# SRAG: Agrupamento e Detecção de Anomalias

**Endhel Lopes de Freitas**
**Celso Henrique Assis Silva**
**Manoel Farias Paixão Júnior**

1

## 1. Introdução

O seguinte relatório tem como objetivo aplicar técnicas de mineração de dados para encontrar grupos e detectar anomalias em uma base de dados sobre casos de síndrome respiratória aguda grave no ano de 2020, disponibilizada pelo governo.

Para tanto, foi coletada uma amostra da base de dados, de 1000 registros aleatórios. Os dados foram pré-processados e em seguida inseridos nos modelos de Machine Learning, KMeans para agrupamento, e KNN para detecção de anomalias. Por último, foram feitas algumas visualizações e conclusões sobre os resultados.

## 2. Desenvolvimento

O projeto foi realizado utilizando a ferramenta Jupyter Notebook, a linguagem de programação Python, as bibliotecas Pandas, Numpy e Statistics para manipulação dos dados, Matplotlib para vizsualização dos dados e a biblioteca Scikit-Learn para aplicação dos modelos de Machine Learning.

### 2.1. Importação dos Dados

A base disponibilizada pelo governo é muito extensa, e por questões de performance e memória, foi coletado uma amostragem de 1000 linhas para ser utilizada no projeto.

O arquivo de amostragem é do tipo csv, e a importação dos dados foi feita utilizando a biblioteca Pandas, e as dimensões do dataframe resultante são de 1000 linhas e 154 colunas.

### 2.2. Seleção dos Dados

Como a base de dados possui muitas variáveis irrelevantes, foram selecionadas apenas as colunas que serão necessárias para detectar anomalias e fazer a clusterização. A seguir, as colunas que serão utilizadas posteriormente:

- NU_IDADE_N - Idade do Paciente
- SG_UF - UF da residência do Paciente
- SURTO_SG - Se o caso é proveniente de SG
- NOSOCOMIAL - Caso de SRAG com infecção adquirida após internação.
- AVE_SUINO - Caso com contato direto com aves ou suínos.
- FEBRE - Se o Paciente apresentou febre
- TOSSE - Se o Paciente apresentou tosse
- GARGANTA - Se o Paciente apresentou dor de garganta

- DISPNEIA - Se o Paciente apresentou dispneia
- DESC_RESP - Se o Paciente apresentou desconforto respiratório
- SATURACAO - Se o Paciente apresentou saturação O2 ¡ 95
- DIARREIA - Se o Paciente apresentou diarreia
- VOMITO - Se o Paciente apresentou vômito
- DOR_ABD - Se o Paciente apresentou dor abdominal
- FADIGA - Se o Paciente apresentou fadiga
- PERD_OLFT - Se o Paciente apresentou perda de olfato
- PERD_PALA - Se o Paciente apresentou perda de paladar
- OUTRO_SIN - Se o Paciente apresentou outros sintomas
- PUERPERA - Se o Paciente era puérpera ou parturiente
- FATOR_RISC - Se o Paciente apresentou algum fator de risco
- CARDIOPATI - Se o Paciente possuia Doença Cardiovascular Crônica
- HEMATOLOGI - Se o Paciente possuia Doença Hematológica Crônica
- SIND_DOWN - Se o Paciente possuia Síndrome de Down
- HEPATICA - Se o Paciente possuia Doença Hepática Crônica
- ASMA - Se o Paciente possuia asma
- DIABETES - Se o Paciente possuia diabetes
- NEUROLOGIC - Se o Paciente possuia Doença Neurológica
- PNEUMOPATI - Paciente possui outra pneumopatia crônica
- IMUNODEPRE - Se o Paciente possuia Imunodeficiência ou Imunodepressão
- RENAL - Se o Paciente possuia Doença Renal Crônica
- OBESIDADE - Se o Paciente possuia obesidade
- OUT_MORBI - Se o Paciente possuia outro fator de risco
- VACINA - Se o Paciente foi vacinado contra a gripe na última campanha
- ANTIVIRAL - Se o Paciente usou antiviral para a gripe
- HOSPITAL - Se o Paciente foi internado
- SG_UF_INTE - UF de internação do paciente
- UTI - Se o Paciente foi internado na UTI
- RAIOX_RES - Resultado do Raio X de tórax
- CLASSI_FIN - Diagnóstico final do caso.
- EVOLUCAO - Evolução do caso
- ID_MN_RESI - Município de residência do paciente

- ID_MN_INTE - Município de internação do paciente

- SEM_PRI - Semana Epidemiológica dos primeiros sintomas do paciente.

- CS_RACA - Raça do paciente

O restante das colunas foram excluídas do dataframe para evitar problemas aos modelos.

### 2.3. Pré-Processamento dos Dados

Antes de aplicar técnicas de mineração de dados, é fundamental preparar e limpar os dados para garantir a eficiência dos algoritmos.

#### 2.3.1. Preenchimento de valores faltantes

Muitas colunas possuiam muitos valores faltantes, e para resolver este problema foi utilizada a função ”fillna”do Pandas. Os dados inseridos são a Moda de cada coluna. A escolha desta medida de centralidade se deve ao fato de os dados representarem valores categóricos,logo, a média aritmética traria dados insignificantes e a mediana provavelmente um valor NAN.

#### 2.3.2. Criação de Arquivos CSV

Foram criados vários arquivos csv, para diferentes propósitos, tanto para a clusterização, quanto para a detecção de anomalias.

Os datasets são: Um para estados, contendo o número de casos e internações; Dois para municípios, um com o número de casos e o outro com o número de internações; Um contendo o número de casos em cada semana epidemiológica; Um com o número de casos para cada classificação final; Um contendo o número de pacientes que tiveram cada sintoma; Um com o número de pacientes que tinham cada fator de risco no momento da doença; Um com as idades de todos os pacientes; Um com a quantidade de pacientes para cada raça; E para pacientes, com sintomas, fatores de risco, etc.

#### 2.3.3. Normaliação

As variáveis estavam em escalas diferentes, então os dados foram normalizados utilizando a técnica de MinMax.

### 2.4. Aplicação dos Modelos

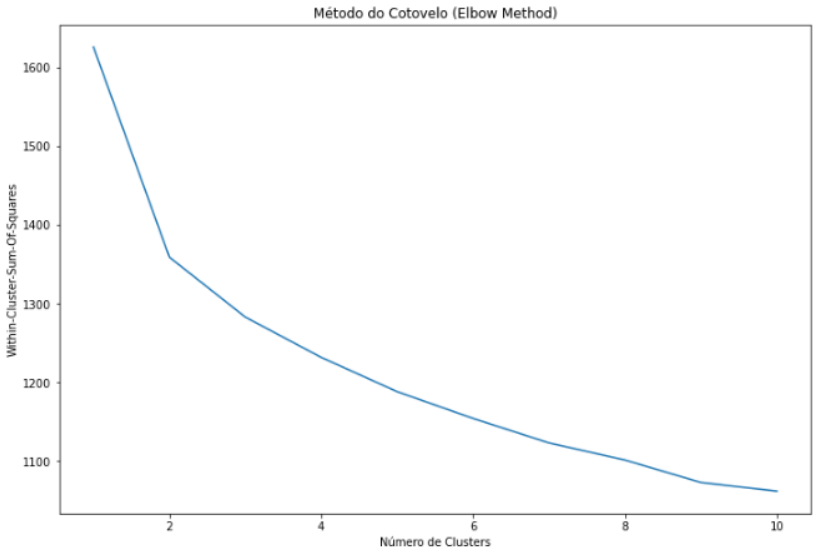
Com os dados preparados, foram aplicados os modelos de clusterização com KMeans, e detecção de anomalias com KNN.

#### 2.4.1. Clusterização

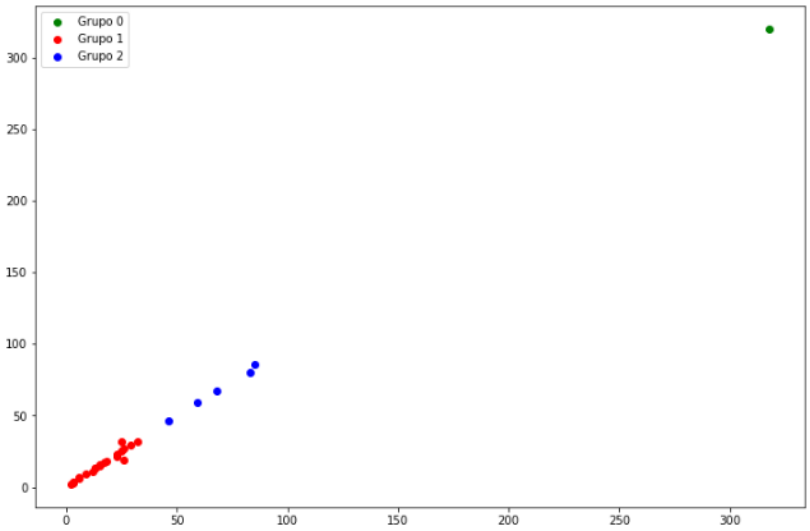
O algoritmo KMeans dependende de um valor K, que representa em quantos clusters serão divididos os dados. Para descobrir quantos usar, foi utilizado uma heurística chamada

Elbow Method (Método do Cotovelo), onde é coletado o erro de agrupamento para cada valor de K e adicionado a uma lista. Em seguida ela é plotada em um gráfico, e o formato que ela produz se parece com um braço, e o valor de K ideal está alinhado ao cotovelo (local onde o ângulo é menor que os demais pontos).

O primeiro agrupamento realizado é o de pacientes. O valor ideal de K encontrado é 2. Portanto, os pacientes foram divididos em dois grupos, onde um possui 625 pacientes, e o outro, 375. As principais características que diferem os dois grupos são a idade do paciente e o diagnóstico final do caso.




O segundo agrupamento foi o de estados mais afetados pela SRAG. As variáveis utilizadas foram o número de casos e o número de internações em cada um deles. O Elbow Method indicou que o valor de K deve ser o 3. Após a aplicação do modelo, foram separados 1 estado para o primeiro grupo, 21 para o segundo e 5 para o terceiro. Isto fica mais claro no gráfico a seguir:




O eixo x representa o número de casos, e o y, o número de internações. Destaque para o grupo 0, que contém apenas o estado de São Paulo, e está muito distante do restante.

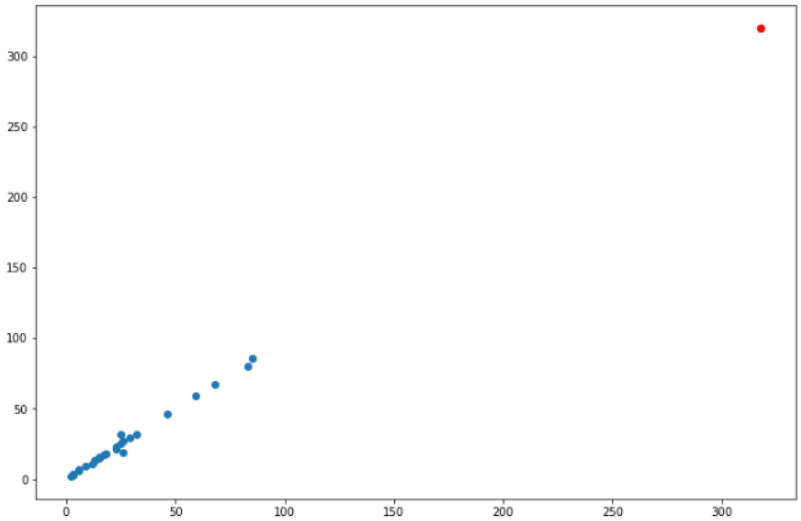
É interessante notar que a maior cidade do país, com o maior número de circulações de pessoas, também é a que mais sofre com esta doença.

### 2.4.2. Detecção de Anomalias

Para encontrar registros anormais na base de dados foi utilizado o algoritmo KNN (K-Nearest Neighbors). O valor utilizado para os 10 casos encontrados foi 3.

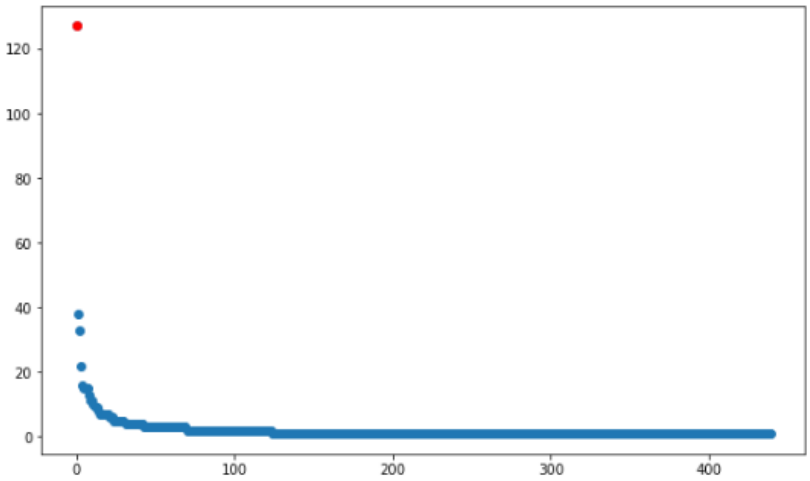
O primeira anomalia foi sobre os dados de pacientes. Infelizmente, o resultado encontrado se refere ao paciente que teve a maioria dos campos do registro ignorados, impedindo uma análise mais interessante sobre os dados.

Na segunda detecção, utilizando novamente os dados sobre os estados, o estado de São Paulo foi, obviamente, o mais anômalo. Visto que, no modelo de agrupamento ficou claro a discrepância deste estado para com os outros, este resultado já era esperado.



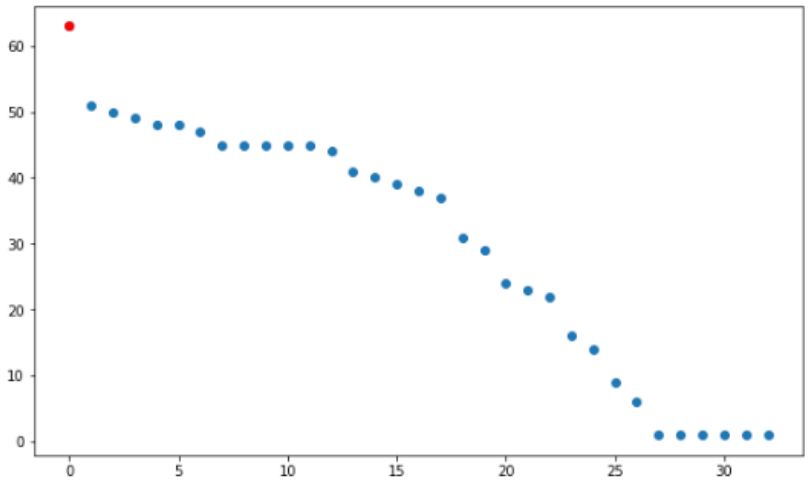

Neste gráfico fica mais claro a anomalia encontrada.

Nas próximas duas aplicações do modelo, onde foram analisados os municípios mais afetados, tanto de casos quanto de internações, o resultado encontrado é similar ao anterior, uma vez que, como o estado de São Paulo foi absurdamente afetado, provavelmente o município mais afetado, para as duas características, está localizado lá. Para ambas, o município de São Paulo se sobresaiu.



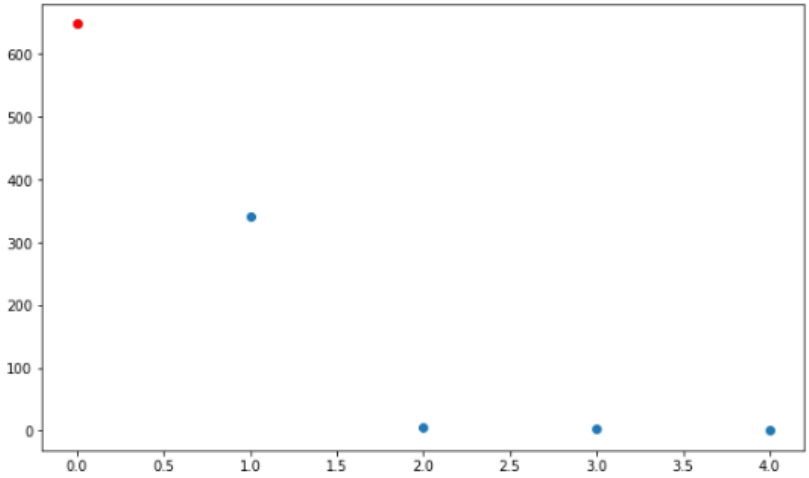

Este gráfico remete ao município com maior número de casos (São Paulo). Percebe-se que dos 1000 pacientes coletados na amostragem, 127 são residentes de lá.

Outra detecção de anomalia realizada buscou descobrir qual a semana epidemiológica mais anormal, em se tratando de casos de SRAG.



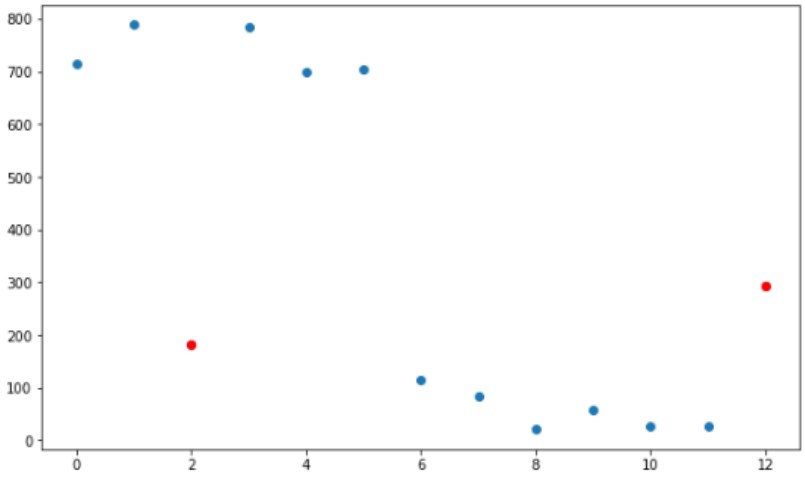

A semana encontrada é a 29. Observar o gráfico pode causar uma confusão ao entendimento do leitor, então só para esclarecer, a semana 29 estava no índice 0 do dataset pois no ato da coleta, foi utilizado a função ”value_counts”do Pandas, que retorna a contagem de cada valor único da coluna e em ordem decrescente. Nesta semana houveram 63 casos de SRAG, lembrando, é claro, que estamos utilizando uma amostra.

A sexta detecção de anomalia utilizou dados dos diagnósticos finais dos casos. O resultado encontrado se refere ao diagnóstico com mais ocorrências no ano de 2020, agravamento causado por COVID-19. Este é só mais um indicativo, dentre muitos outros, da gravidade do coronavírus, e da taxa altíssima de transmissão nestes últimos meses.

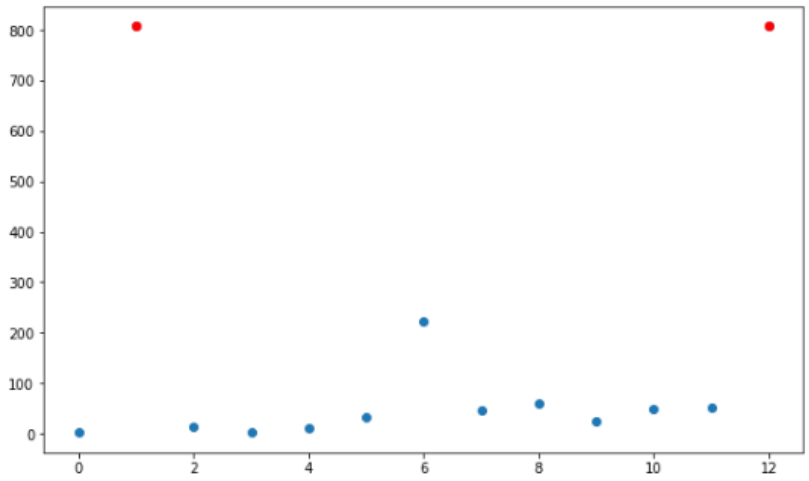



Novamente, como explicado anteriormente, o mais anômalo se encontra no índice 0.

As próximas duas aplicações, se referem a quantidade de ocorrências de sintomas e fatores de risco para os mil pacientes da amostra.

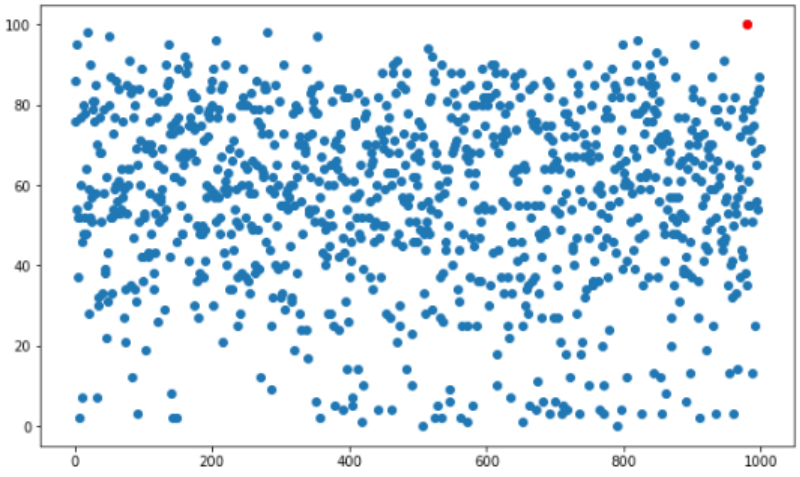



Este gráfico mostra que o mais anômalo, é o índice 12 (outros sintomas). Mas para observar um dos sintomas comuns, foi detectado também a dor de garganta como anormal na base de dados.



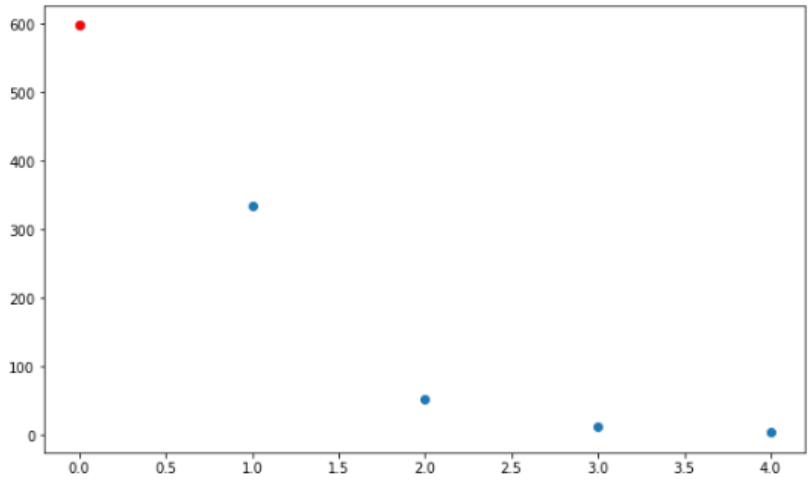

Já este gráfico mostrou que os dois mais anômalos são os que possuem o maior número de ocorrências. Coincidentemente, ambos possuem o mesmo número de casos, 807, e se referem a Doença Cardiovascular Crônica e Outros Fatores.

Outra anomalia encontrada foi sobre a idade de todos os pacientes da amostragem.




O ponto em vermelho é o mais distante dos outros, e representa um paciente que possuia 100 anos na época da coleta dos dados. Realmente incomum!

A última detecção de anomalias foi aplicado em cima da raça dos pacientes. Foram utilizados a quantidade de pacientes com cada raça.




O dado mais anormal encontrado se refere a raça parda, o que é bastante significativo para um país como o Brasil que possui uma das populações mais miscigenadas do mundo.

## 3. Conclusão

Por meio deste trabalho, foi possível aprender mais sobre sobre o processo de descoberta do conhecimento, principalmente sobre as fases de pré-processamento dos dados e mineração de dados. Além disso, concluímos que um projeto na área de dados é muito mais do que aplicar técnicas e algoritmos sobre os dados. É necessário também fazer uma análise crítica acima dos dados tanto na seleção, quanto no pós-processamento, para que seja possível interpretar e encontrar significados sobre eles, afinal, se trata de descobrir conhecimento, e isso nunca foi trivial.