acima de 2 casos por logradouro possui risco elevado e abaixo possui risco baixo. Como observado na tabela 1. Para a regressão utilizou-se: sklearn.linear_model.LogisticRegression.

7 Dificuldades

Uma das dificuldades para desenvolver este projeto foi a falta de dados no banco de dados abertos da prefeitura. Mesmo o *metadados* descrevendo algumas variáveis de usuário a maioria delas não existia em metade do dataframe.

Além de dificuldades para a organização e distribuição de tarefas entre os integrantes da equipe no começo do desenvolvimento do projeto.

Em relação ao modelo houve uma grande dificuldade em juntar um logradouro com o seu respectivo número de casos de dengue relatados, pois ambos eram oriundos de base de dados distintas (Logradouros e Casos de Dengue em 2016) e as variáveis que poderiam ser utilizadas para join estavam vazias ou corrompidas. Nos metadados de Casos de Dengue existem duas variáveis que associam-se com um logradouro: o co_logradouro_residencia(código numérico) e nome_logradouro_residencia(valor textual), co_logradouro_residencial estava vazio e nome_logradouro_residencia estava preenchido, mas apresentava uma série de problemas, dentre eles os caracteres como acentos estavam corrompidos e alguns nomes de logradouros estavam digitados incorretamente.

Para que a junção fosse possível as variáveis nome_logradouro_residencia, de Casos de Dengue em 2016, e nome_logradouro, Logradouros, foram manipuladas e formatadas ao máximo que fosse viável realizar uma comparação. Apesar de

tudo, apenas foi possível obter 6682 logradouros com quantidade de casos de dengue de um total de 20819 logradouros. Caso a qualidade dos dados da variável nome_logradouro_residencia de Casos de Dengue em 2016 fosse melhor formatada a quantidade de logradouros não analisados provavelmente seria menor.

8 Resultados

Como observado na tabela 2, o fato de um logradouro(rua, avenida, etc) não ser pavimentado ou possuir pavimentação do tipo paralelo, ser arborizado, possuir rede de esgoto ou possuir guias/sarjetas é um indicativo de que o mesmo possui grande chance de ocorrer casos de dengue. Outras features também apresentaram p-value abaixo de 0.05, ver tabela 2, não foram destacadas porque as citadas apresentaram p-value quase 0(zero), indicando quase 100% de certeza que as influências das features no surgimento de casos de dengue não é ao acaso.

Na tabela 3 é possível vizualizar as features que possuem p-value acima de 0.05 indicando que possivelmente a influência no número de casos de dengue foi ao acaso. Com destaque para outros tipos de pavimentação e coleta manual diária que apresentação aproximadamente 95% de chance da influencia ser ao acaso.

Pela tabela 4 foi possível verificar os resultados de desempenho do modelo treinado. Ele teve 0.634 de acurácia, 0.264 de sensitividade e 0.880 de especificidade.

Visto as análises do modelo de aprendizagem podemos inferir quais são os principais pontos de infraestrutura que os bairros precisam melhorar para diminuir a incidência de casos de Dengue. Não excluindo, logicamente, as atividades de conscientização e de educação contra o mosquito *Aedes*.

Além disso, podemos também caracterizar a distribuição dos diferentes tipos de dengue e os diferentes públicos que esta atinge. Podendo auxiliar os órgãos públicos competentes ao combate de doenças tropicais.

Feature	p-value
pavimentacao[T.PARALELO]	0.000031
pavimentacao[T.SEM PAVIMENTACAO]	0.000000
coleta[T.CONVENCIONAL DIARIA]	0.019077
limpezaurbana[T.PROGRAMADA SEMANAL]	0.003852
limpezaurbana[T.REGULAR ALTERNADA]	0.033415
limpezaurbana[T.SEM LIMPEZA PUBLICA]	0.007155
arborizacao	0.000009
redeagua	0.002577
codredesgoto	0.000000
guiasesarjetas	0.000073

Table 2: Features importantes para definir um logradouro com alto nível de ocorrência de casos de dengue.

Feature	p-value
pavimentacao[T.CONCRETO]	0.237264
pavimentacao[T.ESCADARIA]	0.113582
pavimentacao[T.OUTROS]	0.930368
pavimentacao[T.POLIEDRO]	0.718698
coleta[T.MANUAL ALTERNADA]	0.097581
coleta[T.MANUAL DIARIA]	0.947745
limpezaurbana[T.REGULAR DIARIA]	0.081583
galeriapluviais	0.105822

Table 3: Features que não importam para definir se uma rua pode ter alta ocorrência de dengue.

	P	F
P	706	394
F	96	141

Table 4: Tabela de contingência.

References

[1] Registro dos casos de Dengue, Zica e Chikungunya com registros nas unidades de saúde, públicas ou particulares. `http://dados.recife.pe.gov.br/dataset/casos-de-dengue-zika-e-chikungunya`, Dezembro, 2016.

[2] Dados sobre a área espacial da cidade do Recife, `http://dados.recife.pe.gov.br/dataset/area-urbana`, Dezembro, 2016.