Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Sistemas de Apoio à Decisão

Caio Damasceno Alves - 18.2.8076
Iago Nuvem Cardoso - 18.2.8003
Thais Souto Damasceno - 18.2.8013

Professor - Helen de Cássia Sousa da Costa Lima

João Monlevade
2022

Sumário

1 Definição do Problema

O problema definido para analisar se há solução foi baseado no surto de COVID-19 no início do ano de 2021 que ocorreu em Manaus.

Em janeiro de 2021, a cidade enfrentou um caos no sistema de saúde por conta da falta de oxigênio nos hospitais. Em decorrência disso, houve um grande aumento no número de óbtitos.

Com base nisso, a proposta é analisar através do agrupamento se casos como esses poderiam ser evitados. Basicamente, analisamos 3 períodos diferentes, sendo eles o período em que houve o incidente, um mês antes (dez/2020) e um mês depois (02/2021) do incidente.

Com base nos dados de janeiro de 2021 em Manaus, é possível fazer alguns grupos de locais que estavam tendo quantidades parecidas de óbitos e entender se o problema poderia ocorrer novamente. Caso houvesse possibilidade de uma nova ocorrência parecida, pode-se entender que esse locais precisariam de maior atenção governamental e possivelmente alguma intervenção para evitar um novo incidente.

2 Ferramentas

2.0.1 Tableau:

Plataforma de análise visual utilizada na área de Business Intelligence. Com essa ferramenta, é possível capturar os dados brutos e transformá-los em análises descomplicadas, facilitando seu entendimento.

Foi utilizado para preparar toda a apresentação visual, incluindo aplicação do algoritmo e todos os tipos de filtros necessários para a análise.

2.0.2 Tableau Prep:

Plataforma que permite para estruturar, preparar e combinar os dados para análise.

Foi utilizado para realizar todo o processamento necessário na base de dados.

3 Pré Processamento

O pré processamento da base de dados foi realizada no Tableau Prep. Nele foi possível analisar toda a base e definir quais campos seriam necessários para a aplicação.

Inicialmente a base de 2020 contava com 153 colunas e a base de 2021 com 166 colunas. Após todo o processo de identificação dos dados necessários, ambas as bases reduziram para 5 colunas de dimensão e uma coluna de valor de medida gerada pelo Tableau que realiza a contagem de linhas. Com base nos dados, não foi realizada nenhuma redução, sendo assim, a análise foi feita com 2.874.607 linhas juntando as duas bases (2020 e 2021).

Antes de realizar a união de linhas das duas bases, foi necessário realizar outra limpeza na base de 2020 que contava também com alguns dados do ano de 2021 e estava gerando redundância quando a análise era realizada.

Foi realizado uma união por linhas das bases de dados. Após essa união convertemos os dados das colunas "Evolução"e "Classificação Final", ambas eram valores de medida e foram convertidas em dimensões.

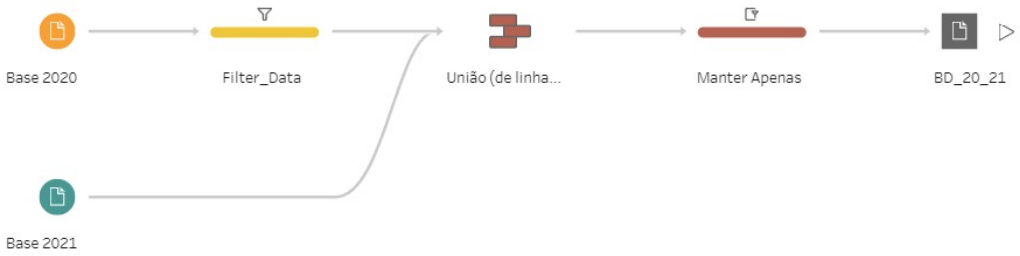

Figura 1: Processo de limpeza de dados

4 Algoritmo

O algoritmo utilizado pelo Tableau é o K-Means que é um algoritmo de clusterização. Ele é um algoritmo de aprendizado não supervisionado que avalia e clusteriza os dados de acordo com suas características.

Ele serviu para nos mostrar em diferentes clusters a distância com relação ao número de óbitos, permitindo que fossem visualizados de maneira mais clara.

Para um determinado número de clusters k, o algoritmo particiona os dados em k clusters. Cada cluster tem um centro (centroide) que é o valor médio de todos os pontos desse cluster. K-Means localiza centros por meio de um procedimento iterativo que minimiza distâncias entre pontos individuais em um cluster e o centro do cluster. (Tableau, 2023)

4.1 Descrição dos Clusters

No processo de clusterização, onde os dados foram agrupados em 5 clusters com base na variável "Contagem de Extract"e no nível de detalhe "Estados", conforme mostra a Figura 2.

O número de clusters é 5, o que significa que os dados foram divididos em 5 grupos distintos. O número de pontos é 27, o que indica o total de itens que foram utilizados na análise. A escala está normalizada, o que sugere que os valores foram ajustados para uma escala comum. A soma dos quadrados entre grupos é 1.1818, o que representa a variação entre os clusters. A soma dos quadrados dentro do grupo é 0.013002, o que indica a variação dentro de cada cluster. A soma total de quadrados é 1.1948, que representa a variação total dos dados. Os centros apresentados para cada cluster representam a média dos valores da variável "Contagem de Extract"para cada grupo. Por exemplo, o Cluster 1 possui 15 itens com uma contagem média de 176.13 para a variável "Contagem de Extract". O Cluster 4 tem apenas 1 item, com uma contagem de 5717.0 para "Contagem de Extract". O Cluster 5 também possui apenas 1 item, com uma contagem de 2076.0.

Por fim, o grupo "Não em cluster"não tem nenhum item que foi atribuído a um cluster específico. Isso pode ocorrer porque esses itens não se enquadram bem

em nenhum dos grupos definidos pela análise de clusterização ou podem ter sido excluídos da análise por algum motivo.

Entradas para clustering

Variáveis:	Contagem de Extract
Nível de detalhe:	Estados
Escala:	Normalizado

Diagnósticos de resumo

Número de clusters:	5
Número de pontos:	27
Soma dos quadrados entre grupos:	1.1818
Soma dos quadrados dentro do grupo:	0.013002
Soma total de quadrados:	1.1948

		Centros
Clusters	**Número de itens**	**Contagem de Extract**
Cluster 1	15	176.13
Cluster 2	2	2606.5
Cluster 3	8	672.62
Cluster 4	1	5717.0
Cluster 5	1	2076.0
Não em cluster	0	

Figura 2: Detalhamento por Estados

Neste processo de clusterização realizado em uma base de dados de municípios, utilizando a variável "Contagem de Extract"normalizada como medida de similaridade entre as observações. O objetivo do processo foi agrupar os municípios em clusters, ou grupos, com base em sua similaridade em relação à variável selecionada, como representa a Figura 4.

Os diagnósticos de resumo apresentam algumas informações importantes sobre o resultado da clusterização. O número de clusters obtidos foi de 5, e o número total de pontos (ou observações) considerados foi de 1287. A "Soma dos quadrados entre grupos"é uma medida de variabilidade que indica o quão diferentes são os clusters uns dos outros em relação à variável selecionada. Já a "Soma dos quadrados dentro do grupo"indica o quão similares são as observações dentro de cada cluster. A "Soma total de quadrados"é a soma dos dois valores anteriores.

Os centros apresentam as características médias de cada cluster em relação à variável selecionada. No Cluster 1, por exemplo, foram encontrados 1249 municípios com uma Contagem de Extract média de 8.7678. Já no Cluster 2, há apenas 35 municípios, mas com uma Contagem de Extract média de 158.74. O Cluster 3, 4 e 5 possuem apenas um município cada, com Contagem de Extract média de 1976.0, 1004.0 e 1542.0, respectivamente.

Por fim, é possível observar que não há nenhum município classificado como "Não em cluster", o que significa que todos os pontos foram alocados em algum dos 5 clusters obtidos.

Entradas para clustering

Variáveis:	Contagem de Extract
Nível de detalhe:	Municipios
Escala:	Normalizado

Diagnósticos de resumo

Número de clusters:	5
Número de pontos:	1287
Soma dos quadrados entre grupos:	2.0317
Soma dos quadrados dentro do grupo:	0.10545
Soma total de quadrados:	2.1371

		Centros
Clusters	**Número de itens**	**Contagem de Extract**
Cluster 1	1249	8.7678
Cluster 2	35	158.74
Cluster 3	1	1976.0
Cluster 4	1	1004.0
Cluster 5	1	1542.0
Não em cluster	0	

Figura 3: Detalhamento por Municípios

5 Resultados

Com base no problema proposto, foi possível identificar quais regiões estavam passando por um momento parecido e que existiram regiões com situações piores que não sofreram o mesmo colapso que ocorreu no Amazonas.

Porém, é fundamental entender que mesmo não entrando em colapso, as regiões precisariam de uma maior atenção e maiores suportes. Isso tudo visando diminuir impactos causados pela COVID-19.

Então, o problema proposto poderia ser analisado utilizando o algoritmo K-Means e utilizando de médias para prever novos surtos nos casos e consequentemente contornando a situação de melhores formas do que as utilizadas na época.

6 Referências

- ANASTACIO, Bruno. Medium, 2020. K-means: o que é, como funciona, aplicações e exemplo em Python. Disponível em: <https://medium.com/programadores-ajudando-programadores/k-means-o-que- %C3%A9-como-funciona-aplica %C3%A7%C3%B5es-e-exemplo-em-python-6021df6e2572>. Acesso em: 19 de março de 2023.

- LIMA, Leanderson. Amazonia Real, 2022. Covid-19: crise de oxigênio em Manaus completa um ano. Disponível em: <https://amazoniareal.com.br/um-ano-da-crise-do-oxigenio/>. Acesso em: 19 de março de 2023.

- TABLEAU. Tableau, 2023. Find Clusters in Data. Disponível em:<https://help.tableau.com/current/pro/desktop/en-us/clustering.htm: :text=Tableau %20uses%20the%20k%2Dmeans,the%20points%20in%20that%20cluster> Acesso em: 19 de março de 2023.