Predição de Casos Endêmicos Mensais de Dengue

Caio Sampaio RA 10391053[1], **Guilherme Picoli RA 10389843**[1]

[1]Ciência da Computação
Faculdade de Computação e Informática
Universidade Presbiteriana Mackenzie
São Paulo – SP – Brasil

10391053@mackenzista.com.br, 10389843@mackenzista.com.br

Resumo. *Foi Identificada a subutilização dos dados de dengue na região da Grande São Paulo para previsão de áreas de foco baseadas nas temperaturas minímas, máximas, casos de dengue por população e volume de chuvas. Utilizando o* ***dataset*** *criado a partir de dados climáticos e infecciosos da doença dos últimos trés anos. Sendo assim sera criada um modelo de regressão múltipla para fazer a predição de casos endêmicos por população.*

1. Introdução

Na grande São Paulo foram registrados mais 35 mil casos dengue nos primeiros três meses de 2024 [Câmara Municipal de São Paulo 2024]. Todavia, trata-se de uma doença com ciclos endêmicos e epidêmicos no Brasil [Agência Fiocruz de Notícias 2013], sendo reportado casos anuais em São Paulo, mas há uma dificuldade em prever cidades com maior risco de contrair a doença e o número possível de casos que ocorrerão em meses seguintes. Com isso, será criada uma aplicação de inteligencia artificial para identificar a quantidade de casos endêmicos de dengue por habitante em municípios da grande São Paulo, com execução de São Lourenço da Serra e Salesópolis, baseados nos dados dos três anos anteriores, serão considerados os seguintes dados: índice pluviométrico do mês, temperatura mínima, média e máxima e incidência de dengue (casos confirmados da doença no município por habitante). Este projeto utilizará a opção *Framework* que empregar uma ferramenta de *Machine Learning* para solucionar um problema de regressão.

2. Descrição do Problema

O Governo Estadual de São Paulo mantém uma base de dados pública com a quantidade de casos notificados e confirmados (autóctones e importados) ao longo do ano, mas esses dados não são utilizados para estimar possíveis focos de dengue, tão pouco é combinado com outras informações relevantes a doença e sua transmissibilidade, como o clima e tamanho da população. Propõe-se criar uma inteligencia artificial que prevê o nível de infecção pela população da grande São Paulo, com execução de São Lourenço da Serra e Salesópolis (que não possuem notificação no período dos dados), nos casos endêmicos da dengue, utilizando como inputs o índice pluviométrico e as temperaturas minímas, máximas e médias. Serão descartados cenários epidêmicos, pois, além não seguirem a curva casos confirmados, busca-se obter uma ferramente de previsão de casos endêmicos para auxiliar na formulação de ações públicas para o combate da doença antes de atingir o ciclo epidêmico.

3. Dataset

O *Dataset* gerado para o projeto é a combinação dos dados públicos do Centro de Vigilância Epidemiológica (CVE) "Prof. Alexandre Vranjac"para a Dengue mensal nos anos de 2023, 2022 e 2021 dividido pela quantidade de habitantes do município (dados do IBGE) com os dados meteorológicos mensais, de temperatura mínima, média e máxima (ºC) e precipitação (mm), da companhia *Weather Spark* para os municípios de São Paulo, Arujá, Barueri, Biritiba Mirim, Caieiras, Cajamar, Carapicuíba, Cotia, Diadema, Embu, Embu-Guaçu, Ferraz de Vasconcelos, Francisco Morato, Franco da Rocha, Guararema, Guarulhos, Itapecerica da Serra, Itapevi, Itaquaquecetuba, Jandira, Juquitiba, Mairiporã, Mauá, Mogi das Cruzes, Osasco, Pirapora do Bom Jesus, Poá, Ribeirão Pires, Rio Grande da Serra, Santa Isabel, Santana do Parnaíba, Santo André, São Bernardo do Campo, São Caetano do Sul, Suzano, Taboão da Serra e Vargem Grande Paulista.

Para a preparação de dados, foram retirados todos os municípios que não fazem da Grande São Paulo, além de duas cidades que não possuíam notificações (São Lourenço da Serra e Salesópolis), os dados de casos notificados e os dados totais (do ano). Posteriormente, os dados de casos confirmados autóctones e importados foram somados em uma única coluna infeccoes) e divididos pelo o número absoluto de habitantes do município do respectivo ano, formando a coluna de incidência (infe_pop). Além disso, foram criadas colunas para cada cidade, sendo o dado originada de uma determinada localidade recebendo o valor 1 na respectiva coluna da cidade e 0 nas demais colunas. Ademais, os dados do ano, do mês, do número de infecções, do tamanho da população, das colunas das cidades, de incidência de dengue e os dados de temperatura mínima, média e máxima (ºC) e precipitação (mm) da Grande São Paulo dos últimos três anos foram agregados em um único arquivo CSV (*dataset*) com as seguintes colunas (atributos): cidade; ano; mes; infeccoes; populacao; infe_pop; max; media; min; ind_pluv; e as colunas de todas as cidades da Grande São Paulo.

Na análise exploratória, identificamos as seguintes informações: Tipo das Colunas (Figura 1); Tipo das Colunas (Figura 2); Tamanho do Conjunto de Dados (Figura 3); Heatmap (Figura 4); Skew de Cada Atributo (Figura 5); Kurtosis de Todos os Atributos (Figura 6); Gráfico de Pontos entre Infecções e População (Figura 7); Histograma de Infecções com a Melhor Curva Associada (Figura 8).

```
cidade                    object
ano                        int64
mes                        int64
infeccoes                  int64
populacao                  int64
infe_pop                 float64
max                        int64
media                      int64
min                        int64
ind_pluv                 float64
 Arujá                     int64
 Barueri                   int64
 Biritiba Mirim            int64
 Caieiras                  int64
 Cajamar                   int64
 Carapicuíba               int64
 Cotia                     int64
 Diadema                   int64
 Embu                      int64
 Embu-Guaçu                int64
 Ferraz de Vasconcelos     int64
 Francisco Morato          int64
 Franco da Rocha           int64
 Guararema                 int64
 Guarulhos                 int64
 Itapecerica da Serra      int64
 Itapevi                   int64
 Itaquaquecetuba           int64
 Jandira                   int64
 Juquitiba                 int64
 Mairiporã                 int64
 Mauá                      int64
 Mogi das Cruzes           int64
 Osasco                    int64
 Pirapora do Bom Jesus     int64
 Poá                       int64
 Ribeirão Pires            int64
 Rio Grande da Serra       int64
 Santa Isabel              int64
 Santana do Parnaíba       int64
 Santo André               int64
 São Bernardo do Campo     int64
 São Caetano do Sul        int64
 Suzano                    int64
 Taboão da Serra           int64
São Paulo                  int64
Vargem Grande Paulista     int64
dtype: object
```

Figura 1. Tipo das Colunas

```
cidade                    1332
ano                       1332
mes                       1332
infeccoes                 1332
populacao                 1332
infe_pop                  1332
max                       1332
media                     1332
min                       1332
ind_pluv                  1332
 Arujá                    1332
 Barueri                  1332
 Biritiba Mirim           1332
 Caieiras                 1332
 Cajamar                  1332
 Carapicuíba              1332
 Cotia                    1332
 Diadema                  1332
 Embu                     1332
 Embu-Guaçu               1332
 Ferraz de Vasconcelos    1332
 Francisco Morato         1332
 Franco da Rocha          1332
 Guararema                1332
 Guarulhos                1332
 Itapecerica da Serra     1332
 Itapevi                  1332
 Itaquaquecetuba          1332
 Jandira                  1332
 Juquitiba                1332
 Mairiporã                1332
 Mauá                     1332
 Mogi das Cruzes          1332
 Osasco                   1332
 Pirapora do Bom Jesus    1332
 Poá                      1332
 Ribeirão Pires           1332
 Rio Grande da Serra      1332
 Santa Isabel             1332
 Santana do Parnaíba      1332
 Santo André              1332
 São Bernardo do Campo    1332
 São Caetano do Sul       1332
 Suzano                   1332
 Taboão da Serra          1332
São Paulo                 1332
Vargem Grande Paulista    1332
dtype: int64
```

Figura 2. Tamanho do Conjunto de Dados

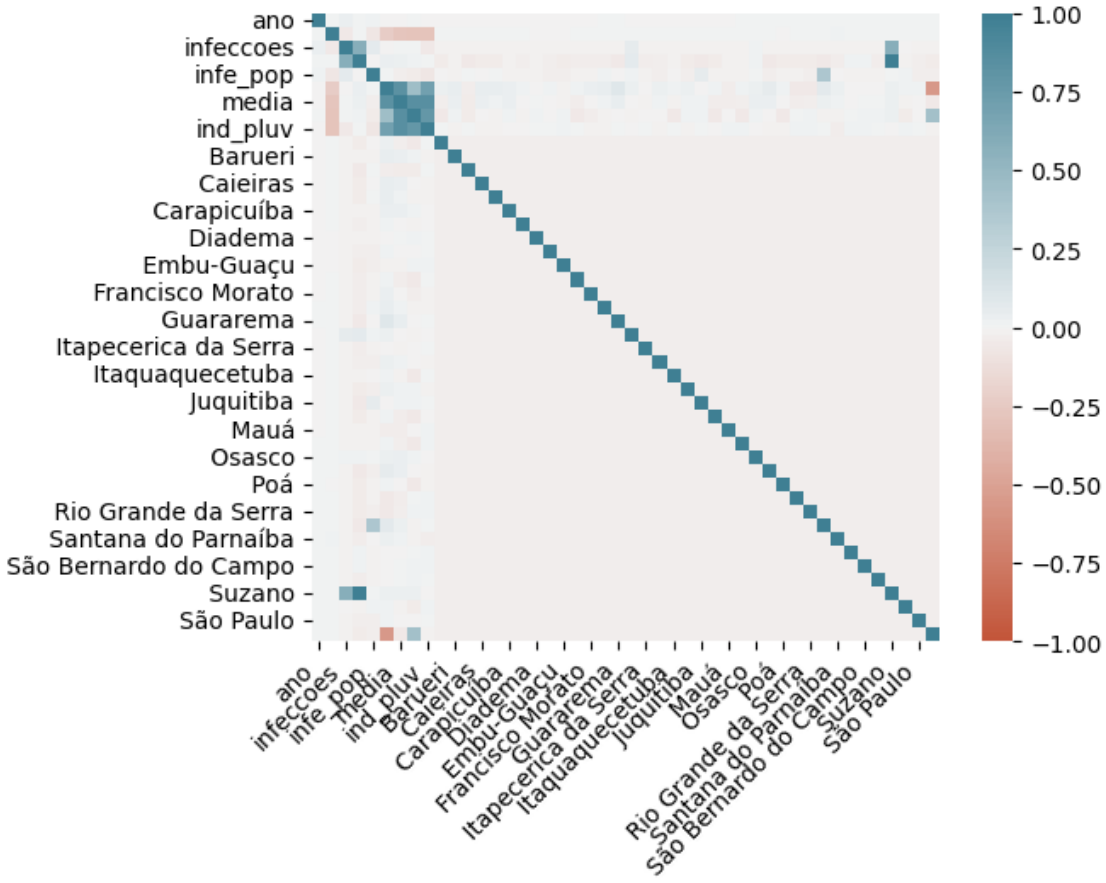

Figura 3. Heatmap

```
ano                          0.000000
mes                          0.000000
infeccoes                   13.126046
populacao                    5.654799
infe_pop                    17.681826
max                         -1.116474
media                       -0.107968
min                          0.271055
ind_pluv                     0.440467
 Arujá                       5.839912
 Barueri                     5.839912
 Biritiba Mirim              5.839912
 Caieiras                    5.839912
 Cajamar                     5.839912
 Carapicuíba                 5.839912
 Cotia                       5.839912
 Diadema                     5.839912
 Embu                        5.839912
 Embu-Guaçu                  5.839912
 Ferraz de Vasconcelos       5.839912
 Francisco Morato            5.839912
 Franco da Rocha             5.839912
 Guararema                   5.839912
 Guarulhos                   5.839912
 Itapecerica da Serra        5.839912
 Itapevi                     5.839912
 Itaquaquecetuba             5.839912
 Jandira                     5.839912
 Juquitiba                   5.839912
 Mairiporã                   5.839912
 Mauá                        5.839912
 Mogi das Cruzes             5.839912
 Osasco                      5.839912
 Pirapora do Bom Jesus       5.839912
 Poá                         5.839912
 Ribeirão Pires              5.839912
 Rio Grande da Serra         5.839912
 Santa Isabel                5.839912
 Santana do Parnaíba         6.121554
 Santo André                 5.839912
 São Bernardo do Campo       5.839912
 São Caetano do Sul          5.839912
 Suzano                      5.839912
 Taboão da Serra             5.839912
São Paulo                    5.839912
Vargem Grande Paulista       5.839912
dtype: float64
```

Figura 4. Skew de Cada Atributo

```
ano                          -1.501128
mes                          -1.216845
infeccoes                   189.861585
populacao                    30.710160
infe_pop                    390.732215
max                           2.897895
media                        -1.290151
min                          -0.011378
ind_pluv                     -1.064591
 Arujá                       32.152846
 Barueri                     32.152846
 Biritiba Mirim              32.152846
 Caieiras                    32.152846
 Cajamar                     32.152846
 Carapicuíba                 32.152846
 Cotia                       32.152846
 Diadema                     32.152846
 Embu                        32.152846
 Embu-Guaçu                  32.152846
 Ferraz de Vasconcelos       32.152846
 Francisco Morato            32.152846
 Franco da Rocha             32.152846
 Guararema                   32.152846
 Guarulhos                   32.152846
 Itapecerica da Serra        32.152846
 Itapevi                     32.152846
 Itaquaquecetuba             32.152846
 Jandira                     32.152846
 Juquitiba                   32.152846
 Mairiporã                   32.152846
 Mauá                        32.152846
 Mogi das Cruzes             32.152846
 Osasco                      32.152846
 Pirapora do Bom Jesus       32.152846
 Poá                         32.152846
 Ribeirão Pires              32.152846
 Rio Grande da Serra         32.152846
 Santa Isabel                32.152846
 Santana do Parnaíba         35.526760
 Santo André                 32.152846
 São Bernardo do Campo       32.152846
 São Caetano do Sul          32.152846
 Suzano                      32.152846
 Taboão da Serra             32.152846
São Paulo                    32.152846
Vargem Grande Paulista       32.152846
dtype: float64
```

Figura 5. Kurtosis de Todos os Atributos

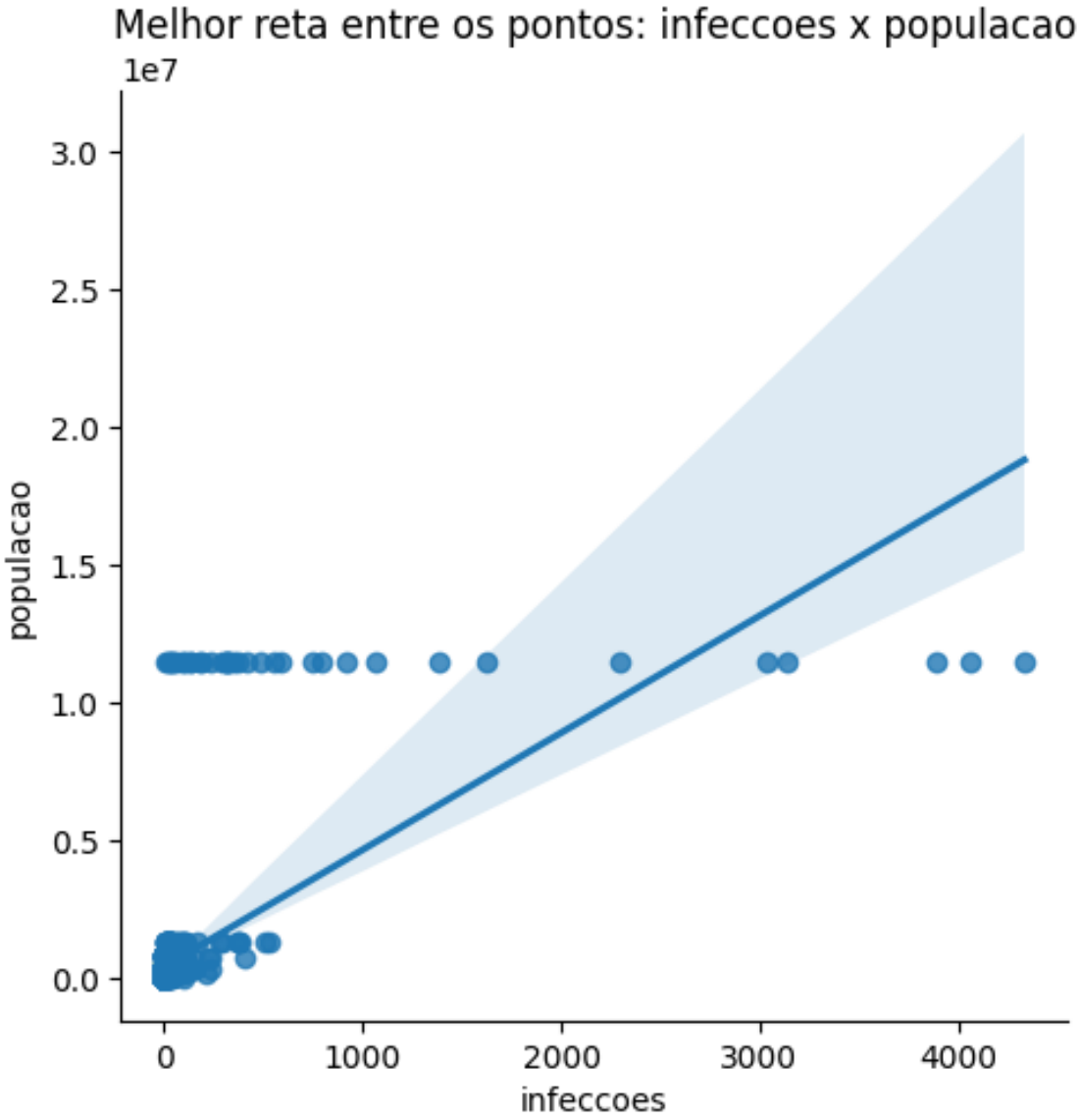

Figura 6. Gráfico de Pontos entre Infecções e População

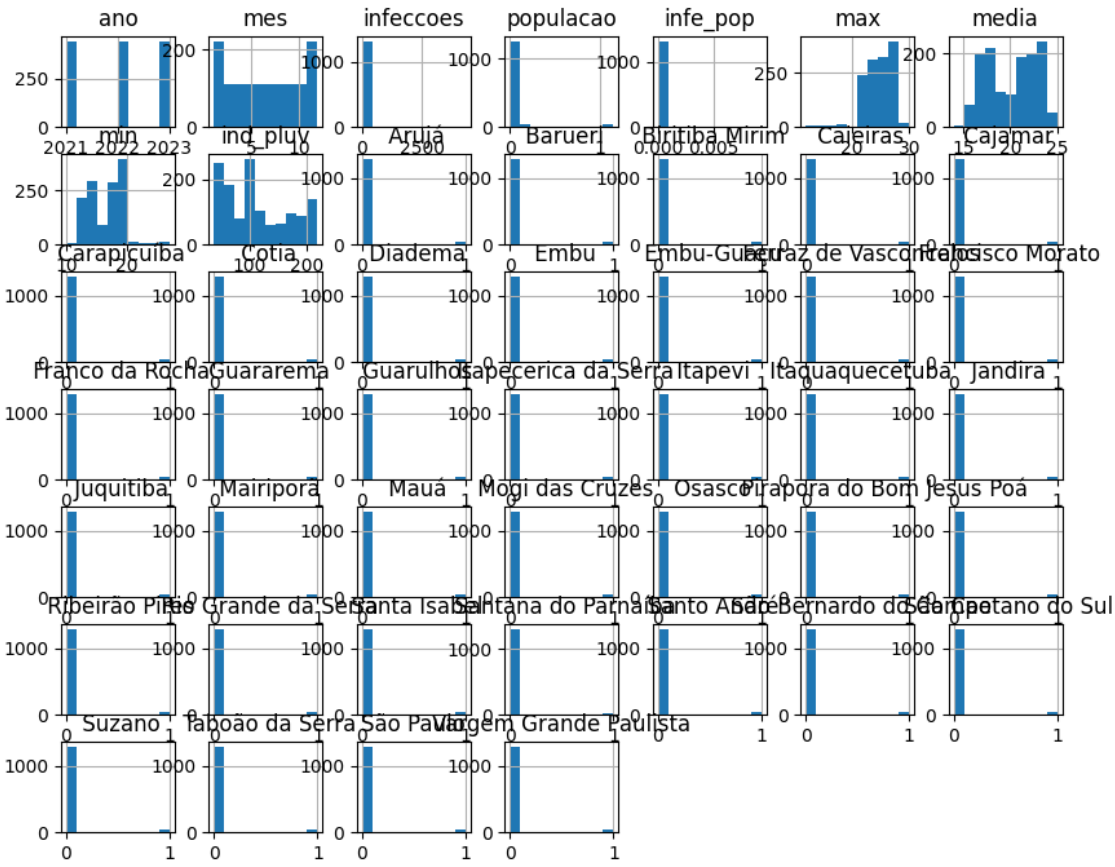

Figura 7. Histograma de Todos os Atributos

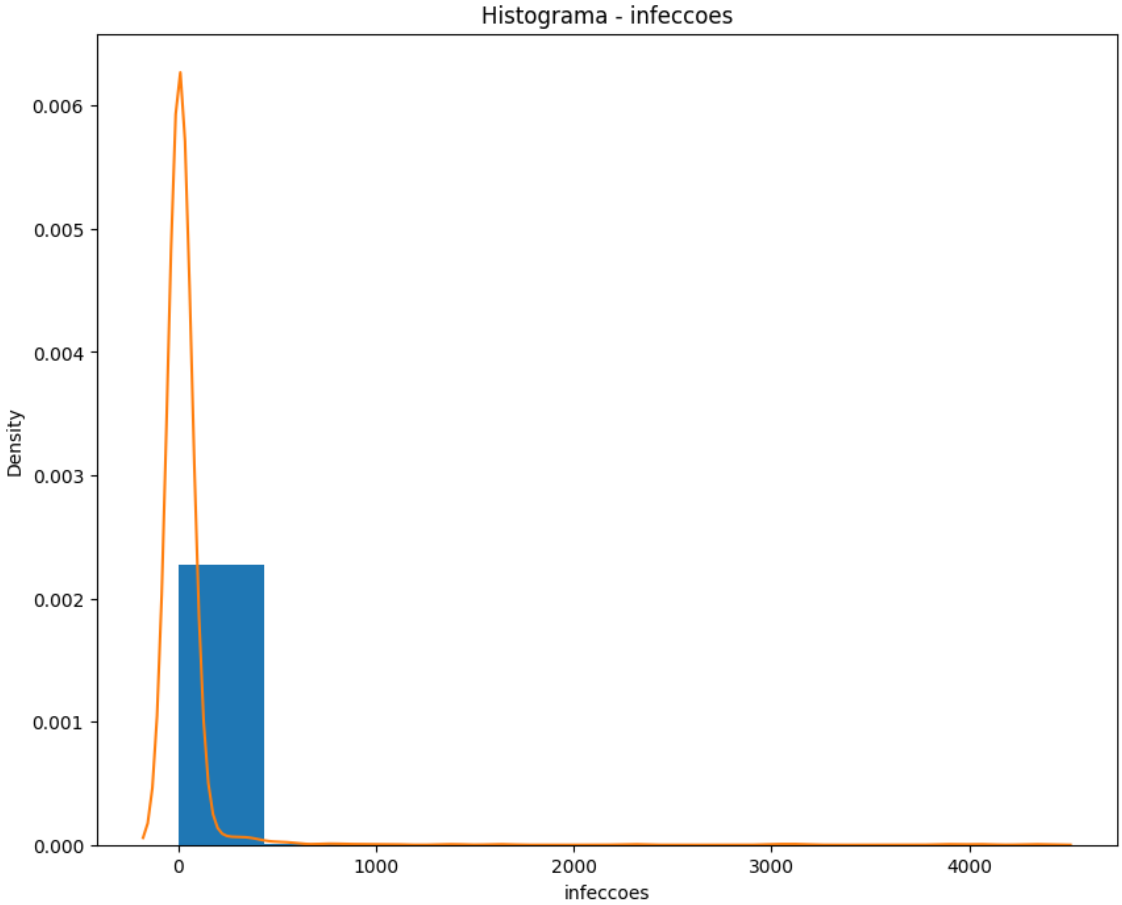

Figura 8. Histograma de Infecções com a Melhor Curva Associada

sk = 13.126046171264715, Ck = 189.8615846142338

Média=35.250750750750754, Mediana=2.0, Moda=0 0

Name: infeccoes, dtype: int6

4. Metodologia

Visto que busca-se prever a incidência de dengue endêmica na Grande São Paulo, foi empregado a Regressão Linear Múltipla para o treinamento do modelo, pois os dados obtidos são numéricos e com várias variáveis. Foram definidos três passos para a geração do modelo: análise exploratória, necessária para compreensão do dados e identificar possíveis demandas de limpeza de dados; análise das variáveis categóricas, para a verificação de quais as melhores variáveis, e suas combinações, para a predição; e a geração e verificação do modelo usando a Regressão Linear.

Inicialmente, foi realizado a Análise Exploratória uma melhor compreensão do dados. Isso nos forneceu informações importantes, como a qualidade do dados: há uma grande quantidade de dados com poucos número de casos de infeções em cidades com baixa população e uma quantidade considerável de dados com alto número de casos de infeções em cidades com grande população, porém há poucos dados entre os dois extremos. Outra informação importante observada foi com relação ao atributo infe_pop (índice de infecção por população) que apresentou um alta dispersão (*kurtosis*) dos valores, assim, decidimos não utilizar essa métrica.

Posteriormente, foi realizado a Análise das Variáveis Categóricas para determinar a melhor combinação das variáveis para a predição do modelo. Foram realizados três testes de combinações: o primeiro com a variável populacao (população) para prever infeccoes (infecções), que resultou em um R^2 de 0.340; o segundo com as variáveis populacao, mes, max, media, min e ind_pluv para prever infeccoes, que resultou em um R^2 de 0.350; o terceiro utilizou as variáveis populacao, mes, max, media, min, ind_pluv, Aruja, Barueri, Biritiba_Mirim, Caieiras, Cajamar, Carapicuiba, Cotia, Diadema, Embu, EmbuGuacu, Ferraz_de_Vasconcelos, Francisco_Morato, Franco_da_Rocha, Guararema, Guarulhos, Itapecerica_da_Serra, Itapevi, Itaquaquecetuba, Jandira, Juquitiba, Mairipora, Maua, Mogi_das_Cruzes, Osasco, Pirapora_do_Bom_Jesus, Poa, Ribeirao_Pires, Rio_Grande_da_Serra, Santa_Isabel, Santana_do_Parnaiba, Santo_Andre, Sao_Bernardo_do_Campo, Sao_Caetano_do_Sul, Suzano, Taboao_da_Serra, Sao_Paulo e Vargem_Grande_Paulista para prever infeccoes, que resultou em um R^2 de 0.354. Desse modo, verificamos que o acréscimo das variáveis de pertencimento à cidade (onde 1 é pertence e 0 é não pertence) não contribuem expressivamente a melhora do modelo. Assim, decidi-se utilizar a combinação do segundo teste.

Por fim, foi gerado o modelo com a Regressão Linear, separando da base de dados em 35% para teste e 65% para o treinamento do modelo. O treinamento, usando o *LinearRegression()* do *sklearn*, resultou em um modelo com um R^2 de 0.34106. O resultado das observações pelos valores previstos dos dados de teste pode ser observado no gráfico da figura que segue (Figura 14).

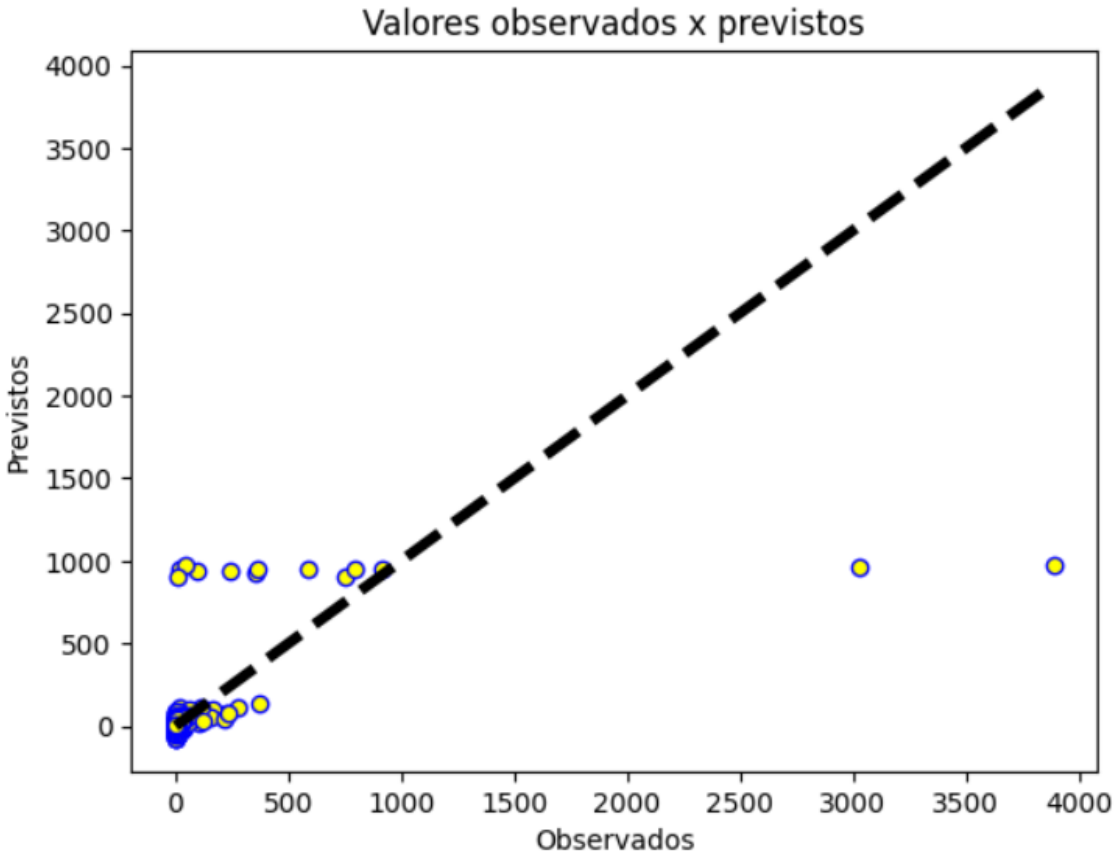

Figura 9. Valores Observados x Previstos

Dessa forma, foram retirados dos dados todos os atributos não utilizados no modelo e gerou-se um novo *dataset* após a limpeza. Com o novo *dataset*, treinou-se e persistiu-se um modelo utilizando a Regressão Linear.

5. Resultados

Após a limpeza dos dados e adequação do modelo, é possível prever a quantidade de casos de dengue na Grande São Paulo ao fornecer o mês, a população, a temperatura máxima (ºC), média (ºC) e mínima (ºC) e precipitação (mm). Assim, foram realizados cinco testes com dados de casos de dengue dos anos de 2018 a 2019. Cada dado de teste foi retirado de uma cidade diferente, com o objetivo de variar o tamanho das populações.

Teste 1:

```
[34] dadosTeste = pd.DataFrame({'mes': [5], 'populacao': [11451999], 'max': [32], 'media': [27], 'min': [20], 'ind_pluv': [108]})
     infeccoesEstimado = modelo_regressor_final.predict(dadosTeste)[0]

     print(f'{infeccoesEstimado} infectados')

1013.5332151744996 infectados
```

- Valor esperado: 1023 infectados
- Valor predito: 1013.53 => 1014

Erro = | 1023 - 1014 | = 9

Figura 10. Teste 1

No Teste 1 (Figura 9), o resultado esperado era de 1023, mas o valor predito foi de 1014 infectados. Assim, o erro foi de 9 infectados na predição.

Teste 2:

```
[35] dadosTeste = pd.DataFrame({'mes': [7], 'populacao': [810729], 'max': [29], 'media': [26], 'min': [24], 'ind_pluv': [112]})
     infeccoesEstimado = modelo_regressor_final.predict(dadosTeste)[0]

     print(f'{infeccoesEstimado} infectados')

117.42141827265904 infectados
```

- Valor esperado: 104 infectados
- Valor predito: 117 infectados

Erro = | 104 - 117 | = 13

Figura 11. Teste 2

No Teste 2 (Figura 10), o resultado esperado era de 104, porém o valor predito foi de 117 infectados. Desse modo, o erro foi de 13 infectados na predição.

Teste 3:

```
[50] dadosTeste = pd.DataFrame({'mes': [4], 'populacao': [1291771], 'max': [24], 'media': [21], 'min': [18], 'ind_pluv': [43]})
     infeccoesEstimado = modelo_regressor_final.predict(dadosTeste)[0]

     print(f'{infeccoesEstimado} infectados')

165.58967956220266 infectados
```

- Valor esperado: 178 infectados
- Valor predito: 166 infectados

Erro = | 178 - 166 | = 13

Figura 12. Teste 3

No Teste 3 (Figura 11), o resultado esperado era de 178, mas o valor predito foi de 166 infectados. Assim, o erro também foi de 13 infectados na predição.

Teste 4:

```
[51] dadosTeste = pd.DataFrame({'mes': [6], 'populacao': [451505], 'max': [32], 'media': [16], 'min': [13], 'ind_pluv': [57]})
     infeccoesEstimado = modelo_regressor_final.predict(dadosTeste)[0]

     print(f'{infeccoesEstimado} infectados')

8.766917348026084 infectados
```

- Valor esperado: 17 infectados
- Valor predito: 9 infectados

Erro = | 17 - 9 | = 8

Figura 13. Teste 4

Já no Teste 4 (Figura 12), o resultado esperado era de 17 infectados, todavia o valor predito foi de 9 infectados, representando um erro de 8 infectados na predição.

Teste 5:

```
[52] dadosTeste = pd.DataFrame({'mes': [4], 'populacao': [158522], 'max': [25], 'media': [21], 'min': [18], 'ind_pluv': [93]})
     infeccoesEstimado = modelo_regressor_final.predict(dadosTeste)[0]

     print(f'{infeccoesEstimado} infectados')

40.070122312684234 infectados
```

- Valor esperado: 34 infectados
- Valor predito: 40 infectados

Erro = | 34 - 40 | = 6

Figura 14. Teste 5

Por último, o Teste 5 (Figura 13) resultou na predição de 40 infectados, quando o valor esperado era de 34 infectados, o que representa um erro de 6 infectados na predição.

6. Conclusão

Foi identificado que os atributos ano, infe_pop e as colunas de cidades não contribuíram para a criação do modelo, sendo retiradas, passou-se a utilizar as informações mes, infeccoes, populacao, max, media, min, ind_pluv, mostrando ter um maior peso no momento do treinamento.

Utilizando os reunidos para comprovar a acurácia da inteligencia artificial, pode-se concluir que o modelo de Regressão Linear criado fornece predições satisfatórias para a quantidade de casos de dengue dada as informações necessárias, podendo ser usada para estimar dados futuros de infectados nas regiões da Grande São Paulo, mas considerando a falta de dados nas proximidades da média de casos, o modelo poderia ser retreinado com novos dados caso existissem, identificamos uma sub notificação por parte de municípios menores, tando que foram retirados São Lourenço da Serra e Salesópolis por não apresentarem dados durante os anos de 2021, 2022 e 2023. Caso o estado de São Paulo apresente dados mais consistentes, a geração de um modelo preditivo mais acretivo não se torna viável.

7. Endereço GitHub

O projeto em *Jupyter Notebook*, o modelo e o *dataset* encontram-se no seguinte repositório do *GitHub*: https://github.com/MrPicoli1/Projeto-N2-07N-Predi-o-de-Casos-End-micos-Mensais-de-Dengue-Caio-Sampaio-e-Guilherme-Picoli

8. Bibliografia

IBGE. Panorama: Brasil / São Paulo. Disponível em: https://cidades.ibge.gov.br/brasil/sp/ . Acesso em: 30 de março de 2024. IBGE, 2023. Página Web on-line.

Weather Spark. O clima de qualquer lugar da Terra durante o ano inteiro. Disponível em: https:pt.weatherspark.com/ . Acesso em: 30 de março de 2024. Weather Spark, 2024. Página Web on-line.

Centro de Vigilância Epidemiológica "Prof. Alexandre Vranjac". Dados Estatísticos Dengue. Disponível em: https://saude.sp.gov.br/cve-centro-de-vigilancia-epidemiologica-prof.-alexandre-vranjac/oldzoonoses/dengue/dados-estatisticos . Acesso em: 30 de março de 2024. Centro de Vigilância Epidemiológica, 2024. Página Web on-line.

Referências

Agência Fiocruz de Notícias (2013). Dengue.

Câmara Municipal de São Paulo (11 de março de 2024). Capital registra mais de 35 mil casos de dengue; zona norte é a região com o maior número de transmissões da doença.