Sistemas Inteligentes

Projeto da Disciplina
Etapa 3

Alunos:

Bruno Henrique Da Silva Lucena

João Victor Voltarelli

Rodrigo Leonello Bellotti

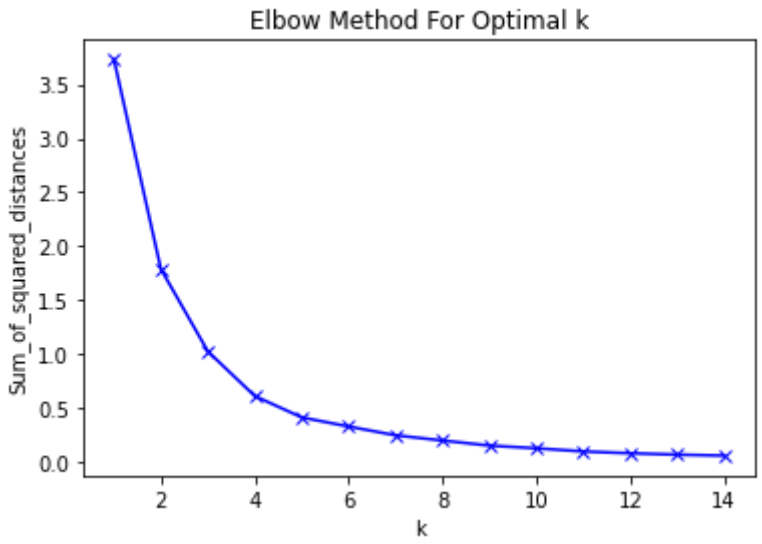
- **Ajuste dos dados**

  Com o mapa de calor e a matriz de correlação gerados através dos algoritmos em Python, analisamos quais os atributos que apresentavam uma relação significativa entre si, dessa forma, removemos também os atributos que não apresentavam relevância para o nosso objetivo. Com isso, tentamos manter na base de dados somente os atributos que representavam diretamente a economia de um estado e os atributos que representavam os números relacionados ao vírus do COVID-19.

- **Técnicas para o uso do algoritmo de clusterização k-means**

  Com base nos dados que seriam utilizados, foram aplicadas algumas técnicas para que fosse possível obter a melhor forma de usar o algoritmo k-means. Utilizamos a linguagem Python e o método de Elbow, assim como o coeficiente de Silhouette para determinarmos a quantidade de clusters que seria ideal para o nosso modelo, como mostram os resultados abaixo:

  - Método Elbow



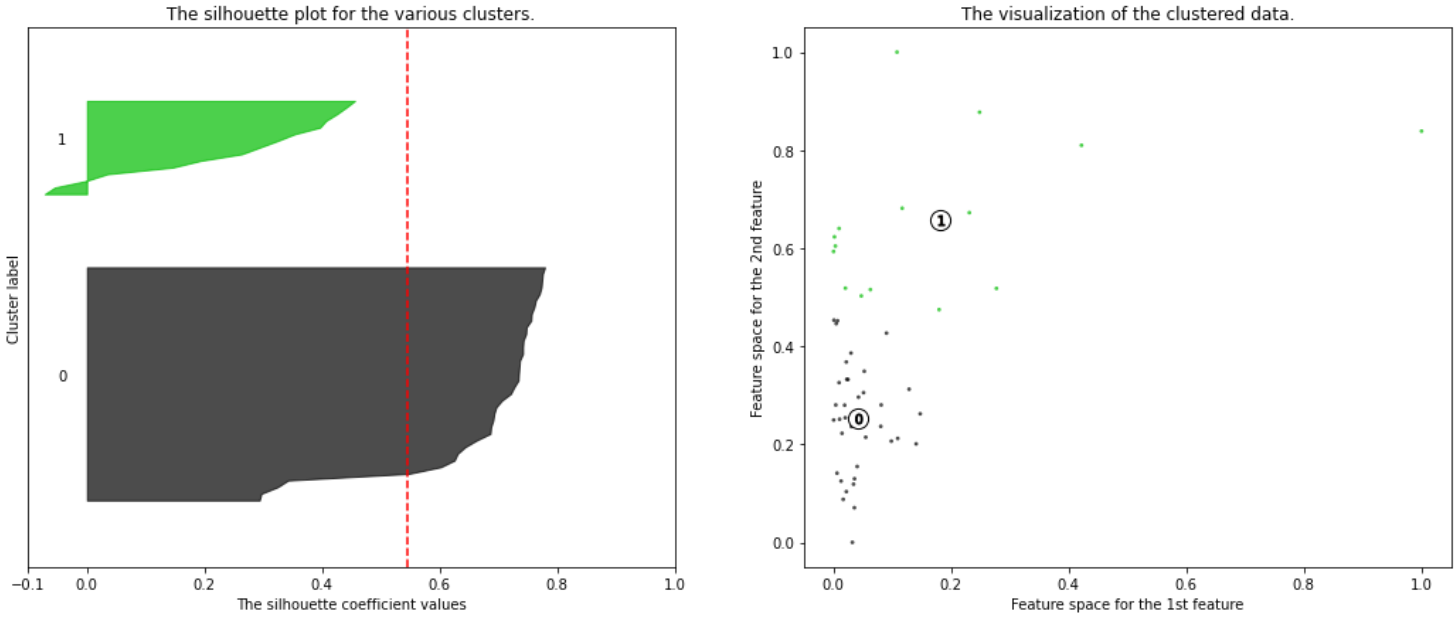
Com os resultados que obtivemos através do método Elbow é possível observar que o número ideal de clusters pode estar aproximadamente entre 2 e 3, visto que seguindo a regra do “cotovelo”, após 3 clusters, os novos clusters que são adicionados não apresentam uma diferença muito grande.

- Coeficiente de Silhouette

Para enxergarmos melhor, foi utilizado também o algoritmo do coeficiente de Silhouette, que também auxilia na escolha do número de clusters baseado no modelo desejado.

```
Para n_clusters = 2 O score_silhouette médio é : 0.544053900235069
Para n_clusters = 3 O score_silhouette médio é : 0.5352424959415405
Para n_clusters = 4 O score_silhouette médio é : 0.4894129458443845
Para n_clusters = 5 O score_silhouette médio é : 0.4603631862324087
Para n_clusters = 6 O score_silhouette médio é : 0.46285956304585585
```

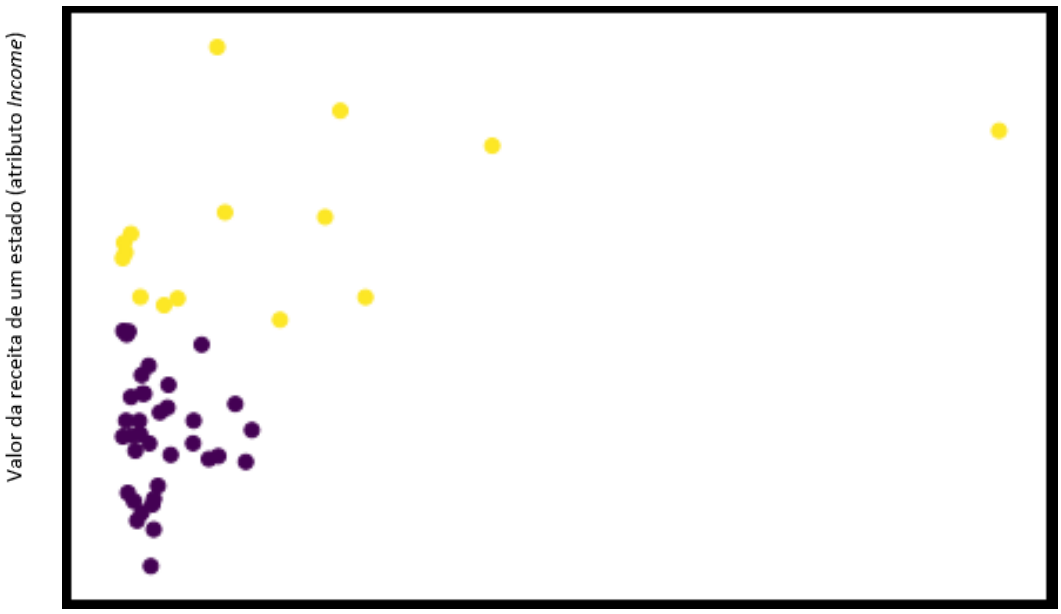
Observando a saída é possível perceber que utilizar 2 clusters parece ser o ideal, o coeficiente de Silhouette quando mais próximo de 1 indica que os pontos estão muito longes dos pontos dos outros clusters, e quando próximo de 0, indica que os pontos estão muito perto ou até interseccionando um outro cluster. A saída do algoritmo também mostra os gráficos com o score da análise e com uma visualização dos clusters.



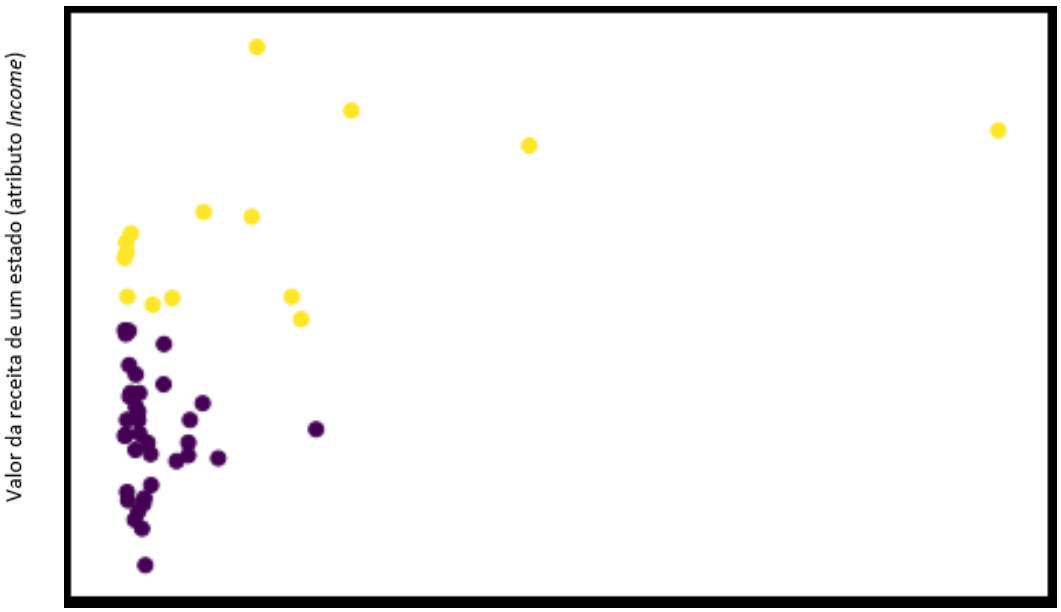
Gráficos para número de clusters igual a 2.

- **Utilização do algoritmo de clusterização**

Após os ajustes feitos nos dados do dataset, e nos parâmetros do algoritmo de clusterização, utilizando a linguagem Python executamos o algoritmo k-means para observarmos quais seriam as saídas e como poderíamos interpretar os dados.
Como foi identificado através do pré-processamento realizado anteriormente, o atributo "*Income"*, que simboliza a renda/receita de um estado é o atributo relacionado a economia que apresenta uma maior ligação com os atributos relacionados aos números do vírus do COVID-19. Partindo disso, ao executarmos o algoritmo utilizando o atributo "*Income"* no eixo Y e o atributo "Infected", que simboliza o número de infectados pelo vírus, no eixo X, obtivemos os seguintes resultados através dos clusters gerados:

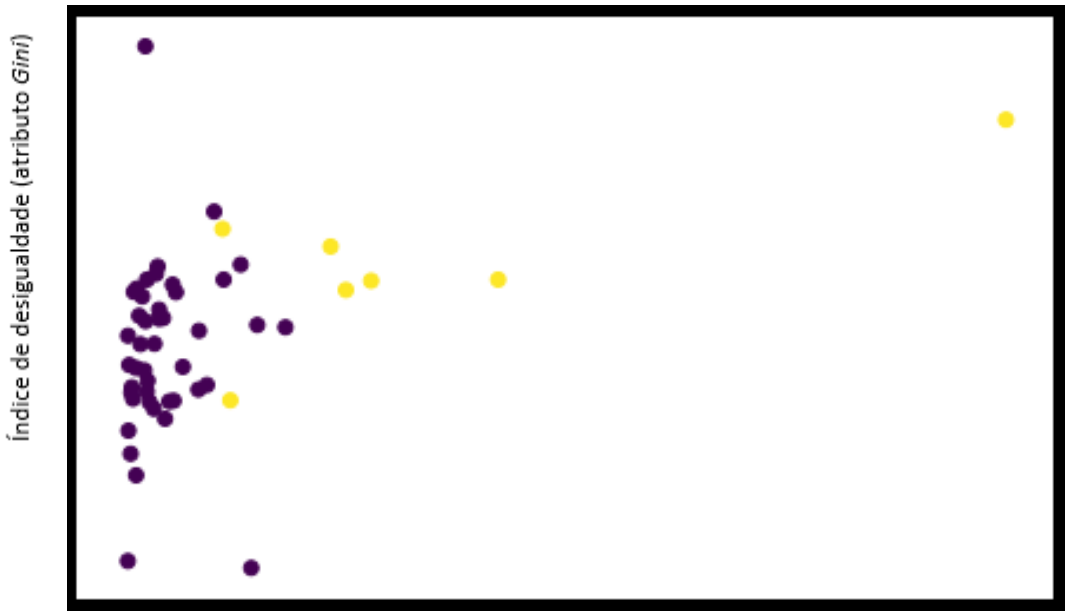


Observando os resultados do pré-processamento também foi possível identificar que o atributo "*Income"* também tem uma relação considerável com o atributo "*Deaths"*, que simboliza o número de mortes causadas pelo vírus. Essa relação com o atributo "*Deaths"* é até um pouco mais significativa do que com o atributo "*Infected"*, com isso ao gerarmos os clusters com estes atributos obtivemos a seguinte saída:

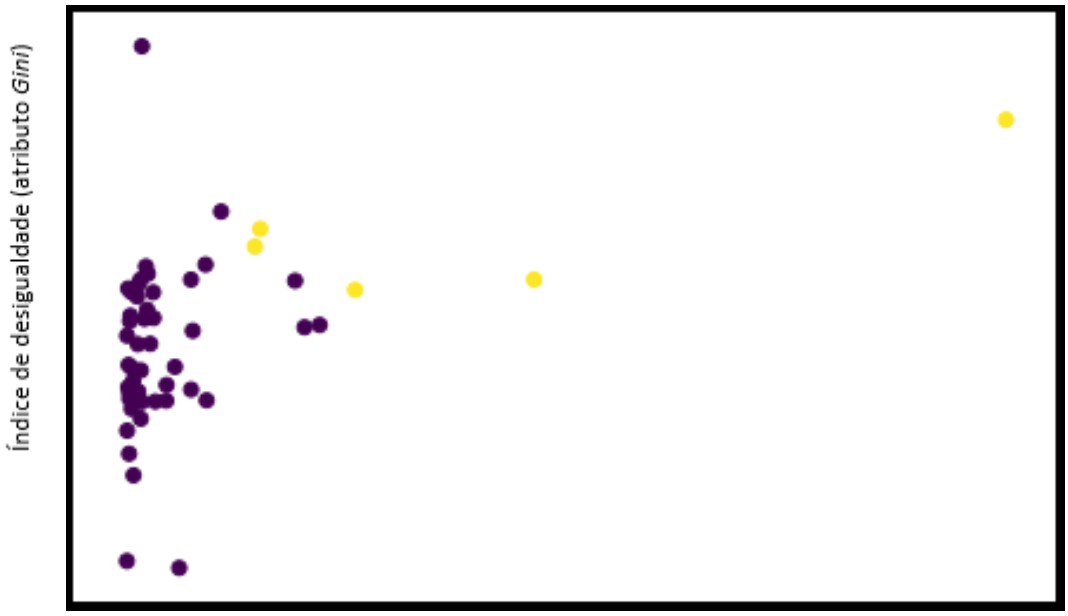


Número de mortos pelo vírus (atributo *Deaths*)

Outro atributo que está relacionado diretamente com a economia é o atributo *"Gini"*, que simboliza o coeficiente de Gini, ou seja, o índice de desigualdade de um determinado estado.
Ele também apresenta um relacionamento razoável com os atributos que representam os números referentes ao vírus. Relacionando este atributo com os atributos "*Income"* e *"Deaths"*, conseguimos observar a seguinte organização dos clusters:



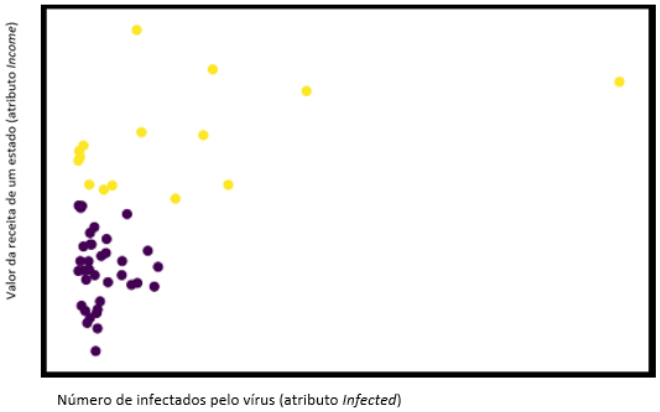
Número de infectados pelo vírus (atributo *Infected*)

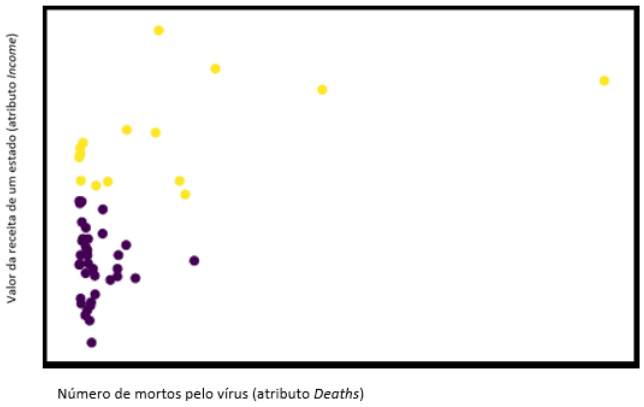


Número de mortos pelo vírus (atributo *Deaths*)

- **Conclusões finais**

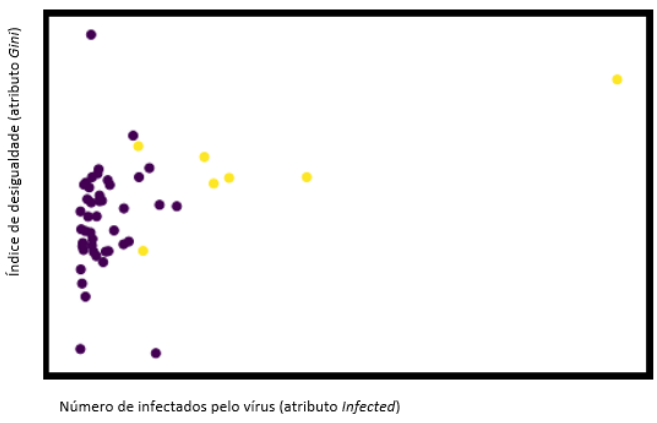
Com base nos dados que utilizamos no desenvolvimento do trabalho e com todos os processos realizados até agora, desde o pré-processamento dos dados até os ajustes dos mesmos, observando, analisando e interpretando todos os resultados que obtivemos em todas as etapas, o objetivo de identificar se a situação econômica de um estado tem influência nos números de infectados e mortos pelo vírus do COVID-19 pode ser concluído da seguinte forma: Analisando principalmente o atributo "*Income"*, que é o atributo relacionado a economia que mais se relaciona com os atributos referentes ao vírus, concluímos com base nos clusters gerados que, por mais que a economia seja um fator observável ao analisarmos os números do vírus, não é um fator predominante, ou seja, por mais que a situação econômica de um estado possa influenciar um pouco nos números de infectados e mortos, não é a economia que determina se os números relacionados ao COVID-19 vão ser maiores ou menores em alguns locais. Existem fatores que influenciam muito mais, por exemplo o número da população, densidade populacional etc.

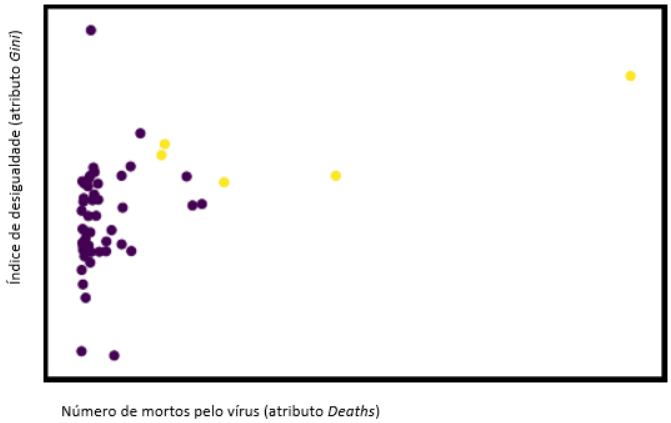




Com base nos dois principais gráficos com os clusters gerados, vemos no primeiro gráfico que alguns estados com uma renda mais alta (eixo Y) , acabam tendo um número de infectados maior (eixo X), porém não necessariamente é uma regra a ser seguida, com isso é possível analisar que não é porque um estado é mais rico que ele terá menos casos de infectados e mortos (gráfico2). É necessário também analisar outros fatores, e não somente os econômicos.

A mesma coisa acontece com os gráficos dos clusters considerando o atributo do índice de desigualdade:





Analisando os clusters percebemos que não necessariamente um estado ter menos desigualdade significa que ele terá menos infectados e mortos. Em alguns casos, estados menos desiguais acabam tendo um número de casos maior, causado pela influência de outros fatores, que não só os econômicos.

- **Resposta do objetivo inicial**

  Resumidamente, com relação a base de dados que utilizamos, os fatores econômicos apesar de serem relativamente observáveis não exercem uma influência tão alta nos números do COVID-19, nessa questão, outros fatores acabam sendo mais importantes e mais influenciáveis.

- **Referências**

https://medium.com/neuronio-br/aprendizado-n%C3%A3o-supervisionado-com-k-means-f4272dee98a0

https://minerandodados.com.br/algoritmo-k-means-python-passo-passo/

https://medium.com/pizzadedados/kmeans-e-metodo-do-cotovelo-94ded9fdf3a9

https://medium.com/@paulo_sampaio/entendendo-k-means-agrupando-dados-e-tirando-camisas-e90ae3157c17

https://www.devmedia.com.br/data-mining-na-pratica-algoritmo-k-means/4584