UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Thiago Sordi

MINERAÇÃO DE DADOS TEMPORAIS DA COVID-19 NO BRASIL: BUSCA DE SIMILARIDADES NAS CARACTERÍSTICAS DOS ESTADOS BRASILEIROS ATRAVÉS DO AGRUPAMENTO HIERÁRQUICO

Santa Maria, RS
2023

Thiago Sordi

MINERAÇÃO DE DADOS TEMPORAIS DA COVID-19 NO BRASIL:
BUSCA DE SIMILARIDADES NAS CARACTERÍSTICAS DOS ESTADOS BRASILEIROS ATRAVÉS DO AGRUPAMENTO HIERÁRQUICO

Trabalho Final de Graduação apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **em Ciência da Computação**.

Orientador: Prof. Joaquim Vinicius Carvalho Assunção

Santa Maria, RS
2023

Thiago Sordi

MINERAÇÃO DE DADOS TEMPORAIS DA COVID-19 NO BRASIL:
BUSCA DE SIMILARIDADES NAS CARACTERÍSTICAS DOS ESTADOS BRASILEIROS ATRAVÉS DO AGRUPAMENTO HIERÁRQUICO

Trabalho Final de Graduação apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **em Ciência da Computação**.

Aprovado em 13 de dezembro de 2023:

Joaquim Vinicius Carvalho Assunçao, Dr. (UFSM)
(Presidente/Orientador)

Sérgio Luis Sardi Mergen, Dr. (UFSM)

Leonardo Ramos Emmendorfer, Dr. (UFSM)

Santa Maria, RS
2023

RESUMO

MINERAÇÃO DE DADOS TEMPORAIS DA COVID-19 NO BRASIL: BUSCA DE SIMILARIDADES NAS CARACTERÍSTICAS DOS ESTADOS BRASILEIROS ATRAVÉS DO AGRUPAMENTO HIERÁRQUICO

AUTOR: Thiago Sordi
Orientador: Joaquim Vinicius Carvalho Assunção

A pandemia da COVID-19 apresentou uma série de desafios sem precedentes para a saúde pública global, e o Brasil foi um dos países severamente afetados. Diversos trabalhos foram feitos para comparar diferentes contextos temporais da COVID-19. Neste trabalho foi realizado algo similar, explorando a dinâmica temporal da COVID-19 no Brasil, buscando possíveis padrões entre estados específicos, comparando-os em determinadas características. Para isso utilizou-se um conjunto de dados que compreende notificações de casos da doença em cada estado brasileiro durante os anos de 2020 e 2021, e exploraram-se métricas de distância como Dynamic Time Warping (DTW) e distância Euclidiana para analisar séries temporais de dados. Além disso, as ferramentas PySpark e Pandas foram utilizadas para manipulação e preparação de dados, facilitando a análise subsequente. Foram empregados métodos como importância de recursos, para identificar e selecionar as características nas notificações que impactavam nas evoluções clínicas, e o Agrupamento Hierárquico para relacionar as séries temporais em suas similaridades.

Palavras-chave: COVID-19. Engenharia de dados. Data Science. Mineração de Dados. Séries temporais.

ABSTRACT

TEMPORAL DATA MINING OF COVID-19 IN BRAZIL: SEARCH FOR SIMILARITIES IN THE CHARACTERISTICS OF BRAZILIAN STATES THROUGH HIERARCHICAL CLUSTERING

AUTHOR: Thiago Sordi
ADVISOR: Joaquim Vinicius Carvalho Assunção

The COVID-19 pandemic posed a serie of unprecedented challenges to global public health, with Brazil being one of the severely affected countries. Various studies have been conducted to compare different temporal contexts of COVID-19. This work carries out a similar exploration, delving into the temporal dynamics of COVID-19 in Brazil, and seeking potential patterns among specific states by comparing them based on certain characteristics. For this purpose, a dataset containing notifications of disease cases in each Brazilian state during the years 2020 and 2021 was used, and distance metrics such as Dynamic Time Warping (DTW) and Euclidean Distance were employed to analyze time series data. Furthermore, tools like PySpark and Pandas were utilized for data manipulation and preparation, thus facilitating subsequent analysis. Methods such as feature importance were applied to identify and select characteristics in the notifications that impacted clinical evolutions, and Hierarchical Clustering was used to relate the time series based on their similarities.

Keywords: COVID-19. Data Engineer. Data Science. Data Mining. Time Series.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

DTW	Dynamic time warping
CSV	Comma-separated values
TPU	Tensor Processing Unit
GPU	Graphics Processing Unit
EUA	Estados Unidos da América

SUMÁRIO

1 INTRODUÇÃO

A pandemia da COVID-19, causada pelo novo coronavírus SARS-CoV-2, emergiu no final de 2019 em Wuhan, China, e rapidamente se tornou uma das crises de saúde pública mais desafiadoras do século XXI. Em poucos meses, o vírus se alastrou pelo mundo, afetando nações em todos os continentes e levando a Organização Mundial da Saúde a declarar a situação como uma pandemia global. Esta crise sanitária não apenas sobrecarregou sistemas de saúde ao redor do mundo, mas também teve severas implicações econômicas, sociais e políticas.

O Brasil, com sua vasta extensão territorial e diversidade populacional, enfrentou desafios únicos durante a pandemia. Em outubro de 2021, o país já havia registrado mais de 600.000 mortes devido à doença (Evandro Furoni and Giulia Alecrim and André Luiz Rosada, 2021), com picos diários alarmantes, como o registro de mais de 4.249 mortes em um único dia (Jonas Valente, 2021). Estes números, por si só, demonstram a gravidade da situação, mas também levantam questões sobre a dinâmica da doença em diferentes regiões do país, as respostas políticas e de saúde pública adotadas e os fatores que podem ter influenciado a disseminação do vírus.

Neste contexto, a análise de dados desempenha um papel crucial. A capacidade de coletar, processar e interpretar grandes volumes de dados em tempo real tornou-se uma ferramenta essencial para monitorar a evolução da pandemia, prever tendências e informar decisões políticas. A partir disso, a ideia deste trabalho surge, com o objetivo de explorar os dados da COVID-19 no Brasil, analisando e comparando características de um conjunto de dados temporal, contendo notificações de casos da doença em cada estado brasileiro.

Muitas análises parecidas já foram feitas ao longo desses 3 últimos anos, principalmente no decorrer da pandemia, no entanto, foram utilizados outros contextos (comparação entre países ou entre estados dos Estados Unidos, por exemplo). Embora este trabalho possa empregar algumas técnicas analíticas utilizadas por eles, ele se distingue por sua ênfase na mineração dos dados da dinâmica da COVID-19 especificamente no contexto brasileiro.

Desse modo, este trabalho tem como foco principal explorar diferentes métodos para comparar séries temporais, formas de agrupamento e representações visuais para esses dados, a fim de demonstrar as relações entre os casos de óbitos nos estados brasileiros selecionados. Assim, obtendo como resultado mais concreto, o desenvolvimento completo de um *pipeline*, capaz de extrair dados de páginas web, processá-los e transformá-los, preparando-os assim para a análise de agrupamento.

Para o desenvolvimento, foi escolhido o ambiente Python em nuvem do Google Colaboratory, que oferece uma integração com Google Drive, para salvar os datasets extraídos e processados. Os dados que serviram como base para esse trabalho são os datasets

(2020 e 2021) de notificações de COVID-19 disponibilizados pelo OpenDataSUS[1]. Com os dados selecionados e o ambiente preparado, foram feitas as extrações desses conjuntos de dados utilizando *web scrapping*, aplicadas transformações neles utilizando Dataframes Pyspark, e por fim aplicados algoritmos nesses dados, após serem preparados com Pandas. Para comparar e medir as séries temporais, foi aplicado o Dynamic Time Warping (DTW) e a Distância Euclidiana. Após essa medida, então foi realizado o agrupamento e análise dos óbitos nos estados brasileiros, e para isso foi utilizado o agrupamento hierárquico e dendrogramas para visualização.

Em suma, a pandemia da COVID-19 trouxe à tona a importância da ciência de dados no enfrentamento de crises de saúde pública. Este trabalho busca contribuir para este campo emergente, oferecendo uma perspectiva de cenários diante da situação no Brasil e destacando as técnicas e ferramentas que podem ser usadas em problemas semelhantes no futuro.

1.1 OBJETIVO GERAL

Este trabalho visa utilizar todas técnicas mencionadas para obter resultados que demonstrem possíveis padrões de relacionamentos entre determinados estados brasileiros em caracteristicas selecionadas (idade, sintomas, condições pré-existentes e sexo) das notificações que resultaram em óbitos, para os anos de 2020 e 2021[2].

1.2 OBJETIVOS ESPECIFICOS

- Criar um fluxo para extração (*web scrapping*) e processamento dos dados brutos;
- Minerar os dados, aplicar medidas de distância de séries temporais e algoritmo de agrupamento hierárquico em diferentes cenários de caracteristicas para os anos 2020 e 2021;
- Medir os resultados através do Coeficiente de Silhueta;
- Mostrar os resultados através de dendrogramas e analisá-los.

[1] https://opendatasus.saude.gov.br/dataset?tags=covid

[2] Ressalta-se que este trabalho não é da área da saúde, e não cabem análises mais profundas dentro desse tema.

1.3 ORGANIZAÇÃO DO TEXTO

O trabalho está estruturado da seguinte forma: no Capítulo 2 estão descritos os conceitos e as tecnologias empregadas no desenvolvimento do trabalho, assim como trabalhos relacionados com o tema; no Capítulo 3 é apresentada a forma com que o trabalho foi desenvolvido, esclarecendo todos os passos para a obtenção dos objetivos propostos; no Capítulo 4 são apresentados os experimentos e resultados, além das interpretações dos resultados, e por fim, no Capítulo 5, as conclusões do trabalho.

2 REFERENCIAL TEÓRICO

Neste capítulo, são apresentados algumas das principais tecnologias e técnicas que fundamentam o desenvolvimento deste trabalho. Além disso, são citados trabalhos e livros relevantes que abordam e fornecem explicações detalhadas sobre essas tecnologias e técnicas. São apresentadas tecnologias como Pyspark e Pandas, técnicas para medida de distância como o DTW, distância Euclidiana e correlação de Pearson, além de explicar dois tipos de agrupamentos em detalhes: agrupamento hierárquico e agrupamento particional. Por fim, também são relacionados alguns trabalhos que utilizam dessas técnicas para temas similares ao deste trabalho.

2.1 AGRUPAMENTO

O agrupamento, ou clustering, é uma técnica fundamental em análise de dados e aplicações de mineração de dados. Seu principal objetivo é agrupar um conjunto de objetos de tal forma que objetos no mesmo grupo (ou cluster) sejam mais semelhantes entre si do que com aqueles em outros grupos.

> "Dado um conjunto de pontos de dados, separe-os em um conjunto de grupos que sejam o mais similar possível"(Aggarwal, Charu C and Reddy, Chandan K, 2013).

Existem diversos métodos de agrupamento, cada um com suas particularidades e aplicações. Os principais tipos incluem:

- **Agrupamento Particional:** Conforme descrito por Aggarwal, Charu C and Reddy, Chandan K (2013), este método divide o conjunto de dados em um número pré-definido de clusters sem qualquer estrutura hierárquica, os dados somente são colocados na partição (grupo) em que se encaixam melhor. O algoritmo K-Means é um exemplo popular deste tipo de agrupamento;
- **Agrupamento Hierárquico:** Ao contrário do agrupamento particional, o agrupamento hierárquico não exige a especificação prévia do número de clusters. Aggarwal, Charu C and Reddy, Chandan K (2013) descreveram que ele cria uma árvore de clusters, onde cada nível da árvore representa uma divisão dos dados. O resultado é frequentemente visualizado como um dendrograma, que mostra a sequência e o nível da relação dos clusters;

- **Agrupamento Baseado em Densidade:** Diferentemente dos dois citados anteriormente, que são algoritmos baseados em distância, este é um algoritmo próprio, baseado na densidade. Este método agrupa os dados com base na densidade de pontos no espaço de características. O algoritmo DBSCAN é um exemplo típico deste tipo de agrupamento, (Aggarwal, Charu C and Reddy, Chandan K, 2013).

No contexto deste trabalho, buscando como os diferentes grupos se relacionam em diferentes níveis, que novos agrupamentos são feitos hierarquicamente, foi escolhido utilizar o agrupamento hierárquico.

2.1.1 K-Means

De acordo com Jain, A. K. and Murty, M. N. and Flynn, P. J. (1999), o algoritmo K-Means é um método de agrupamento particional amplamente utilizado em análise de dados, caracterizado por sua simplicidade e eficiência. Inclusive sendo utilizado em um dos trabalhos relacionados deste trabalho. Ele é empregado para a segmentação de dados em um número pré-definido de clusters, K, baseando-se em atributos numéricos dos dados.

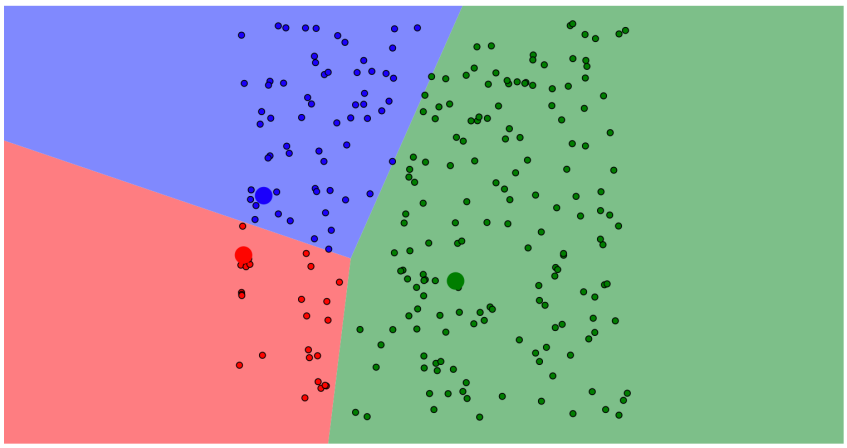
Para montar esses grupos o algoritmo começa escolhendo aleatoriamente K pontos como centros iniciais dos clusters, denominados centróides. Estes pontos podem ser escolhidos aleatoriamente dentre os dados ou por outros métodos mais sofisticados para a inicialização do K-Means.

Cada ponto no conjunto de dados é atribuído ao cluster cujo centróide é o mais próximo, demonstrado na Figura 1. A proximidade é geralmente medida usando a distância Euclidiana (subseção 2.3.2). Após a atribuição de todos os pontos a um cluster, os centróides são recalculados. Isso é feito tomando a média de todos os pontos atribuídos a cada cluster, redefinindo assim a posição do centróide. Matematicamente, o novo centróide c de um cluster é calculado com a seguinte equação 2.1, onde x é cada ponto pertecente ao cluster e $|S|$) é o número total de pontos no cluster.

$$c = \frac{1}{|S|} \sum_{x \in S} x \tag{2.1}$$

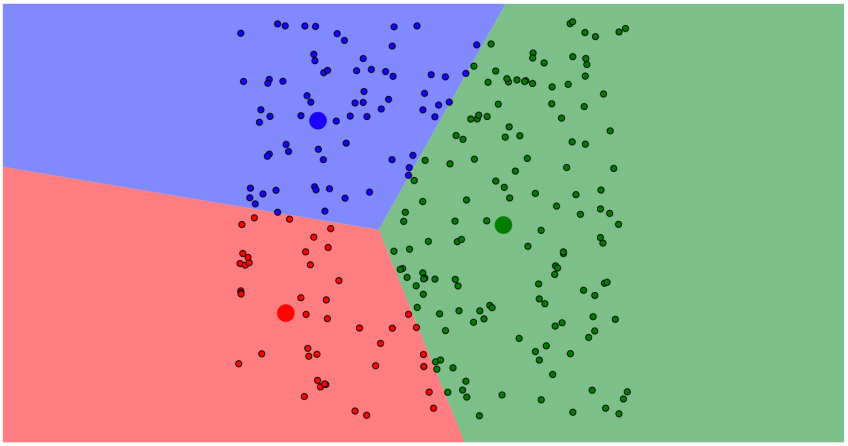
Esse processo de reatribuição dos pontos e calcular os novos centróides (demonstrado na Figura 2) é repetido até que uma condição de parada seja satisfeita. Essa condição pode ser um número fixo de iterações, uma mudança mínima nos centróides, ou uma mínima alteração na soma das distâncias quadradas dentro do cluster.

Figura 1 – Simulação da execução do K-Means - Inicialização aleatória dos centróides



Fonte: Naftali Harris.[1]

Figura 2 – Simulação da execução do K-Means - Pontos reatribuídos e cálculo dos novos centróides



Fonte: Naftali Harris.

Por isso, este algoritmo é muito sensível a escolha inicial dos centróides, e isso pode culminar em um mínimo local da função especificada, (Jain, A. K. and Murty, M. N. and Flynn, P. J., 1999). Assim, é recomendado que sejam realizadas algumas iterações para evitar uma falsa convergência.

2.1.2 Agrupamento Hierárquico

O agrupamento hierárquico é uma técnica de análise de cluster que busca construir uma hierarquia entre os clusters. Esta abordagem não exige a especificação prévia do número de clusters, diferentemente de outros tipos de agrupamento, principalmente o particional. Em vez disso, ele cria uma série de clusters que variam desde um único cluster

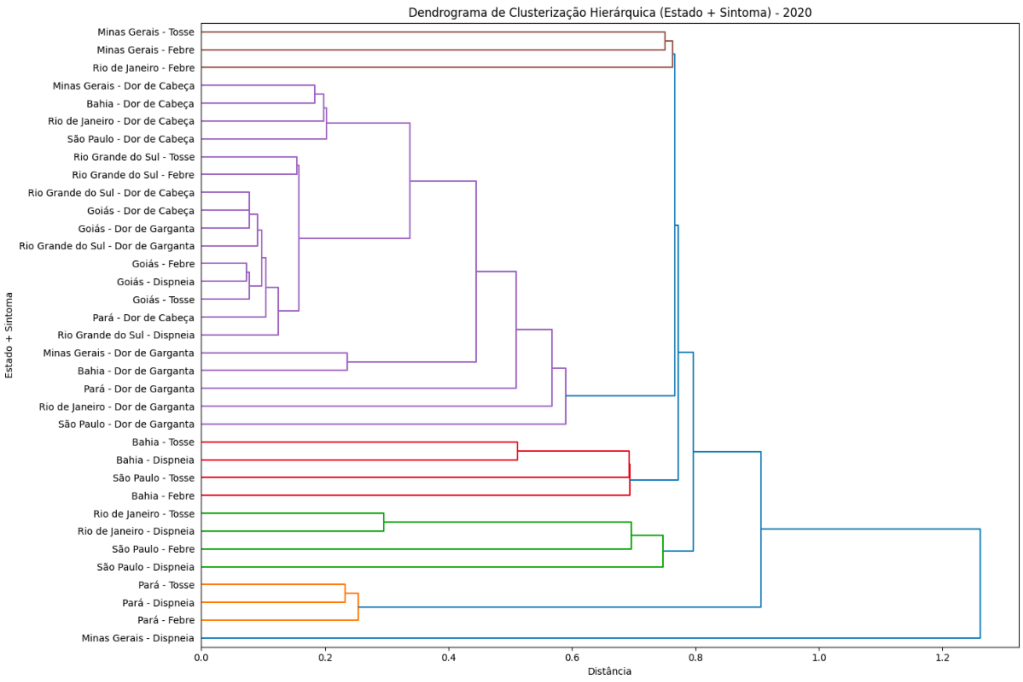
[1] Disponivel em: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

contendo todas as observações até n clusters, cada um contendo uma única observação. Segundo Johnson (1967) em seu trabalho sobre agrupamento hierárquico, esta técnica é especialmente útil quando se deseja entender estruturas de dados complexas e relações hierárquicas inerentes.

Esta técnica começa considerando cada ponto de dado como um cluster individual e, progressivamente, funde-os com base na medida de similaridade ou dissimilaridade. Essa fusão de clusters no agrupamento hierárquico segue um processo iterativo, onde, em cada etapa, os dois clusters mais próximos são combinados. Esta proximidade é determinada por funções de *linkage*, que definem como a distância entre conjuntos de clusters é calculada.

Os resultados deste tipo de agrupamento são frequentemente visualizados por meio de um dendrograma. Esta representação gráfica em forma de árvore ilustra a disposição dos clusters formados pelo agrupamento hierárquico. Através do dendrograma, é possível observar a sequência e a estrutura das fusões ou divisões de clusters. A altura na qual duas ramificações se unem no dendrograma indica a distância (ou dissimilaridade) entre esses dois clusters, conforme descrito por Murtagh e Legendre (2014) em sua revisão sobre métodos de agrupamento hierárquico. Esse dendrograma também pode ser cortado em determinada altura, este corte em um nível específico de dissimilaridade determina a formação final dos clusters. Na Figura 3 pode-se ver isso exemplificado com um dendrograma orientado horizontalmente, onde no eixo horizontal estão as distâncias, e o eixo vertical todos os grupos, e as linhas conectam sempre dois grupos em cada nível.

Figura 3 – Dendrograma exemplificando saída de um agrupamento hierárquico do trabalho.



Fonte: Próprio autor

Como mencionado, dentro do agrupamento hierárquico, a escolha da função de *linkage* é crucial, pois determina como a similaridade entre dois clusters é calculada. Algumas das funções de *linkage* mais comuns incluem (os clusters estão representados por X e Y):

- Linkage Simples (Single Linkage): Calcula a distância entre os dois clusters como a menor distância entre um ponto em um cluster e um ponto no outro cluster, conforme a equação 2.2, onde x e y são pontos pertecentes aos clusters X e Y. Este método tende a produzir clusters alongados ou em cadeia;

$$D(X,Y) = \min\{d(x,y) : x \in X, y \in Y\} \tag{2.2}$$

- Linkage Completa (Complete Linkage): Calcula a distância entre os dois clusters como a maior distância entre um ponto em um cluster e um ponto no outro cluster, conforme a equação 2.3, onde x e y são pontos pertecentes aos clusters X e Y. Este método tende a produzir clusters mais compactos e esféricos;

$$D(X,Y) = \max\{d(x,y) : x \in X, y \in Y\} \tag{2.3}$$

- Linkage Média (Average Linkage): Calcula a distância entre os dois clusters como a média das distâncias entre todos os pares de pontos, onde um ponto pertence a um

cluster e o outro ponto pertence ao outro cluster. Na seguinte equação 2.4, $|X|$ e $|Y|$ são os tamanhos dos clusters X e Y, enquanto x e y são pontos pertecentes aos clusters X e Y;

$$D(X,Y) = \frac{1}{|X| \times |Y|} \sum_{x \in X, y \in Y} d(x,y) \tag{2.4}$$

- Linkage de Ward: Minimiza o aumento total da soma do quadrado das distâncias entre os pontos de dados e a média dos clusters aos quais pertencem. Essa abordagem tende a formar clusters de tamanho mais uniforme e é particularmente útil para identificar padrões naturais nos dados. Nesta equação 2.5, $|X|$ e $|Y|$ são os tamanhos dos clusters X e Y, n é o número de dimensões e $\bar{x}_i$ $\bar{y}_i$ são as médias dos clusters em cada dimensão.

$$D(X,Y) = \sqrt{\frac{2|X||Y|}{|X| + |Y|} \sum_{i=1}^{n} (\bar{x}_i - \bar{y}_i)^2} \tag{2.5}$$

Esta abordagem de agrupamento tem sido fundamental em várias áreas da ciência de dados, desde a biologia até às ciências sociais, devido à sua capacidade de revelar estruturas complexas em conjuntos de dados, no contexto deste trabalho busca-se essas relações entre os estados ao longo do tempo, sob determinadas características.

2.1.3 Avaliando um agrupamento com Coeficiente de Silhueta

O Coeficiente de Silhueta é uma métrica valiosa na avaliação da eficácia dos clusters formados em análises de agrupamento, ou seja, conforme descrito por Rousseeuw (1987), é avaliada a proximidade de um elemento com outros em seu próprio cluster, em comparação com os elementos do cluster mais próximo. Esta métrica varia entre -1 e 1, onde valores mais próximos de 1 indicam que os pontos estão bem agrupados dentro de seus clusters e mal ajustados aos clusters vizinhos, enquanto valores próximos de -1 indicam o oposto. Um score próximo de 0 sugere sobreposição entre os clusters.

O cálculo do Coeficiente de Silhueta para cada ponto em um conjunto de dados é feito considerando duas medidas principais: a coerência dentro do cluster e a separação entre os clusters. A coerência dentro do cluster, denotada por $a(i)$, é a distância média entre o ponto i e todos os outros pontos no mesmo cluster. Por outro lado, a separação entre os clusters, representada por $b(i)$, é a distância média entre o ponto i e os pontos no cluster mais próximo.

O coeficiente para o ponto i, denotado por $s(i)$, é então calculado pela equação 2.6.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.6}$$

O valor resultante indica quão bem o ponto i está ajustado ao seu cluster em comparação com o cluster mais próximo.

Ao calcular o coeficiente médio de silhueta para todos os pontos em um conjunto de dados, obtem-se uma visão geral da qualidade do agrupamento. Coeficientes mais altos indicam uma estrutura de cluster clara e bem definida, enquanto scores mais baixos podem indicar clusters sobrepostos ou mal definidos. O uso dessa métrica é particularmente útil para validar a escolha do número de clusters em métodos como o K-Means, onde essa escolha não é óbvia. Mas também util ao agrupamento hierárquico ao determinar a melhor altura (distância) para o corte do dendrograma, e assim gerando o melhor número de clusters para a situação.

2.2 NORMALIZAÇÃO DE DADOS

A normalização de dados é um procedimento de pré-processamento aplicado nos dados de entrada, pois normalmente o intervalo de algumas variáveis estão muito diferentes, e dessa forma, evita-se que a escala numérica tenha importância, mas sim a oscilação dos valores (HEIDARI; SOBATI; MOVAHEDIRAD, 2016). Este processo ajusta os valores das variáveis numéricas no conjunto de dados para uma escala comum, sem distorcer as diferenças nos intervalos de valores ou perder informações. A normalização é importante porque algoritmos de aprendizado de máquina e mineração de dados muitas vezes têm melhor desempenho com dados de entrada que estão na mesma escala. Aqui são apresentadas as técnicas de MinMaxScaler, uma das mais comuns e também usada neste trabalho, e também a Z-Score que foi usada por outros trabalhos.

2.2.1 MinMaxScaler

O MinMaxScaler é uma técnica de normalização que transforma os recursos redimensionando cada um para um intervalo dado, geralmente entre zero e um. A implementação para o MinMaxScaler é dada pela equação 2.7 (HEIDARI; SOBATI; MOVAHEDIRAD, 2016).

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{2.7}$$

onde X é o valor original, X_{min} é o valor mínimo do recurso, e X_{max} é o valor máximo

do recurso.

Esta técnica é útil quando precisa-se de uma normalização que não distorça os dados, como em imagens onde os valores de pixel devem permanecer entre 0 e 255.

2.2.2 Z-Score

O Z-Score, ou normalização por padronização, é outra técnica de normalização de dados que envolve a re-escala dos dados para ter uma média de zero e um desvio padrão de um. A equação 2.8 serve para calcular o Z-Score de um ponto de dados (ALBERTO-OLIVARES et al., 2019).

$$Z = \frac{(X - \mu)}{\sigma} \tag{2.8}$$

onde X é o valor do ponto de dados, μ é a média dos dados, e σ é o desvio padrão dos dados.

Esta técnica é particularmente útil quando os dados têm uma distribuição normal, e você deseja comparar pontos de diferentes conjuntos de dados em uma escala comum.

2.3 MEDIDAS DE DISTÂNCIA EM SÉRIES TEMPORAIS

Séries temporais são sequências de pontos de dados, medidos tipicamente em intervalos de tempo sucessivos. Atualmente, elas são encontradas, e até mesmo utilizadas amplamentes em uma variedade de áreas, incluindo finanças, medicina, meteorologia e, no caso deste trabalho, na análise das características de uma doença ao longo do tempo. Uma das tarefas fundamentais na análise de séries temporais é determinar a similaridade ou dissimilaridade entre duas ou mais séries, pois é com isso que realizam-se previsões, detecções de anomalias ou agrupamentos.

A medida de distância entre séries temporais é uma ferramenta essencial para quantificar a similaridade entre elas. O objetivo é comparar os valores das séries em diferentes instantes e consolidar essas diferenças em uma única métrica. No entanto, devido à natureza dinâmica das séries temporais, onde os padrões podem mudar ao longo do tempo, e à possibilidade de deslocamentos temporais entre séries, uma comparação direta ponto a ponto pode ser enganosa, pois pode não capturar a verdadeira similaridade subjacente entre as séries. Portanto, para obter uma representação mais precisa da distância entre séries temporais, métodos mais sofisticados e adaptativos são frequentemente empregados.

Existem várias métricas de distância que foram propostas para séries temporais,

cada uma com suas próprias vantagens e limitações. Algumas métricas, como a distância Euclidiana, são simples e rápidas de calcular, mas podem não capturar adequadamente as nuances e padrões nas séries. Outras, como o Dynamic Time Warping (DTW), são mais flexíveis e podem alinhar melhor séries temporais que estão com comprimento diferente, mas são computacionalmente mais intensivas. Além dessas, a Correlação de Pearson, apesar de não ser diretamente usada neste trabalho, é util por sua aplicabilidade em determinar a força e direção das relações lineares entre séries temporais. Nas próximas subseções, cada uma dessas métricas será explorada em detalhes, ilustrando sua relevância e aplicabilidade em diferentes cenários de análise de séries temporais.

2.3.1 Correlação de Pearson

A Correlação de Pearson surge como uma ferramenta estatística essencial para medir a relação linear entre duas variáveis (PEARSON, 1895). Ela é amplamente aplicada para investigar o grau de associação entre conjuntos de dados em geral, sendo particularmente útil em séries temporais onde as relações entre diferentes séries ou entre diferentes pontos temporais dentro de uma série são de interesse.

A correlação de Pearson, representada pelo coeficiente r, varia de -1 a 1. Um valor de $r = 1$ indica uma correlação positiva perfeita, onde o aumento em uma variável está sempre associado ao aumento na outra. Inversamente, $r = -1$ denota uma correlação negativa perfeita, sinalizando que o aumento em uma variável sempre corresponde à diminuição na outra. Um valor de $r = 0$ sugere que não há relação linear entre as variáveis.

Esse coeficiente pode ser obitido conforme a equação 2.9. Nesta equação, X_i e Y_i representam os valores individuais nas séries temporais que estão sendo comparadas, enquanto $\bar{X}$ e $\bar{Y}$ são as médias dessas séries. O numerador da equação, $\sum(X_i - \bar{X})(Y_i - \bar{Y})$, calcula o produto da diferença entre cada valor da série e sua média, fornecendo uma medida agregada do relacionamento linear. O denominador normaliza o resultado, garantindo que o coeficiente não seja influenciado pela escala dos dados.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{2.9}$$

Na análise de séries temporais, a correlação de Pearson pode ser utilizada para detectar se existe alguma relação linear entre duas séries temporais. Por exemplo, ela pode ser aplicada para entender se as tendências em uma série temporal de preços de ações estão de alguma forma relacionadas às tendências em uma série temporal de índices econômicos. É importante, contudo, reconhecer que a correlação de Pearson só detecta relações lineares e pode não ser adequada para identificar relações não lineares complexas.

2.3.2 Distância Euclidiana

A distância Euclidiana é uma métrica fundamental em matemática e ciência da computação, originária da geometria Euclidiana. Ela representa a distância em linha reta entre dois pontos em um espaço, ela é calculada pela raiz quadrada da soma das diferenças quadradas entre as coordenadas dos pontos. A equação 2.10 demonstra isso, onde n é o número total de pontos, e p_i e q_i são pontos de um mesmo plano, ou até dois planos. Em análise de dados, a distância Euclidiana é frequentemente usada para medir a similaridade entre vetores de características, sendo uma escolha popular devido à sua interpretabilidade e eficiência computacional. Já no contexto de séries temporais, essa métrica pode ajudar a quantificar a diferença entre dois pontos de dados em momentos específicos.

$$d(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{2.10}$$

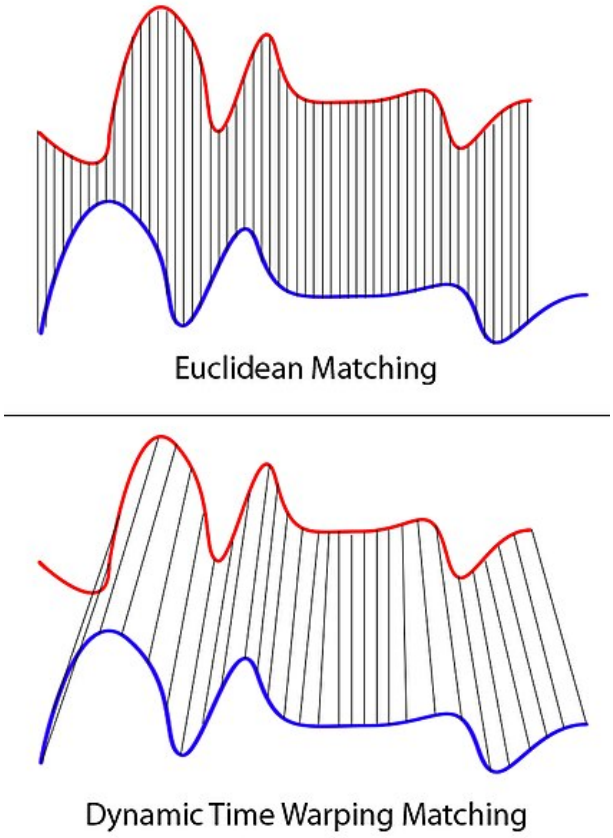
Em estudos sobre séries temporais, a distância Euclidiana tem sido amplamente utilizada como medida de comparação, como em aplicações de finanças, saúde, em geral em áreas com dados temporais. No entanto, alguns pesquisadores perceberam a importância da normalização dessa distância para melhorar os resultados de suas análises. Como é o caso de Berthold e Höppner (2016), que investigaram o agrupamento de dados de séries temporais com uma distância Euclidiana normalizada. Eles descobriram que uma distância Euclidiana quadrada normalizada por z-score (subseção 2.2.2) é, na verdade, equivalente a uma distância baseada no coeficiente de Correlação de Pearson (subseção 2.3.1). Esta descoberta tem implicações significativas, pois permite que métodos baseados em distância Euclidiana utilizem o coeficiente de Correlação de Pearson simplesmente através de uma normalização adequada dos dados de entrada. Além disso, o algoritmo K-Means, frequentemente utilizado para agrupamento, precisa ser modificado para manter a interpretação como Correlação de Pearson estritamente válida. Experimentos realizados pelos autores mostraram que, em muitos casos, o algoritmo K-Means padrão e a versão baseada na Correlação de Pearson produzem resultados semelhantes.

Assim, ao utilizar a distância Euclidiana para comparar séries temporais, é essencial considerar as características específicas das séries e, se necessário, aplicar técnicas de normalização ou outras métricas complementares para obter resultados mais precisos e significativos. No entanto, diferentemente da técnica a seguir, a distância Euclidiana tem uma limitação. Ela pode não capturar adequadamente as similaridades dinâmicas entre séries temporais que podem estar deslocadas no tempo, ou seja, séries temporais com comprimentos diferentes.

2.3.3 Dynamic Time Warping (DTW)

O Dynamic Time Warping (DTW) é uma técnica que se destaca quando tem-se que comparar séries temporais que não estão alinhadas (ou com comprimentos diferentes). De acordo com Sakoe e Chiba, o DTW foi inicialmente introduzido no contexto do reconhecimento automático de fala, onde a técnica foi empregada para alinhar e comparar diferentes padrões de fala, levando em consideração as variações na velocidade da fala entre os falantes(SAKOE; CHIBA, 1978). Pode-se ver na Figura 4 como essas duas técnicas se comportam nas mesmas duas séries temporais (desalinhadas).

Figura 4 – Diferença entre Distância Euclidiana e Dynamic Time Warping (DTW).

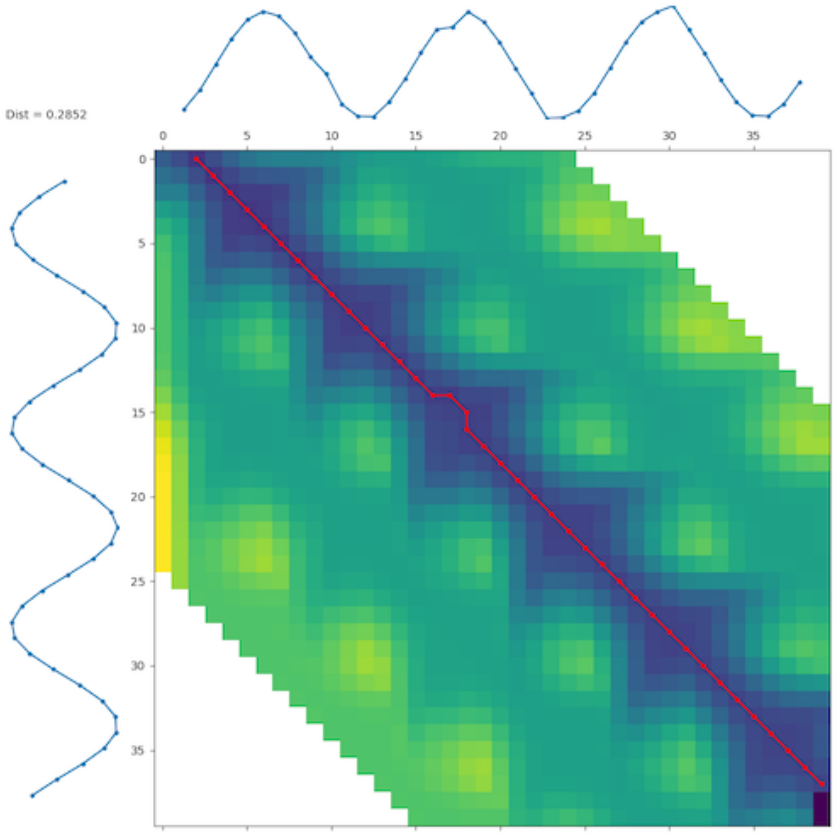


Fonte: (XANTACROSS, 2011).

O DTW trabalha construindo uma matriz que representa todos os possíveis pontos de alinhamento entre as duas séries temporais. Cada elemento da matriz calcula a distância entre um ponto em uma série e um ponto na outra. O objetivo é encontrar o caminho através desta matriz que minimize a distância total, permitindo que as séries sejam “esticadas” ou “comprimidas” para alcançar o melhor alinhamento.

Este processo pode ser visualizado através de um ’grid’ ou matriz, onde o caminho ótimo escolhido pelo DTW é destacado, mostrando como os pontos de uma série são mapeados aos pontos da outra. A Figura 5 serve para ilustrar este conceito, com uma representação gráfica do caminho de alinhamento do DTW sobre a matriz de distâncias.

Figura 5 – Grid gerado pela biblioteca dtaidistance.



Fonte: Documentação dtaidistance.[2]

A essência do DTW reside em sua capacidade de estabelecer um alinhamento ótimo entre duas sequências temporais, permitindo que as sequências sejam “esticadas” ou “comprimidas” para alcançar a melhor correspondência possível. Essa flexibilidade inerente do DTW o torna particularmente útil em aplicações que envolvem séries temporais, onde as variações temporais são comuns e muitas vezes inevitáveis(MüLLER, 2007).

Além do reconhecimento de fala, o DTW tem encontrado aplicações em uma variedade de campos, incluindo a análise de séries temporais financeiras, reconhecimento de gestos e bioinformática. A capacidade do DTW de lidar com deformações temporais e diferentes velocidades torna-o uma ferramenta valiosa para a análise de séries temporais em muitos tipos de aplicações (KEOGH; RATANAMAHATANA, 2005).

Justamente por essa flexibilidade temporal e seu uso em variados contextos, o DTW acaba sendo uma medida interessante de se utilizar ao estudar o comportamento de uma doença ao longo de tempo, ou no caso deste trabalho, para análise do comportamento da COVID-19 ao longo do tempo, comparando justamente as séries temporais entre os estados. Dessa maneira, o DTW pode oferecer insights valiosos ao identificar similaridades temporais entre diferentes regiões ou períodos.

[2]Disponível em: https://dtaidistance.readthedocs.io/en/latest/usage/dtw.html

2.4 FEATURE IMPORTANCE

A *feature importance* (ou importância de características) é empregada para identificar e classificar a relevância relativa de diferentes variáveis (recursos) em um conjunto de dados, em relação ao seu impacto sobre o resultado de um modelo preditivo (SAARELA; JAUHIAINEN, 2021). Esta técnica é uma etapa de pré-processamento opcional nas áreas de análise de dados e aprendizado de máquina.

O conceito de *feature importance* baseia-se na ideia de que nem todas as variáveis em um conjunto de dados contribuem igualmente para a precisão ou eficácia de um modelo. Algumas variáveis podem ter uma influência significativa na previsão ou classificação realizada pelo modelo, enquanto outras podem ter pouca ou nenhuma influência. Entender quais recursos são mais importantes pode ajudar a simplificar modelos, tornando-os mais eficientes e fáceis de interpretar, além de destacar áreas que podem necessitar de mais dados ou análise.

Para avaliar a *feature importance*, são utilizadas técnicas que medem o impacto de cada variável na performance do modelo. Estas técnicas podem variar desde abordagens simples, como a análise de coeficientes em modelos estatísticos, até métodos mais complexos que envolvem a manipulação e reavaliação do modelo com diferentes combinações de recursos. Por exemplo, uma abordagem comum é alterar ou remover um recurso do conjunto de dados e observar como essa alteração afeta a performance do modelo.

2.5 FERRAMENTAS

Nessa seção são apresentadas algumas ferramentas utilizadas para trabalhar com os dados, todas muito conhecidas na área da ciência de dados.

2.5.1 Notebooks Python

Como uma evolução dos ambientes de programação tradicionais, os *notebooks* representam uma ferramenta interativa e versátil para cientistas de dados, pesquisadores e programadores. Esses *notebooks* oferecem um ambiente de computação interativo, onde é possível escrever e executar blocos de código de forma modular (chamados de células), inserir anotações explicativas (texto rico), visualizar dados e compartilhar resultados de forma integrada e intuitiva. Essa abordagem facilita a experimentação, a análise exploratória de dados e a prototipagem rápida, tornando-os uma ferramenta ideal para análise de dados e aprendizado de máquina.

Existem diversos tipos de ambientes de *notebook*, e suportam várias linguagens de

programação. O mais conhecido é o Jupyter Notebook, que se destaca pela sua adaptabilidade e interatividade, sendo amplamente utilizado em ciência de dados, educação, análise de dados e muitas outras áreas. Além do Jupyter, o Google Colab é outra plataforma de destaque, oferecendo um ambiente de notebook baseado em nuvem que é totalmente integrado com o Google Drive e outros serviços Google. O Google Colab é particularmente apreciado por sua facilidade de acesso e pela capacidade de executar códigos Python em GPUs (Graphics Processing Unit) e TPUs (Tensor Processing Unit), sem a necessidade de uma configuração complexa.

2.5.1.1 Google Colaboratory

O Google Colab, ou Colaboratory, é uma plataforma de pesquisa desenvolvida pelo Google que oferece um ambiente de desenvolvimento semelhante ao Jupyter Notebook. Ele foi projetado para ajudar pesquisadores e cientistas de dados a desenvolver e executar códigos Python em um ambiente de nuvem, sem a necessidade de qualquer configuração adicional ou instalação de software.

Uma das principais vantagens do Google Colab é a capacidade de executar códigos em GPUs e TPUs fornecidas pelo Google, o que pode acelerar significativamente operações computacionais intensivas, como treinamento de modelos de aprendizado de máquina. Porém essas opções mais robustas somente estão disponíveis na versão paga (Pro).

Além disso, o Google Colab tem a integração com o Google Drive. Os notebooks do Colab podem ser salvos diretamente no Google Drive, compartilhados como qualquer outro arquivo do Drive e até mesmo incorporados a sites ou blogs. Além disso, o Colab suporta várias bibliotecas e *frameworks* populares de ciência de dados e aprendizado de máquina, tornando-o uma ferramenta versátil para uma variedade de aplicações.

2.5.2 PySpark

PySpark é a interface do Python para o Apache Spark, uma poderosa plataforma de análise de dados. O Apache Spark é uma ferramenta de processamento de dados em larga escala que oferece paralelismo para processamento de dados distribuídos. De acordo com o artigo “Apache Spark: A Unified Engine for Big Data Processing”, o Spark é projetado para ser altamente acessível, oferecendo APIs simples para Python, Java, Scala e SQL, além de bibliotecas ricas para aprendizado de máquina, gráficos, *streaming* e muito mais (ZAHARIA et al., 2016). O PySpark combina a capacidade de computação distribuída do Spark com a flexibilidade e capacidade de integração do Python, tornando-o

uma ferramenta valiosa para cientistas de dados, ainda mais se utilizado em ambientes como um *notebook* python (subseção 2.5.1).

2.5.3 Pandas

O Pandas é uma biblioteca Python amplamente utilizada para análise de dados também. McKinney (2010) em seu artigo “Data Structures for Statistical Computing in Python” descreve que o Pandas oferece estruturas de dados e funções robustas necessárias para operar com dados estruturados ou tabulares. A biblioteca é especialmente adequada para dados tabulares com colunas de tipos heterogêneos. Uma das principais características do Pandas é sua capacidade de ler uma variedade de formatos de arquivo e transformá-los em Dataframes, uma estrutura bidimensional semelhante a uma planilha.

O Pandas tem um fluxo de uso muito mais simples, desde a instalação até o uso. Porém se torna pouco eficiente quando tem-se conjuntos de dados muito grandes, e algumas vezes até distribuídos (que não foi o caso deste trabalho). A combinação de PySpark e Pandas permite uma abordagem híbrida, onde os dados podem ser pré-processados e filtrados em grande escala com PySpark e, em seguida, analisados em detalhe com Pandas.

2.6 TRABALHOS RELACIONADOS

A pandemia da COVID-19 gerou uma necessidade urgente de análise e interpretação de dados para entender a propagação e impacto do vírus em diferentes regiões ao longo do tempo. Muitos desses trabalhos foram feitos em tempo real (durante a pandemia) e para terem efeitos imediatos (ou ações imediatas) a partir de suas análises. Desde então, diversos estudos têm se dedicado a analisar a dinâmica da doença. Seja temporalmente, seja espacialmente, o melhor cenário é considerar ambos aspectos (temporal e espacial), como é visto nos seguintes trabalhos.

No estudo “Estimation of COVID-19 dynamics in the different states of the United States using Time-Series Clustering”, os autores Rojas-Valenzuela et al. (2021) propuseram uma metodologia paramétrica para analisar conjuntamente pessoas infectadas e falecidas. Esta abordagem utilizou a técnica DTW como métrica, permitindo comparar as séries temporais de diferentes comprimentos, o que é particularmente relevante para estudar os EUA (Estados Unidos da América), onde o vírus não se espalhou simultaneamente em todos os estados. Após determinar a similaridade entre as séries temporais dos estados dos EUA, um agrupamento hierárquico foi criado. Com essa metodologia, nove clusters diferentes foram identificados, mostrando comportamentos distintos nas zonas leste e oeste

dos EUA. Para prever a evolução da COVID-19 nos estados, foram utilizados os modelos Logístico, Gompertz e SIR.

Para construir o agrupamento hierárquico, os autores calcularam a distância entre todos os estados dos EUA usando a distância DTW, selecionaram os dois estados mais semelhantes e os agruparam, atualizando a matriz de distância e repetindo o processo até que todos os estados pertencessem a um grupo.

Em conclusão, o estudo destacou a utilidade do agrupamento através do clustering para a análise de séries temporais, sendo um processo não supervisionado que busca encontrar similaridades comportamentais entre as diferentes séries temporais analisadas. A métrica paramétrica proposta, baseada na distância DTW, mostrou-se robusta para diferentes comprimentos de sequências de dados, considerando o início diferente da pandemia nos diversos estados dos EUA.

Outro artigo, intitulado “Comparing the dynamics of COVID-19 infection and mortality in the United States, India, and Brazil”, os pesquisadores James, Menzies e Bondell (2022) se aprofundaram na análise da propagação e impacto da COVID-19 nesses três países, que foram fortemente afetados pela pandemia. O foco foi analisar a similaridade estrutural e as anomalias nas trajetórias de casos e mortes como séries temporais multivariadas. Para isso, foram empregados cinco métodos de otimização distintos, cada um utilizando dados dos estados de cada país, para estimar um deslocamento apropriado entre as séries temporais de casos e mortes para os EUA, Índia e Brasil como um todo.

Os autores introduziram novas abordagens para quantificar cuidadosamente a extensão da heterogeneidade em uma série temporal multivariada. Essas abordagens foram projetadas para lidar bem com a existência de elementos discrepantes, proporcionando uma análise mais precisa e confiável. Além disso, o estudo apresentou uma abordagem para estudar várias outras séries temporais onde se espera um desfasamento, como é o caso da progressão de casos para mortes na COVID-19.

Os resultados do estudo revelaram insights significativos sobre a dinâmica da COVID-19 nos três países. Foi observado que, quando os países se tornam sobrecarregados com casos de COVID-19, o comprimento da progressão de caso para morte diminui. Isso pode ser atribuído a sistemas hospitalares sobrecarregados, tratamento médico subótimo, acesso limitado a recursos médicos, como ventiladores, e um aumento nos casos não detectados. Em suma, o estudo ofereceu importantes contribuições metodológicas e descobertas não triviais sobre a heterogeneidade entre os estados em cada federação, especialmente em relação às taxas de mortalidade.

Essa análise da evolução da pandemia da COVID-19, em diferentes regiões e países, é crucial para entender as variações na disseminação do vírus e na eficácia das medidas de controle. Nisso, um estudo (CASSãO et al., 2022) semelhante aos citados acima (inclusive utilizando o (ROJAS-VALENZUELA et al., 2021) como trabalho relacionado também), buscou extrair informações e descobrir padrões a partir de um vasto conjunto de

dados sobre a disseminação da Covid-19 nos estados brasileiros. Utilizando uma técnica de agrupamento de séries temporais baseada em uma variação do K-Means e o DTW como métrica de similaridade, o estudo identificou três padrões distintos de resposta à pandemia entre os estados brasileiros.

Os dados analisados foram obtidos de um repositório que compila indicadores públicos de saúde sobre a epidemia, provenientes de fontes oficiais, como o Ministério da Saúde do Brasil. Através das análises, foi possível observar variações na resposta à pandemia, desde estados com maior controle até aqueles com menor controle da situação. No entanto, um aumento significativo no número de mortes foi observado em todos os clusters nos meses mais próximos à escrita do artigo, ressaltando a complexidade e os desafios contínuos enfrentados pelo Brasil na gestão da crise.

Esses estudos, tanto os internacionais quanto o focado no Brasil, destacam a importância de técnicas de análise avançada para entender a evolução da pandemia. A identificação de padrões e tendências podem fornecer *insights* valiosos.

Neste trabalho seguiremos algumas abordagens semelhantes ao que foi feito nesses três estudos, como o uso do DTW e a ideia de agrupamento dessas séries temporais, porém utilizando o agrupamento hierárquico e uma fonte de dados mais microscópica, ou seja, contendo mais detalhes das notificações em cada local. O trabalho também utiliza a Distância Euclidiana para medir a distância entre séries temporais também. Essas diferenças e similaridades entre os trabalhos podem ser observadas no Quadro 1.

Trabalho	**Dataset**	**Medida distância temporal**	**Agrupamento**
Estimation of COVID-19 dynamics in the different states of the United States using Time-Series Clustering	Óbitos diários nos estados dos EUA	DTW	Agrupamento hierárquico
Comparing the dynamics of COVID-19 infection and mortality in the United States, India, and Brazil	Óbitos diários nos estados dos países EUA, India e Brasil	Cinco metódos explicados no artigo: *Affinity matrices*, *Probability density function*, *Wasserstein distance*, *Energy distance* e *Normalised inner product*	Agrupamento hierárquico
Unsupervised analysis of COVID-19 pandemic evolution in brazilian states	Óbitos diários nos estados do Brasil	DTW	K-Means
Presente trabalho	Notificações diárias detalhadas de pacientes nos estados brasileiros	DTW e Distância Euclidiana	Agrupamento hierárquico

Quadro 1 – Comparação entre os trabalhos relacionados e o presente trabalho.

3 METODOLOGIA

Neste capítulo são abordadas e descritas as diferentes etapas pelas quais este trabalho passou. Desde a sua concepção, onde foram buscadas fontes de dados até o uso desses dados efetivamente no agrupamento hierárquico.

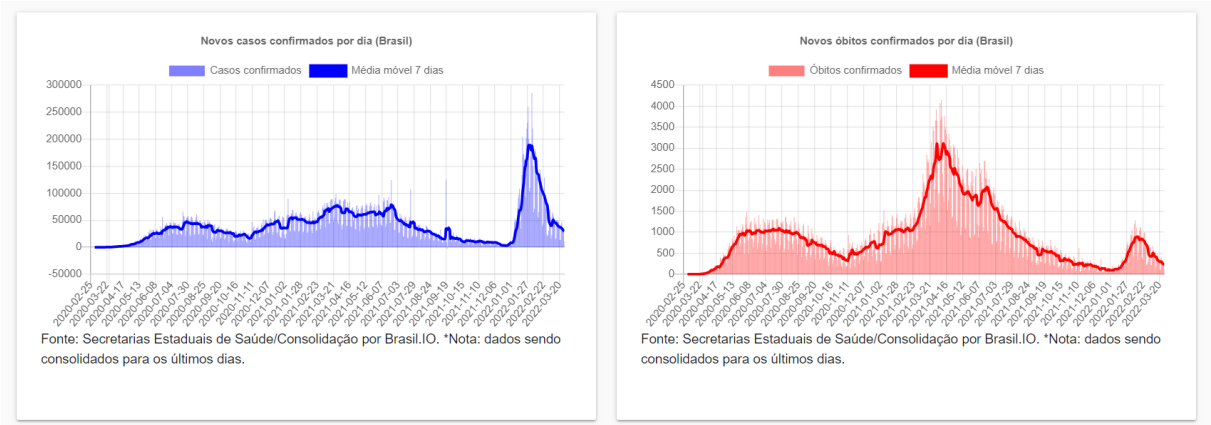
3.1 OBTENÇÃO DE DADOS

3.1.1 Fonte de dados

A fase inicial deste trabalho envolveu uma busca para identificar conjuntos de dados relevantes sobre a COVID-19 no Brasil. Era essencial selecionar dados que proporcionassem não apenas uma perspectiva temporal, mas também a informação espacial, ou pelo menos, uma caracteristica que distinguisse entre os estados brasileiros. Durante esta busca, uma variedade de conjuntos de dados foi examinada, dentre as fontes analisadas, destacam-se o conjunto de dados desenvolvido por Wesley Cota[1], o portal Brasil.io[2], e os registros fornecidos pelas secretarias estaduais de saúde. Notavelmente, os dois primeiros recursos disponibilizam interfaces gráficas pelo navegador, e assim, demonstrar alguns *insights* interessantes, como na Figura 6 onde tem-se dois gráficos de novos casos confirmados por dia no Brasil e também novos óbitos confirmados por dia no Brasil. Além de ter gráficos que relacionam caracteristicas, como a distribuição espacial de óbitos, gráficos de vacinação que podem ser comparados com de óbitos ou casos, entre outros.

[1] Disponivel em: https://github.com/wcota/covid19br
[2] Disponivel em: https://brasil.io/dataset/covid19/files/

Figura 6 – Captura de tela na página do Brasil.io sobre COVID-19.



Fonte: Brasil.io

A avaliação da adequação dos *datasets* selecionados foi um componente crucial deste trabalho, para se ter certeza de que serviriam eles precisariam ser submetidos a um processo de verificação abrangente. Este processo incluía utilizar esses dados em um *pipeline* analítico ainda em desenvolvimento, bem como a obtenção de resultados promissores em análises preliminares. Neste contexto, a análise de trabalhos relacionados desempenhou um papel fundamental, especialmente na avaliação dos formatos de dados empregados em estudos semelhantes. Estas leituras levaram à descoberta de que o *dataset* desenvolvido por Wesley Cota, já aplicado em pesquisas comparativas anteriores, possuía características macroscópicas, ou seja, era uma abordagem de série temporal e distribuição espacial dos dados, focando principalmente na contagem de casos e óbitos da COVID-19 em geral.

Por outro lado, os *datasets* providenciados por cada secretaria estadual de saúde apresentavam informações mais detalhadas, abrangendo notificações individuais dos casos, com dados anônimos de pacientes, sintomas relatados e a evolução clínica de cada caso. Entretanto, uma dificuldade emergiu devido à variação na forma como esses dados eram disponibilizados pelas diferentes secretarias, além de não possuirem um formato padronizado. Essa diversidade demandava uma etapa adicional de padronização no processo de transformação e limpeza dos conjuntos de dados, para garantir a consistência e a precisão na análise subsequente.

Em vista dessas considerações, optou-se pela utilização dos conjuntos de dados disponibilizados pelo OpenDataSUS. Este portal é uma fonte abrangente de informações de saúde no Brasil, cobrindo uma variedade de tópicos. No que diz respeito à COVID-19, o OpenDataSUS disponibiliza dados detalhados que são comparáveis aos fornecidos pelas secretarias estaduais de saúde, com a vantagem adicional de estarem organizados por estado e padronizados. Contudo, é importante mencionar uma ressalva destacada pelo próprio portal: a existência de uma potencial discrepância entre os dados disponíveis

em sua plataforma e os registros efetivos das secretarias de saúde de cada estado. Tal divergência pode ser atribuída ao processo de integração dos dados, que ainda está em andamento e enfrenta desafios decorrentes dos diferentes sistemas de registro adotados pelos estados.

Apesar das eventuais ausências nos conjuntos de dados do OpenDataSUS, a relevância deste estudo não reside na precisão absoluta das contagens de casos e óbitos, mas sim na mineração das relações entre diversas características e como estas se manifestam nos diferentes estados brasileiros ao longo do tempo. Exemplos pertinentes dessa abordagem incluem a análise da distribuição de óbitos por faixa etária, sintomas, sexo, condições, entre outras.

3.1.2 Extração dos dados

Após a definição do conjunto de dados a ser utilizado, o próximo passo envolveu o desenvolvimento de uma metodologia programática para a obtenção desses dados via internet, um processo tecnicamente conhecido como *web scraping*. A necessidade dessa etapa decorre do fato de que o OpenDataSUS fornece links individuais para cada estado brasileiro, e dentro de cada estado, existem múltiplos links que correspondem a diferentes lotes de dados. Além disso, a constante atualização desses dados pelo OpenDataSUS, devido à inclusão de novas informações e correção de dados existentes, implica que as extrações realizadas em momentos distintos podem apresentar variações no volume de dados. Por exemplo, a extração inicial não possuía a mesma quantidade de dados que a mais recente[3].

A fim de otimizar o processo de aquisição de dados, evitando a necessidade de um trabalho manual exaustivo para coletar, concatenar (nos casos dos lotes de dados de cada estado) e armazenar as informações, a plataforma Google Colab foi adotada nesta etapa. A escolha desta ferramenta deveu-se à sua capacidade de oferecer um aumento significativo de memória RAM e expansão do espaço de armazenamento disponível em nuvem. Tais recursos revelaram-se essenciais para a realização eficiente das tarefas de manipulação de dados em grande escala.

O Google Colab, análogo ao Jupyter Notebook, é estruturado em células (subseção 2.5.1.1). Neste projeto, cada célula foi designada para a tarefa específica de extrair dados de um ano determinado, especificamente 2020 e 2021. O processo de *web scraping* desenvolvido, envolveu acessar elementos e links nas páginas do OpenDataSUS, permitindo o download dos lotes de dados de cada estado. Posteriormente, esses dados foram concatenados por estado.

Após a conclusão do processo de extração e concatenação, todos os arquivos CSV

[3]Realizada em 26/08/2023

(Comma-Separated Value) resultantes foram armazenados diretamente no Google Drive. Esta estratégia foi adotada para garantir a integração e acessibilidade eficiente dos dados nos notebooks subsequentes, dedicados às etapas de transformação e análise. Cabe destacar que, em média, cada arquivo CSV, correspondente a um estado específico, continha aproximadamente 1 GB de dados. Esta abordagem facilitou significativamente a manipulação dos dados, otimizando o fluxo de trabalho e garantindo a consistência das informações ao longo do projeto.

3.1.3 Critério para seleção do período

A escolha dos anos 2020 e 2021 como foco deste estudo foi impulsionada pelos significativos desafios enfrentados pelo Brasil durante o período pandêmico. O ano de 2020 marcou o surgimento e a rápida propagação da COVID-19 no território nacional. Uma conjuntura particularmente crítica ocorreu em outubro de 2021, quando o país ultrapassou o número de 600 mil mortes devido à doença (Evandro Furoni and Giulia Alecrim and André Luiz Rosada, 2021). Além disso, o país vivenciou momentos de extrema gravidade, como em abril de 2021, quando foi registrado o alarmante número de mais de 4.249 mortes em um único dia(Jonas Valente, 2021).

Enquanto o ano de 2021 não apenas representou um período de adversidades, mas também marcou um momento crucial na luta contra a pandemia: o início da campanha de vacinação contra a COVID-19 no Brasil. Em janeiro deste ano, o país deu seus primeiros passos na imunização de sua população. Contudo, apesar do avanço representado pela vacinação, o processo enfrentou desafios substanciais, principalmente devido à limitada disponibilidade de doses e às complexidades associadas à logística de distribuição (Heloisa Cristaldo and Marcelo Brandão, 2021).

3.2 TRANSFORMAÇÃO DOS DADOS

A necessidade de processar os dados coletados surgiu mesmo com a organização prévia por estados e a padronização dos formatos. Desafios como a presença de registros com valores nulos, a existência de colunas contendo informações irrelevantes ou formatadas de maneira complexa para análise, exigiram atenção adicional. Para atender a essas demandas, foi desenvolvido um novo arquivo na plataforma Google Colab, dedicado especificamente ao processamento e refinamento dos dados.

3.2.1 Usando Pyspark

Inicialmente, houve uma tentativa de empregar Dataframes Pandas para a leitura e armazenamento dos *datasets* diretamente na memória. Contudo, tornou-se evidente a necessidade de recorrer à versão Pro do Google Colab, a fim de acessar uma maior capacidade de memória RAM, devido ao considerável tamanho dos *datasets*. Apesar dessa adaptação, os tempos de leitura e processamento permaneceram elevados, indicando que a simples expansão da memória RAM não era suficiente. Era essencial otimizar não apenas a capacidade de armazenamento, mas também a eficiência na leitura, escrita e transformação dos Dataframes. Por conseguinte, optou-se pelo uso do PySpark, uma ferramenta notavelmente adequada para o manuseio de grandes volumes de dados. Uma das principais vantagens do PySpark reside na sua capacidade de paralelizar o processamento de dados, uma funcionalidade indispensável para este projeto.

É importante destacar que o ambiente do Google Colab não possui integração nativa com essa ferramenta, devido à necessidade de instalações prévias do Apache Spark e do Java. Dessa forma, o primeiro passo consistiu na instalação do Java[4], seguido pela instalação do Apache Spark[5]. Posteriormente, foram realizadas configurações específicas no ambiente, viabilizando o uso da biblioteca Findspark. Esta biblioteca simplifica a localização e integração do Spark no ambiente Python do Google Colab, e consequentemente, facilita o processo de configuração e utilização do PySpark.

Para dar início à execução de operações utilizando o Spark, é necessária a inicialização de uma sessão. Esta sessão é configurada para operar em um ambiente local e é armazenada em uma variável designada. Este procedimento assegura que todas as operações subsequentes realizadas com o Spark sejam efetuadas por meio desta sessão previamente estabelecida e armazenada.

3.2.2 Transformações

Nesta etapa do processo, foi essencial consultar o dicionário de dados[6] para uma compreensão detalhada do significado de cada coluna presente no *dataset*. Com base nesse entendimento, iniciaram-se as transformações necessárias nos dados. A primeira etapa de filtragem abordou uma coluna específica que indicava se um registro havia sido excluído e, portanto, não era mais válido. Consequentemente, todos os registros marcados como excluídos foram removidos do dataframe. Após essa ação inicial, procedeu-se com a remoção de outras colunas que serviam apenas como controle ou que não continham

[4] JDK 8

[5] versão 3.1.2

[6] Disponivel em: https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SGL/pdfs/dicionario-de-dados_e-sus-notifica-opendatasus-1.pdf

informações relevantes para a análise.

Outra etapa no processamento dos dados envolveu a complementação das informações de município e estado, que estavam ausentes em uma parcela dos registros. Essa lacuna nos dados representava um desafio, pois a localização geográfica é um fator essencial para o agrupamento das notificações. Para resolver essa questão, foi necessário recorrer a um conjunto adicional de dados, que continha as informações de município e estado vinculadas aos respectivos códigos do IBGE[7]. Este conjunto de dados permitiu a realização de uma intersecção eficaz entre os diferentes conjuntos de dados: os registros originalmente incompletos, contendo apenas o código do IBGE, foram enriquecidos com as informações detalhadas de localização.

Como etapa final do processo de transformação dos dados, implementou-se a técnica de *explode*[8], aplicada especificamente a duas colunas textuais referentes a 'sintomas' e 'condições'. Esta abordagem envolveu a conversão dessas colunas em múltiplas colunas, seguindo o método conhecido como One-Hot Encoding. Cada sintoma e condição foi representado por uma coluna individual, onde os valores 0 ou 1 indicavam, respectivamente, a ausência ou presença de determinado sintoma ou condição em um registro. Após essa transformação, as colunas originais foram removidas, mantendo-se apenas as novas colunas geradas por este processo, facilitando assim análises subsequentes.

Durante esta etapa do processamento dos dados, algumas colunas foram mantidas no conjunto de dados, apesar da incerteza inicial sobre a sua relevância para a análise. Estas colunas contêm informações variadas, incluindo os tipos de testes realizados, as datas de administração das doses da vacina, e os diferentes tipos de vacinas utilizadas. Esta decisão foi tomada considerando a possibilidade de que tais dados pudessem revelar-se úteis em etapas posteriores do trabalho. Além disso, foram preservadas colunas que já haviam sido identificadas como essenciais para a análise, como a idade dos pacientes, o gênero, a data da notificação do caso e a evolução clínica do mesmo.

3.3 ANÁLISE DOS DADOS

A última etapa deste trabalho consiste em utilizar os dados já processados para extrair informações para o objetivo do mesmo. Esta etapa envolve a criação de cenários específicos, o cálculo das distâncias entre as séries temporais para identificar padrões e similaridades, metrificar qual o melhor número de clusters e a apresentação desses resultados através de gráficos.

[7] Disponível em https://github.com/leogermani/estados-e-municipios-ibge

[8] Operação em processamento de dados onde uma coluna com listas é transformada, expandindo cada item da lista em uma nova linha.

3.3.1 Montagem de cenários

Para a aplicação efetiva de algoritmos que calculam medidas como a distância Euclidiana e o Dynamic Time Warping (DTW), é necessário adaptar o conjunto de dados às necessidades específicas dessas técnicas. Isso envolve selecionar e transformar as informações disponíveis em séries temporais estruturadas em torno de contagens específicas. Estas contagens representam os cenários que serão analisados, como, por exemplo, a incidência de óbitos diários por faixa etária em cada estado ou a ocorrência de óbitos diários associados a certos sintomas em cada estado. A utilização desta abordagem permite a especificação de uma variedade de cenários, embora se deva notar que um aumento nas combinações de variáveis resultará em um acréscimo proporcional no número de registros no novo *dataset*, por essa razão, foram feitas apenas combinações de até duas variáveis.

Para selecionar os estados que foram utilizados na montagem de cenários (e eventualmente no agrupamento) foi feita uma contagem de notificações da COVID-19 nos estados de cada região do Brasil. Com base nessa análise, representada na Tabela 1, selecionaram-se os estados com o maior número de notificações para as que se tivessem uma variedade de tipos de dados. Esta seleção permitiu realizar análises comparativas entre estados, mantendo a consideração regional, mas restringindo o foco aos estados mais representativos em termos de volume de dados. Os estados escolhidos foram o Rio Grande do Sul, representando a região Sul; Goiás, pelo Centro-oeste; Pará, pelo Norte; Bahia, pelo Nordeste; e Rio de Janeiro, Minas Gerais e São Paulo, representando a região Sudeste. Estes últimos três foram incluídos por serem os estados com o maior número de notificações no país.

Tabela 1 – Quantidade de registros nos datasets de 2020.

Região	Estado	Quantidade
Norte	AC	134.746
Norte	AM	517.084
Norte	AP	118.520
Norte	**PA**	**642.070**
Norte	RO	354.060
Norte	RR	168.684
Norte	TO	237.102
Nordeste	AL	365.706
Nordeste	**BA**	**1.791.659**
Nordeste	CE	955.567
Nordeste	MA	576.306
Nordeste	PE	871.716
Nordeste	PI	432.571
Nordeste	RN	573.523
Nordeste	SE	263.922
Centro-oeste	DF	406.490
Centro-oeste	**GO**	**1.024.096**
Centro-oeste	MS	569.474
Centro-oeste	MT	347.531
Sudeste	ES	352.639
Sudeste	**MG**	**2.497.491**
Sudeste	**RJ**	**1.969.837**
Sudeste	**SP**	**3.637.723**
Sul	PR	1.147.099
Sul	**RS**	**2.405.838**
Sul	SC	2.185.503

Fonte: Próprio autor

3.3.1.1 Seleção de características

Após a seleção de sete estados, cada um representado por um *dataset* anual, procedeu-se à concatenação desses conjuntos de dados por ano, formando um único Dataframe. Dentro deste Dataframe, uma etapa a mais foi feita, a identificação das características mais relevantes para a evolução dos casos de COVID-19, conhecida como *feature importance*. Para tal, foram necessárias transformações em colunas textuais, utili-

zando ferramentas específicas do PySpark, como o *StringIndexer* e o *VectorAssembler*. O *StringIndexer* converteu textos em valores numéricos, enquanto o *VectorAssembler* combinou as colunas para formar vetores de características para cada uma. Finalmente, com os dados adequadamente preparados, aplicou-se um modelo de *RandomForestClassifier*. Dessa forma, obtém-se uma atribuição de pesos de importância para cada característica (os pesos são normalizada de modo que a soma de todas as importâncias das características seja igual a 1), de acordo com a ajuda para melhorar a precisão das previsões das árvores de decisão dentro do modelo, elucidando assim os fatores mais significativos na evolução dos casos.

Os resultados detalhados desta etapa de análise são abordados e discutidos na seção 4.1 do trabalho. As características identificadas e selecionadas para o estudo incluem a idade do paciente, os sintomas apresentados, as condições pré-existentes, o sexo, e a informação sobre se o indivíduo é um profissional da saúde.

3.3.1.2 Agrupamento por características

No agrupamento pelas características, utilizou-se o Dataframe original (apenas com os estados concatenados) como base e procedeu-se à seleção das colunas correspondentes às características previamente determinadas como relevantes. Adicionalmente, implementou-se um filtro para selecionar às notificações que culminaram em óbito.

Nesta parte, com a redução significativa tanto no número de colunas quanto de linhas, tornou-se viável a conversão do conjunto de dados para um dataframe Pandas. Isso facilitou o manuseio e a manipulação dos dados. As Tabelas 2 e 3 ilustram detalhadamente o número de notificações de óbitos para cada um dos estados selecionados.

Estado	**Notificações de óbito**
Bahia	10.497
São Paulo	6.508
Minas Gerais	5.646
Rio de Janeiro	3.657
Pará	3.147
Rio Grande do Sul	818
Goiás	354

Tabela 2 – Número de notificações de óbito por estado em 2020

Estado	Notificações de óbito
Bahia	14222
São Paulo	14404
Minas Gerais	13526
Rio de Janeiro	6663
Pará	4548
Rio Grande do Sul	1696
Goiás	1127

Tabela 3 – Número de notificações de óbito por estado em 2021

A partir do Dataframe Pandas, foram desenvolvidos diversos Dataframes secundários, cada um representando a contagem semanal de óbitos para cada estado em uma característica específica anualmente. Dessa forma, obtendo cinco Dataframes distintos (um para cada característica) para cada ano. Para padronizar e facilitar as comparações entre os estados, aplicou-se a cada um desses Dataframes a normalização *MinMaxScaler*. Assim, os dados estavam prontos para o cálculo das distâncias entre as séries temporais e, posteriormente, para a realização do agrupamento hierárquico.

3.3.2 Aplicação das medidas de distância

Conforme estabelecido, os algoritmos selecionados para calcular as distâncias em séries temporais foram a distância Euclidiana e o DTW. Nesta etapa, os Dataframes gerados anteriormente para cada característica são utilizados, agora convertidos para o formato de Dataframes do Pandas. Esta conversão foi realizada para assegurar uma maior compatibilidade com bibliotecas especializadas, como a numpy[9] e a dtaidistance[10], esta última necessária para o cálculo do DTW. Para cada Dataframe e característica selecionada, são realizados cálculos de medida de distância tanto pela distância Euclidiana quanto pelo DTW, resultando em um total de 10 medidas distintas de distância para cada ano analisado.

No processo de cálculo da distância Euclidiana, utiliza-se o Dataframe correspondente como um dos argumentos da função *linkage()*[11]. Esta função, por padrão, emprega a distância Euclidiana em sua implementação. Conforme discutido na subseção 2.1.2, é necessário especificar um tipo de função *linkage*, que constitui outro argumento necesário para ser passado. Para esta análise baseada na distância Euclidiana, optou-se pelo método de Ward como a função de *linkage*. Esta escolha justifica-se pelo fato de que o método de Ward se concentra em minimizar a variação interna dos clusters, uma abordagem que

[9] https://numpy.org/

[10] https://pypi.org/project/dtaidistance/

[11] Função para cálculo do linkage do pacote SciPy, disponível em: https://docs.scipy.org/doc/scipy/

é coerente com a lógica da distância Euclidiana.

Diferentemente do cálculo da distância Euclidiana, o processo para calcular a distância DTW exige uma etapa adicional antes da aplicação da função *linkage()*. Foi desenvolvida uma função específica para construir a matriz de distâncias DTW, a qual calcula a distância entre pares de séries temporais, detalhada no Algoritmo 1. Esta matriz de distâncias obtida é então utilizada como argumento para a função *linkage()*, porém, neste caso, seleciona-se o tipo *single* como método de *linkage*. A escolha do método *single*, que foca na menor distância entre quaisquer dois pontos, é feita para preservar as características únicas capturadas pelo DTW, garantindo assim a fidelidade do agrupamento às nuances das séries temporais analisadas.

Algoritmo 1: Cálculo da Distância DTW

Função `distance_dtw(`*dados*`)`**:**

Entrada: dados // uma matriz de seriés temporais para calcular a distância DTW

Saída : dist_matrix // matriz de distância DTW

n $\leftarrow$ número de linhas em dados

dist_matrix $\leftarrow$ criar matriz $n \times n$, inicializada com zeros

para *cada par de linhas* i, j *em dados* **faça**

dtwDistância $\leftarrow$ calcular DTW entre a linha i e a linha j

dist_matrix[i, j] $\leftarrow$ dtwDistância

fim

return *dist_matrix*

Após a execução dos cálculos de distância, obtiveram-se as matrizes de ligação, que representam o agrupamento hierárquico dos dados. Cada uma dessas matrizes foi utilizada como entrada para uma função dedicada à determinação do número ideal de clusters. Neste processo, explorou-se a formação de clusters variando de 2 a 12, e para cada um, calculou-se o coeficiente de silhueta usando a função *silhouette_score()*[12]. O número de clusters que resultava no coeficiente de silhueta mais elevado era selecionado como o ideal. Essa função, portanto, retornava o melhor número de clusters (entre os 11 testados), o valor correspondente do coeficiente de silhueta, e o *threshold* necessário para cortar o dendrograma no eixo das distâncias e formar o número desejado de clusters. Estes resultados são detalhados na seção 4.2. Um exemplo do código utilizado é encontrado no Algoritmo 2.

[12]Disponível no pacote scikit-learn: https://scikit-learn.org/stable/

Algoritmo 2: Determinação de Limiar e Clusters

Função `get_threshold_and_clusters(`*linked, original_dataframe, metric*`)`**:**

Entrada: linked // matriz de ligação do agrupamento hierárquico
original_dataframe - dataframe original
metric // métrica utilizada para calcular o score de silhueta

Saída : Número de clusters, limiar e a melhor pontuação

melhorPontuacao $\leftarrow -1$
numClusters, limiar $\leftarrow$ *None*
para i *de* 2 *a* 12 **faça**
  rotulosClusters $\leftarrow$ cortar dendrograma com i clusters
  mediaSilhouette $\leftarrow$ calcular média do Silhouette Score com rotulosClusters
  se *mediaSilhouette > melhorPontuacao* **então**
    melhorPontuacao $\leftarrow$ mediaSilhouette
    numClusters $\leftarrow i$
  fim
fim
limiar $\leftarrow$ determinar limiar a partir do dendrograma
return *numClusters, limiar, melhorPontuacao*

Por fim, com as matrizes de distância e dos *thresholds* definidos, procedeu-se à exibição dos dendrogramas utilizando a função *dendrogram()* da biblioteca SciPy. Esta etapa serviu para ilustrar visualmente a estrutura do agrupamento hierárquico e então serem feitas as devidas observações. Os dendrogramas também foram configurados para incorporar os *thresholds* calculados, permitindo assim a diferenciação cromática de cada um dos grupos principais para cada cenário, dentro dos quais se localizam os subgrupos hierárquicos. Esta abordagem não apenas preserva as relações e similaridades entre os dados, mas também destaca padrões potenciais que sugerem a formação de diversos clusters maiores, como é discutido no seguinte capítulo.

4 RESULTADOS

Os resultados apresentados e discutidos neste capítulo estão organizados em duas seções distintas. A primeira seção é uma justificativa para a escolha das características selecionadas no capítulo anterior, as quais foram estabelecidas, portanto, no objetivo deste trabalho. Enquanto a segunda seção está elaborada para demonstrar os resultados finais, que visam cumrpir efetivamente com o objetivo deste trabalho.

4.1 RESULTADOS DA FEATURE IMPORTANCE

Como descrito no capítulo da metodologia, a execução de uma análise de *feature importance* (Subseção 3.3.1.1) revelou-se essencial para identificar as características mais influentes na evolução clínica de cada caso de COVID-19 entre os datasets utilizados. A partir dessa informação, foi possível estruturar cenários distintos, agrupando os estados brasileiros com base em várias dessas características. Este procedimento estabeleceu a base para a realização de agrupamentos hierárquicos. Os resultados desta análise são apresentados e discutidos para cada ano analisado, 2020 e 2021.

4.1.1 Análise de 2020

Os resultados das características com os maiores pesos da análise de *feature importance* para o ano de 2020 são apresentados no Quadro 2. Observou-se que a característica 'racaCor_indexed'[1] emergiu como a mais significativa, com um valor de importância de 0,2919. Tal resultado indica uma possível correlação substancial entre a raça/cor e a evolução dos casos de COVID-19. Entretanto, devido à complexidade inerente a esta característica e às nuances que ela implica, decidiu-se não incluí-la na análise subsequente. As características 'idade' e 'profissionalSaude_indexed' seguiram em importância, com valores de 0,1598 e 0,1569, respectivamente. Estes resultados demonstram a influência da idade dos pacientes e o risco aumentado enfrentado pelos profissionais de saúde devido à sua exposição direta ao vírus.

Dentro da análise, sintomas como 'tosse' e 'outros' emergiram como elementos de relevância, apresentando valores de importância de 0,0580 e 0,0459, respectivamente. No entanto, para a análise de sintomas e condições, optou-se por excluir os registros marcados como 'outros', priorizando os cinco sintomas e condições mais comumente reportados

[1] O sufixo 'indexed' indica a conversão de campos textuais para valores numéricos.

nas notificações. É importante notar que a característica 'sexo_indexed' também se destacou em termos de importância, revelando-se significativamente influente na análise, mais do que outras características subsequentes na lista.

Além disso, condições de saúde adicionais, tais como 'imunossupressão' e 'doenças cardíacas crônicas', foram identificadas como fatores relevantes na análise. Apesar de apresentarem valores de importância menores em comparação com outros fatores.

Feature	**Importance**
racaCor_indexed	0,2919
idade	0,1598
profissionalSaude_indexed	0,1569
condicao: outros	0,0848
sintoma: tosse	0,0580
sintoma: outros	0,0459
sexo_indexed	0,0345
condicao: imunossupressao	0,0235
profissionalSeguranca_indexed	0,0196
sintoma: dor de garganta	0,0119
sintoma: febre	0,0115
sintoma: dificuldade de respirar	0,0115
condicao: puerpera (ate 45 dias do parto)	0,0111
condicao: gestante de alto risco	0,0109
sintoma: coriza	0,0095
sintoma: assintomatico	0,0078
sintoma: dispneia	0,0075
condicao: doencas cardiacas cronicas	0,0069
condicao: doencas respiratorias cronicas descompensadas	0,0064
condicao: diabetesdoencas respiratorias cronicas descompensadas	0,0055
condicao: gestante	0,0054
condicao: obesidade	0,0045
sintoma: dor de cabeca	0,0034
sintoma: disturbios olfativos	0,0033
condicao: portador de doencas cromossomicas ou estado de fragilidade imunologica	0,0030
condicao: doencas cardoacas cronicas	0,0021
condicao: doencas renais cronicas em estagio avancado (graus 3;4 ou 5)	0,0014
condicao: diabetes	0,0008
condicao: doencas renais cronicas	0,0004
condicao: doencas cardoacas cronicasdoencas respiratorias cronicas descompensadas	0,0002
sintoma: disturbios gustativos	0,0000

Quadro 2 – Feature Importance para o ano de 2020

4.1.2 Análise de 2021

Os resultados das características com os maiores pesos do ano de 2021 são apresentados no Quadro 3. Observou-se uma redução geral no número de recursos considerados como importantes, embora alguns tenham mantido sua relevância, ainda que com alterações nos graus de importância. Notavelmente, 'racaCor_indexed' permaneceu como a característica de maior importância, com um valor de 0,3449, seguida pela 'idade', com um valor de 0,2241.

Observou-se um aumento na relevância de certos sintomas, especificamente 'dor de garganta', 'febre' e 'dor de cabeça', que apresentaram valores de importância de 0,1306, 0,1030 e 0,0774, respectivamente. Esta tendência está em consonância com a observação de que esses sintomas estão entre os cinco mais comumente reportados, justificando assim sua inclusão e consideração no processo de agrupamento hierárquico.

Feature	**Importance**
racaCor_indexed	0,3449
idade	0,2241
sintoma: dor de garganta	0,1306
sintoma: febre	0,1030
sintoma: dor de cabeca	0,0774
sintoma: coriza	0,0541
sintoma: assintomatico	0,0266
condicao: gestante	0,0119
sintoma: outros	0,0111
sintoma: dispineia	0,0100
profissionalSeguranca_indexed	0,0023
condicao: doencas respiratorias cronicas descompensadas	0,0022
sexo_indexed	0,0011
condicao: outros	0,0007

Quadro 3 – Feature Importance para o ano de 2021

4.1.3 Análise geral

A análise dos recursos identificados no estudo possibilitou a definição das características mais determinantes na evolução clínica dos casos de COVID-19. A idade, destacando-se como um dos elementos mais relevantes, foi prontamente selecionada. Quanto aos sintomas e condições, uma abordagem detalhada de cada valor foi adotada para avaliar seus impactos, levando à consideração dessas variáveis como um todo para

o agrupamento. Posteriormente, foram filtrados apenas os casos que apresentavam pelo menos um dos cinco sintomas ou condições mais comuns entre todas notificações. O sexo, por outro lado, foi incorporado mais como um elemento de validação dos modelos analíticos, dado que não influencia diretamente a evolução clínica. Assim, esperou-se que os agrupamentos relativos ao sexo deveriam ser agrupados por estado brasileiro.

4.2 RESULTADOS DOS AGRUPAMENTOS

Os quadros subsequentes (4 e 5) apresentam os resultados da fase de análise em que os dados foram submetidos a um processo de avaliação do número de clusters, variando o número de clusters entre 2 a 12. Para cada configuração de clusters, foi avaliado o coeficiente de silhueta, sendo o valor mais elevado utilizado para determinar o melhor número de clusters. Associado a cada número ideal de clusters, é determinado um valor de *threshold*, que indica o ponto de corte no dendrograma para a definição dos clusters. É importante ressaltar que esta etapa de avaliação sucede o cálculo do *linkage*, o que significa que o agrupamento hierárquico já está estabelecido, tanto para os cálculos baseados na distância Euclidiana quanto no DTW.

Com o melhor número de clusters definido para cada situação, procedeu-se à elaboração dos dendrogramas para uma visualização detalhada das relações e distâncias, destacando-se os grupos distintos por meio de cores. A análise subsequente concentra-se nos dendrogramas referentes à faixa etária, sintomas, condições e sexo. Para os demais dendrogramas, os leitores são encorajados a consultar o apêndice, onde estão disponibilizados para consulta.

Também foram elaborados matrizes de contagem para cada análise, assim demonstrando as relações diretas entre caracterísitcas, ou seja, as relações de primeira ordem entre os registros.

Agrupamento	Dist. Euclidiana		DTW	
	Coef. de silhueta	**Núm. clusters**	**Coef. de silhueta**	**Núm. clusters**
Estados + faixa etária	0,3365	5	0,3494	2
Estados + sintomas	0,5491	2	0,4916	5
Estados + condições	0,6301	2	0,7207	2
Estados + sexo	0,5496	2	0,5201	2
Estados + profissional da saúde	0,7494	2	0,7832	2

Quadro 4 – Métricas dos agrupamentos para 2020

Agrupamento	Dist. Euclidiana		DTW	
	Coef. de silhueta	**Núm. clusters**	**Coef. de silhueta**	**Núm. clusters**
Estados + faixa etária	0,2666	2	0,4709	2
Estados + sintomas	0,5084	4	0,4562	2
Estados + condições	0,6254	2	0,5255	6
Estados + sexo	0,4059	5	0,4599	2
Estados + profissional da saúde	0,5544	3	0,6549	2

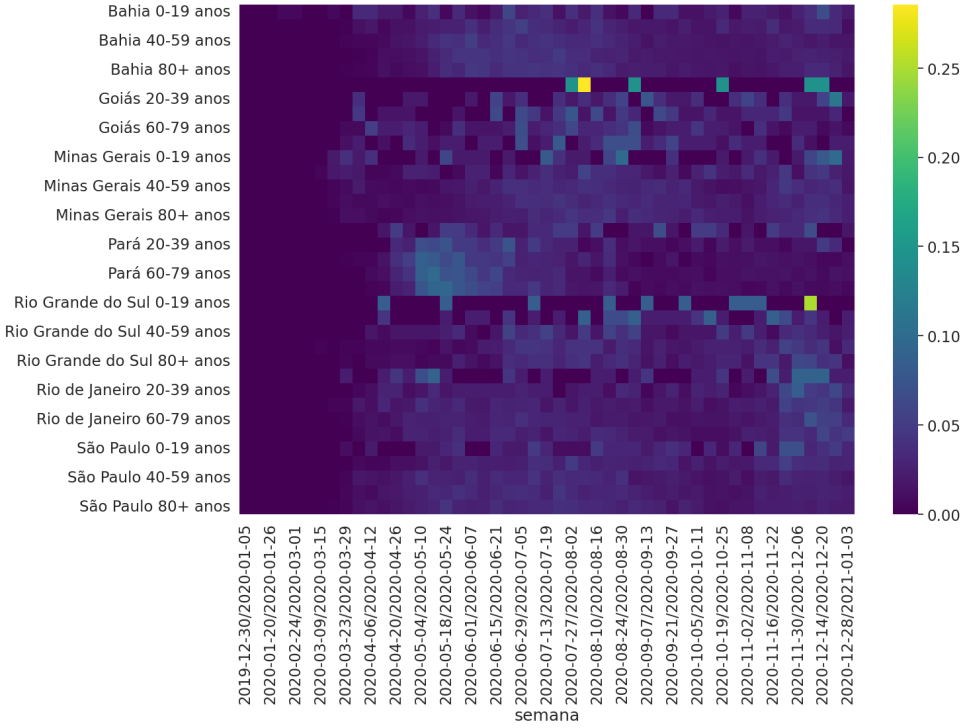
Quadro 5 – Métricas dos agrupamentos para 2021

4.2.1 Agrupamento estado + faixa etária

No contexto do agrupamento de estados com faixas etárias, mesmo após a normalização dos dados, identificou-se a presença de casos com distribuições extremamente desiguais, com ocorrências esporádicas, como apenas uma vez por mês. Tais casos foram classificados como outliers e, consequentemente, resultaram na formação de dois grupos

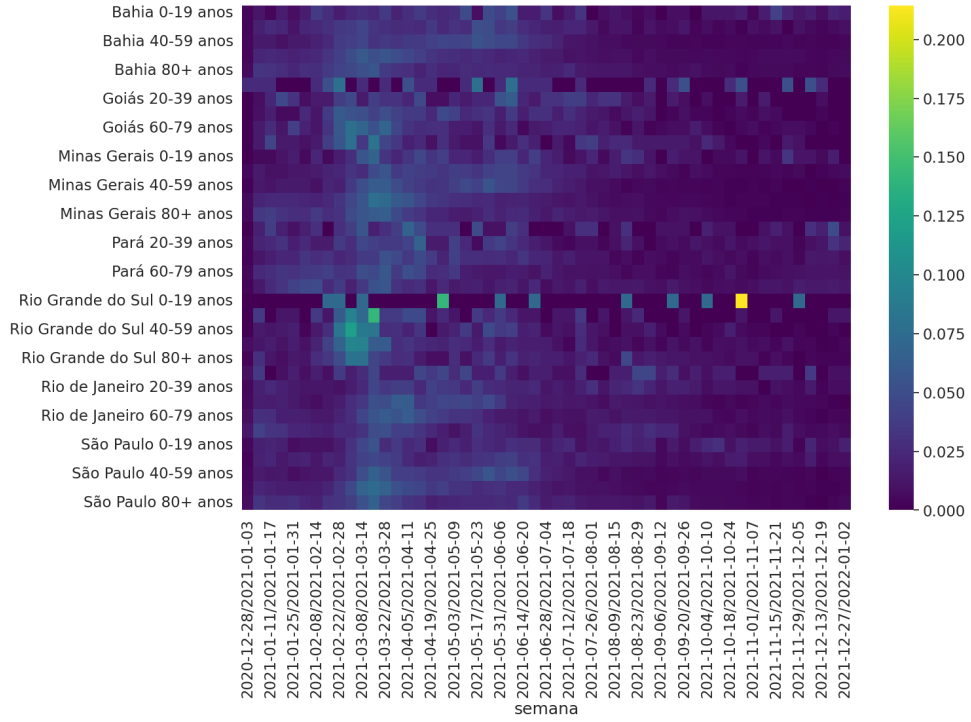
distintos nos agrupamentos hierárquicos, um dos quais representando exclusivamente esses outliers. Esta particularidade é evidenciada nos mapas de calor (Figura 7 e Figura 8), que ilustram as frequências de óbitos por estado e faixa etária ao longo das semanas em 2020 e 2021. Notadamente, dois grupos, especificamente Goiás e Rio Grande do Sul na faixa etária de 0-19 anos, apresentaram padrões anômalos, com múltiplos pontos escuros (indicativos de ausência de casos) com alguns pontos coloridos espaçados. Em vista dessas discrepâncias, optou-se pela remoção desses grupos dos dataframes de ambos os anos para a análise subsequente.

Figura 7 – Heatmap com distribuição dos óbitos ao longo das semanas para cada estado + faixa etária 2020.



Fonte: Próprio autor.

Figura 8 – Heatmap com distribuição dos óbitos ao longo das semanas para cada estado + faixa etária 2021.



Fonte: Próprio autor.

Os dendrogramas com os quadros dos respectivos valores estão separados em duas subseções para cada um dos anos ser detalhados. O conjunto de estado + faixa etária é referido como registro, e grupo refere-se aos grandes grupos montados nos dendrogramas.

4.2.1.1 Ano 2020

A análise do dendrograma (Figura 9), produzido com base na distância Euclidiana, revelou a formação de cinco grandes clusters, que englobam os agrupamentos hierárquicos internos, mantendo as similaridades entre eles. É interessante notar que os grupos 4 e 5 são constituídos por apenas uma combinação de estado e faixa etária cada: Rio Grande do Sul na faixa etária de 20-39 anos e Rio de Janeiro na faixa etária de 0-19 anos, respectivamente. Os três grupos restantes abrangem registros que demonstram maior similaridade interna, conforme indicado pelas distâncias menores entre eles. Por exemplo, o grupo 1 é exclusivamente formado por registros do estado do Pará. Já o grupo 3, que apresenta o maior número de registros, exibe uma interessante característica em seus níveis inferiores no dendrograma, sugerindo a preservação das relações internas de cada estado com suas respectivas faixas etárias.

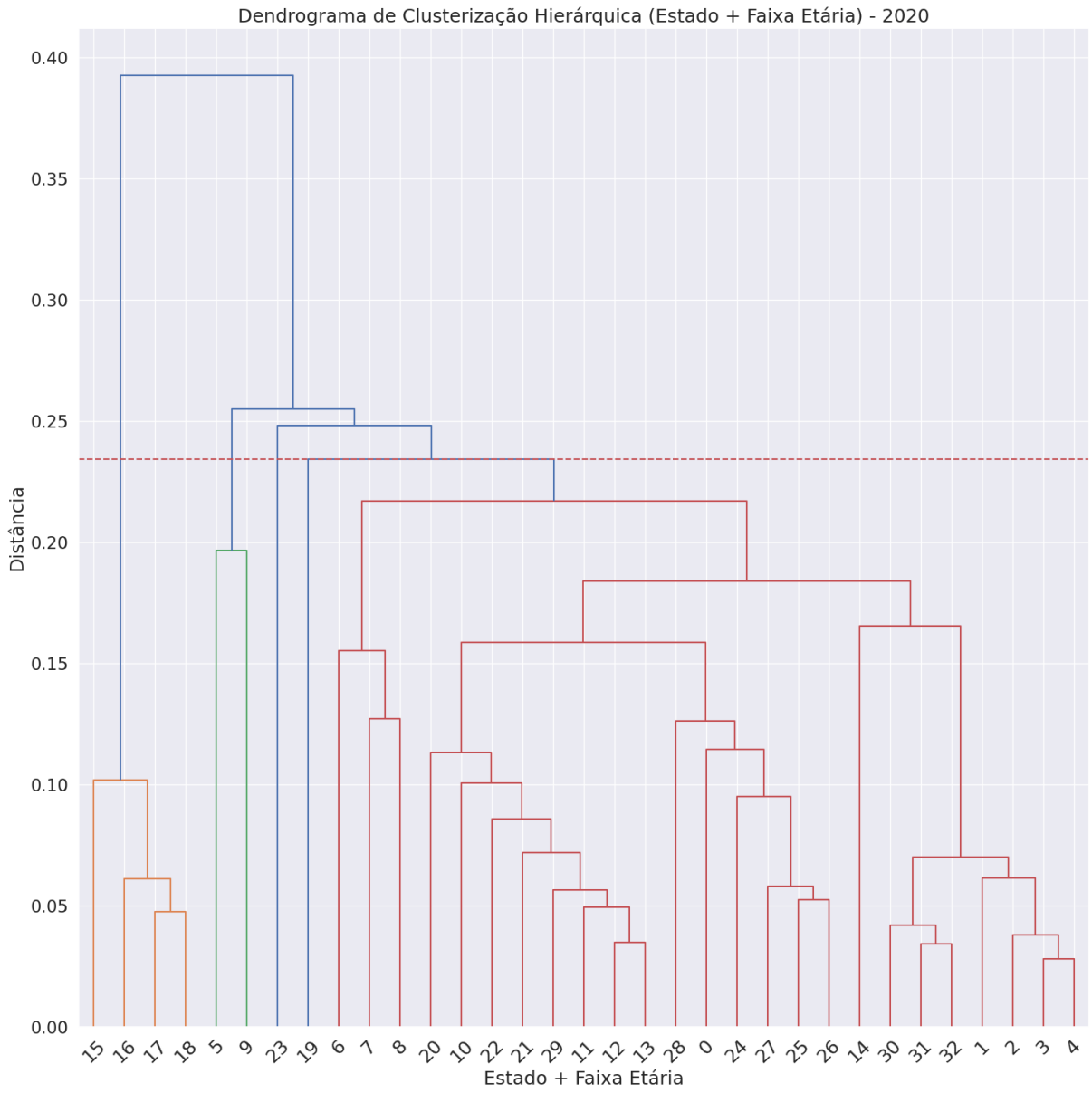
No que concerne à análise utilizando o DTW (Figura 10), observou-se a definição

de apenas dois grupos principais. Esta configuração foi influenciada pela presença de um outlier. Mesmo considerando a remoção desse registro específico, a estrutura de dois grupos do dendrograma permaneceria inalterada, pois percebe-se um formato escalonado no mesmo. Tal disposição resulta em uma complexidade adicional na interpretação dos níveis inferiores do dendrograma, onde as relações entre os dados são menos discerníveis e, consequentemente, dificultam a extração de insights claros.

Uma possível explicação para as dificuldades encontradas na elaboração de um agrupamento hierárquico eficaz para o ano de 2020 reside na significativa discrepância na contagem de óbitos entre os diferentes estados. Esta variação é evidenciada na Tabela 2, onde se observam diferenças substanciais nos números de óbitos reportados por cada estado, fator que pode ter impactado diretamente na formação dos agrupamentos e na clareza das relações estabelecidas entre eles.

De qualquer forma, consegue-se observar algumas relações diretas através das matrizes de contagem (Figuras 11 e 12), em que foram feitas as contagens de relações de primeira ordem entre as faixas etárias para cada medida de distância (Euclidiana e DTW). Uma característica evidente em ambas é a relação direta da faixa etária dos 60-79 anos com 80+ sendo a mais forte, e essas faixas etárias se relacionando dentro do mesmo estado.

Figura 9 – Dendrograma dos dados de estado + faixa etária 2020 com medida de distância temporal sendo distância Euclidiana.

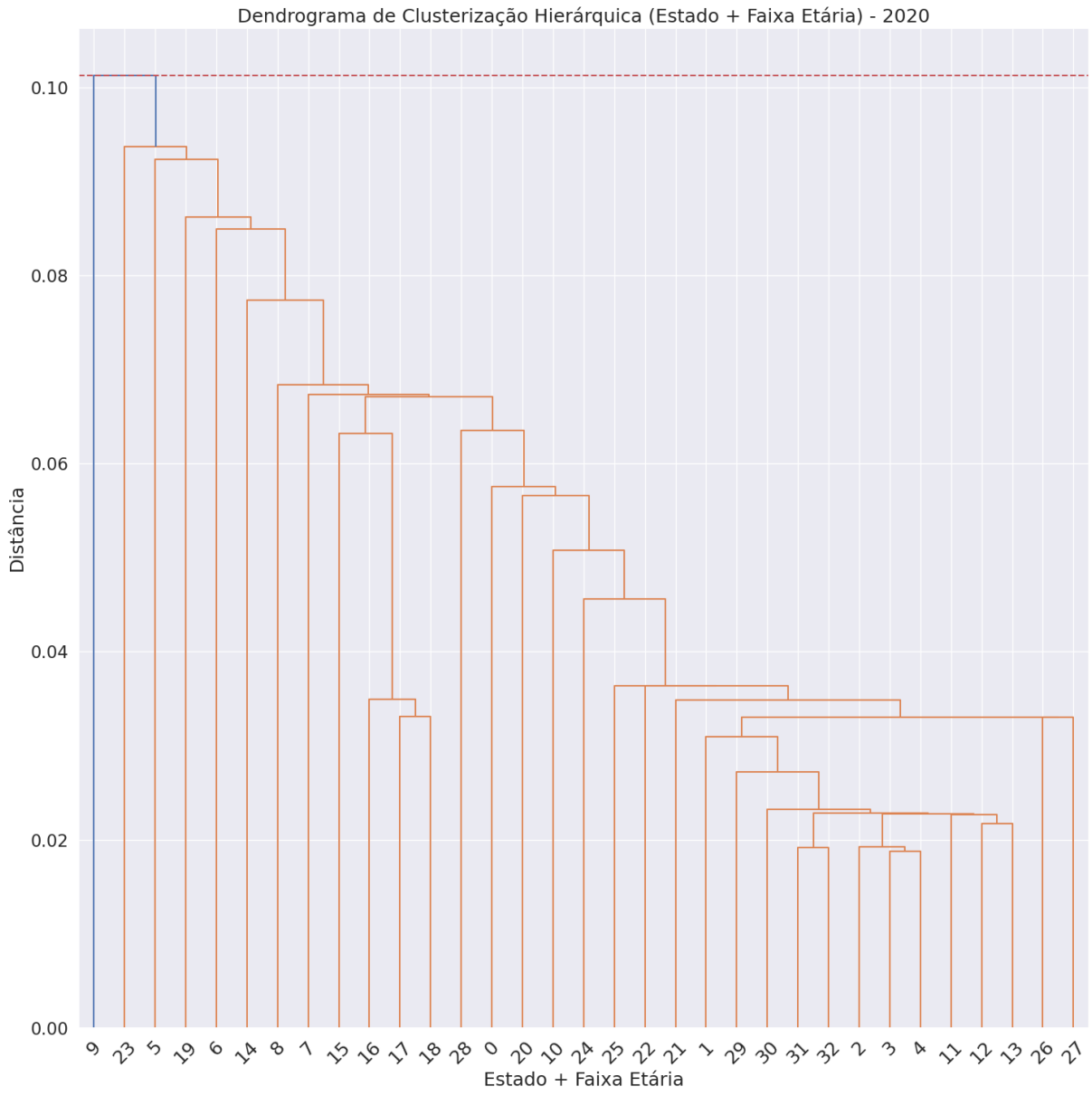


Fonte: Próprio autor.

Cluster	Registros
1	**15:** Pará 20-39 anos, **16:** Pará 40-59 anos, **17:** Pará 60-79 anos, **18:** Pará 80+ anos
2	**5:** Goiás 20-39 anos, **9:** Minas Gerais 0-19 anos
3	**0:** Bahia 0-19 anos, **1:** Bahia 20-39 anos, **2:** Bahia 40-59 anos, **3:** Bahia 60-79 anos, **4:** Bahia 80+ anos, **6:** Goiás 40-59 anos, **7:** Goiás 60-79 anos, **8:** Goiás 80+ anos, **10:** Minas Gerais 20-39 anos, **11:** Minas Gerais 40-59 anos, **12:** Minas Gerais 60-79 anos, **13:** Minas Gerais 80+ anos, **20:** Rio Grande do Sul 40-59 anos, **21:** Rio Grande do Sul 60-79 anos, **22:** Rio Grande do Sul 80+ anos, **24:** Rio de Janeiro 20-39 anos, **25:** Rio de Janeiro 40-59 anos, **26:** Rio de Janeiro 60-79 anos, **27:** Rio de Janeiro 80+ anos, **28:** São Paulo 0-19 anos, **29:** São Paulo 20-39 anos, **30:** São Paulo 40-59 anos, **31:** São Paulo 60-79 anos, **32:** São Paulo 80+ anos
4	**19:** Rio Grande do Sul 20-39 anos
5	**23:** Rio de Janeiro 0-19 anos

Quadro 6 – Quadro com a separação dos registros da Figura 9.

Figura 10 – Dendrograma dos dados de estado + faixa etária 2020 com medida de distância temporal sendo DTW.



Fonte: Próprio autor.

Cluster	Registros
1	**0:** Bahia 0-19 anos, **1:** Bahia 20-39 anos, **2:** Bahia 40-59 anos, **3:** Bahia 60-79 anos, **4:** Bahia 80+ anos, **5:** Goiás 20-39 anos, **6:** Goiás 40-59 anos, **7:** Goiás 60-79 anos, **8:** Goiás 80+ anos, **10:** Minas Gerais 20-39 anos, **11:** Minas Gerais 40-59 anos, **12:** Minas Gerais 60-79 anos, **13:** Minas Gerais 80+ anos, **14:** Pará 0-19 anos, **15:** Pará 20-39 anos, **16:** Pará 40-59 anos, **17:** Pará 60-79 anos, **18:** Pará 80+ anos, **19:** Rio Grande do Sul 20-39 anos, **20:** Rio Grande do Sul 40-59 anos, **21:** Rio Grande do Sul 60-79 anos, **22:** Rio Grande do Sul 80+ anos, **23:** Rio de Janeiro 0-19 anos, **24:** Rio de Janeiro 20-39 anos, **25:** Rio de Janeiro 40-59 anos, **26:** Rio de Janeiro 60-79 anos, **27:** Rio de Janeiro 80+ anos, **28:** São Paulo 0-19 anos, **29:** São Paulo 20-39 anos, **30:** São Paulo 40-59 anos, **31:** São Paulo 60-79 anos, **32:** São Paulo 80+ anos
2	**9:** Minas Gerais 0-19 anos

Quadro 7 – Quadro com a separação dos registros da Figura 10.

Figura 11 – Matriz de contagem dos relacionamentos de primeira ordem entre faixas etárias para distância Euclidiana em 2020.

	0-19 anos	20-39 anos	40-59 anos	60-79 anos	80+ anos
0-19 anos	0				
20-39 anos	1	0			
40-59 anos	0	0	0		
60-79 anos	0	0	1	0	
80+ anos	0	0	0	5	0

Fonte: Próprio autor.

Figura 12 – Matriz de contagem dos relacionamentos de primeira ordem entre faixas etárias para DTW em 2020.

	0-19 anos	20-39 anos	40-59 anos	60-79 anos	80+ anos
0-19 anos	0				
20-39 anos	0	0			
40-59 anos	0	0	0		
60-79 anos	0	0	0	0	
80+ anos	0	0	0	5	0

Fonte: Próprio autor.

4.2.1.2 Ano 2021

No dendrograma referente ao ano de 2021 (Figura 13), utilizando a distância Euclidiana, identificaram-se dois clusters distintos. O primeiro grupo abrange predominantemente os estados do Rio Grande do Sul e Goiás, destacando-se a presença dos registros (19) Rio Grande do Sul 20-39 anos e (20) Rio Grande do Sul 40-59 anos. Contudo, a análise sugere que a conexão neste grupo é mais fortemente influenciada pelas faixas etárias acima de 60 anos. Por outro lado, o segundo grupo revelou-se consideravelmente mais amplo. Uma observação detalhada das relações em níveis mais baixos do dendrograma revela padrões interessantes, tais como a associação de diferentes estados dentro da mesma faixa etária. Alguns exemplos incluem:

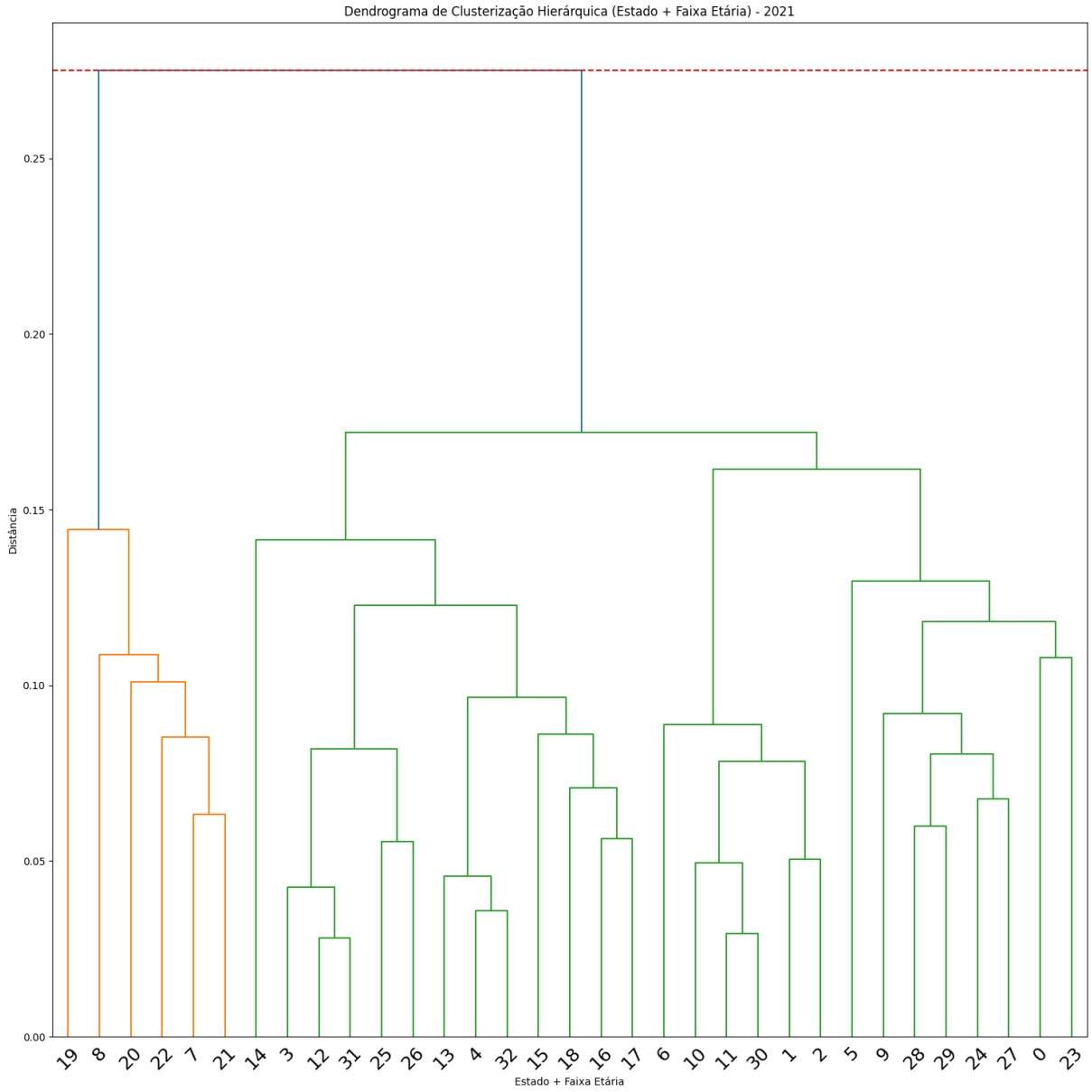
- (12) Minas Gerais 60-79 anos, (31) São Paulo 60-79 anos e (3) Bahia 60-79 anos;
- (4) Bahia 80+ anos e (32) São Paulo 80+ anos;
- (11) Minas Gerais 40-59 anos (30) São Paulo 40-59 anos;

Em relação ao dendrograma gerado pelo DTW para o ano de 2021 (Figura 14), observou-se novamente a formação de apenas dois grupos principais, caracterizados por uma estrutura escalonada semelhante à do ano anterior. No entanto, uma análise mais interna do grupo 1 revela uma tendência interessante: as faixas etárias mais elevadas tendem a estar inter-relacionadas entre os diferentes estados. Esta relação se manifesta de maneira gradativa, partindo dos subgrupos à direita do dendrograma (mais internos), que englobam as faixas etárias mais avançadas, progredindo em direção à esquerda, onde se localizam as faixas etárias mais jovens.

Esse padrão de faixas etárias relacionadas entre estados pode ser observado mais claramente na matriz de contagem da distância Euclidiana (Figura 15), onde percebe-se

que a diagonal principal (mesma faixa etária entre os estados) está preenchida, lembrando que são contadas somente relações diretas. O DTW (Figura 16) acabou tendo a matriz de contagem (ou relações diretas) prejudicada por estar no formato de relações gradativas.

Figura 13 – Dendrograma dos dados de estado + faixa etária 2021 com medida de distância temporal sendo distância Euclidiana.

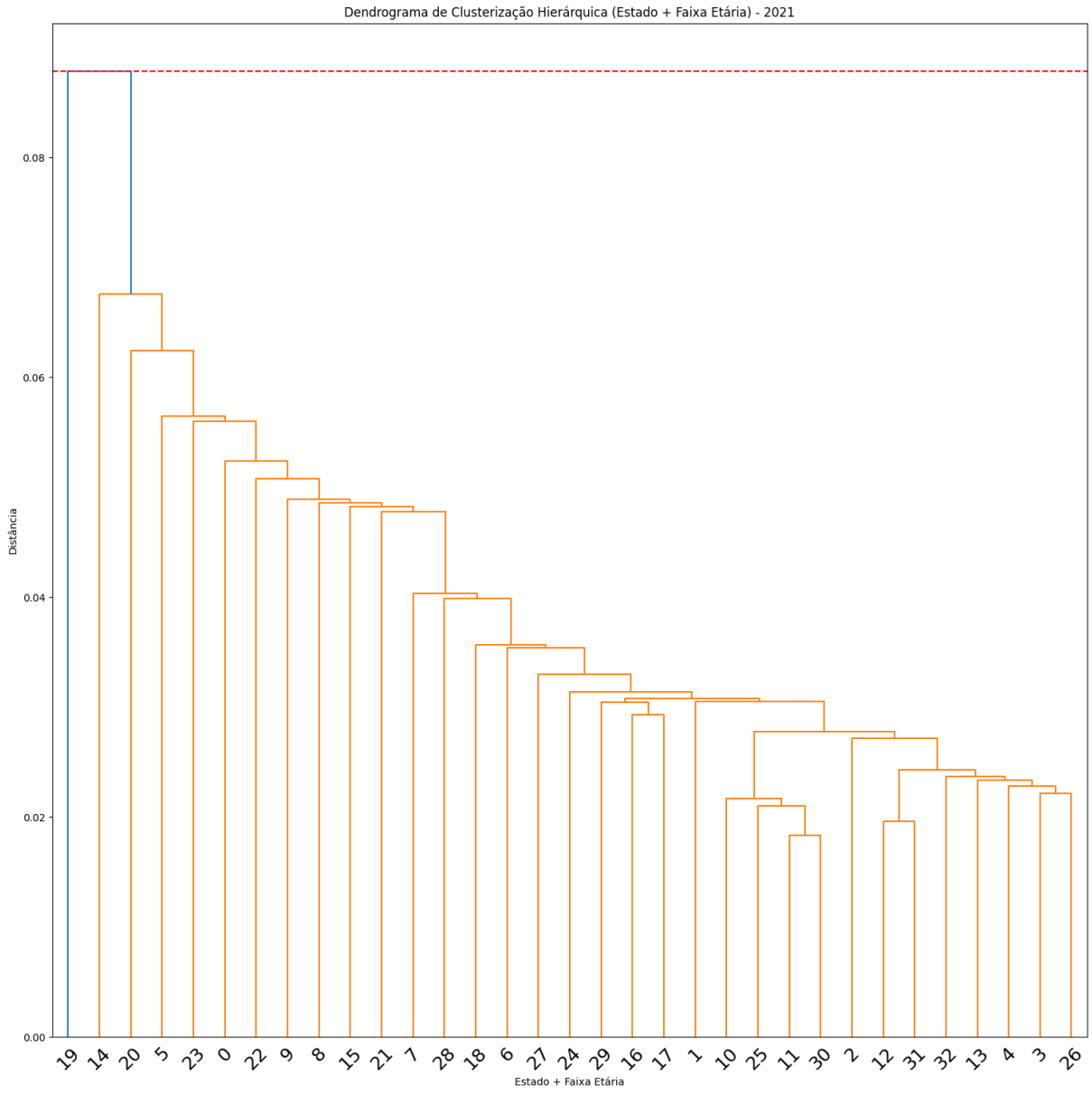


Fonte: Próprio autor.

Cluster	Registros
1	**7:** Goiás 60-79 anos, **8:** Goiás 80+ anos, **19:** Rio Grande do Sul 20-39 anos, **20:** Rio Grande do Sul 40-59 anos, **21:** Rio Grande do Sul 60-79 anos, **22:** Rio Grande do Sul 80+ anos
2	**0:** Bahia 0-19 anos, **1:** Bahia 20-39 anos, **2:** Bahia 40-59 anos, **3:** Bahia 60-79 anos, **4:** Bahia 80+ anos, **5:** Goiás 20-39 anos, **6:** Goiás 40-59 anos, **9:** Minas Gerais 0-19 anos, **10:** Minas Gerais 20-39 anos, **11:** Minas Gerais 40-59 anos, **12:** Minas Gerais 60-79 anos, **13:** Minas Gerais 80+ anos, **14:** Pará 0-19 anos, **15:** Pará 20-39 anos, **16:** Pará 40-59 anos, **17:** Pará 60-79 anos, **18:** Pará 80+ anos, **23:** Rio de Janeiro 0-19 anos, **24:** Rio de Janeiro 20-39 anos, **25:** Rio de Janeiro 40-59 anos, **26:** Rio de Janeiro 60-79 anos, **27:** Rio de Janeiro 80+ anos, **28:** São Paulo 0-19 anos, **29:** São Paulo 20-39 anos, **30:** São Paulo 40-59 anos, **31:** São Paulo 60-79 anos, **32:** São Paulo 80+ anos

Quadro 8 – Quadro com a separação dos registros da Figura 13.

Figura 14 – Dendrograma dos dados de estado + faixa etária 2021 com medida de distância temporal sendo DTW.



Fonte: Próprio autor.

Cluster	Registros
1	**0:** Bahia 0-19 anos, **1:** Bahia 20-39 anos, **2:** Bahia 40-59 anos, **3:** Bahia 60-79 anos, **4:** Bahia 80+ anos, **5:** Goiás 20-39 anos, **6:** Goiás 40-59 anos, **7:** Goiás 60-79 anos, **8:** Goiás 80+ anos, **9:** Minas Gerais 0-19 anos, **10:** Minas Gerais 20-39 anos, **11:** Minas Gerais 40-59 anos, **12:** Minas Gerais 60-79 anos, **13:** Minas Gerais 80+ anos, **14:** Pará 0-19 anos, **15:** Pará 20-39 anos, **16:** Pará 40-59 anos, **17:** Pará 60-79 anos, **18:** Pará 80+ anos, **20:** Rio Grande do Sul 40-59 anos, **21:** Rio Grande do Sul 60-79 anos, **22:** Rio Grande do Sul 80+ anos, **23:** Rio de Janeiro 0-19 anos, **24:** Rio de Janeiro 20-39 anos, **25:** Rio de Janeiro 40-59 anos, **26:** Rio de Janeiro 60-79 anos, **27:** Rio de Janeiro 80+ anos, **28:** São Paulo 0-19 anos, **29:** São Paulo 20-39 anos, **30:** São Paulo 40-59 anos, **31:** São Paulo 60-79 anos, **32:** São Paulo 80+ anos
2	**19:** Rio Grande do Sul 20-39 anos

Quadro 9 – Quadro com a separação dos registros da Figura 14.

Figura 15 – Matriz de contagem dos relacionamentos de primeira ordem entre faixas etárias para distância Euclidiana em 2021.

	0-19 anos	20-39 anos	40-59 anos	60-79 anos	80+ anos
0-19 anos	1				
20-39 anos	1	0			
40-59 anos	0	1	1		
60-79 anos	0	0	2	2	
80+ anos	0	1	0	0	1

Fonte: Próprio autor.

Figura 16 – Matriz de contagem dos relacionamentos de primeira ordem entre faixas etárias para DTW em 2021.

	0-19 anos	20-39 anos	40-59 anos	60-79 anos	80+ anos
0-19 anos	0				
20-39 anos	0	0			
40-59 anos	0	0	1		
60-79 anos	0	0	1	2	
80+ anos	0	0	0	0	0

Fonte: Próprio autor.

4.2.2 Agrupamento estado + sintoma

No agrupamento de estados combinados com sintomas, os dendrogramas sugerem uma disposição dos grupos mais equilibrada e distribuída. As relações hierárquicas dentro desses agrupamentos aparentam ser mais nítidas e bem definidas. É importante ressaltar que essa análise focou nos cinco sintomas mais comuns identificados em todos os estados. A partir dessa seleção, realizou-se o agrupamento hierárquico combinando esses sintomas com os respectivos estados.

Os dendrogramas com os quadros dos respectivos valores estão separados em duas subseções para cada um dos anos ser detalhados. O conjunto de estado + sintoma é referido como registro, e grupo refere-se aos grandes grupos montados nos dendrogramas.

4.2.2.1 Ano 2020

Na avaliação do agrupamento hierárquico baseado na distância Euclidiana para o ano de 2020 (Figura 17), identificaram-se dois grupos. O primeiro cluster é formado pelos estados de São Paulo, Minas Gerais e Bahia, mas especificamente em relação a três sintomas particulares: dispneia, febre e tosse. Essa forte correlação entre esses estados pode estar associada à notável diferença no número de óbitos registrados nesses estados em comparação com os demais para este ano. Pois observa-se que o grupo 2 apresenta algumas semelhanças, especialmente em relação a esses mesmos sintomas para cada estado, sugerindo a existência de um padrão consistente.

- (15) Pará - Dispneia, (18) Pará - Febre e (19) Pará - Tosse;
- (25) Rio de Janeiro - Dispneia, (28) Rio de Janeiro - Febre e (29) Rio de Janeiro - Tosse;

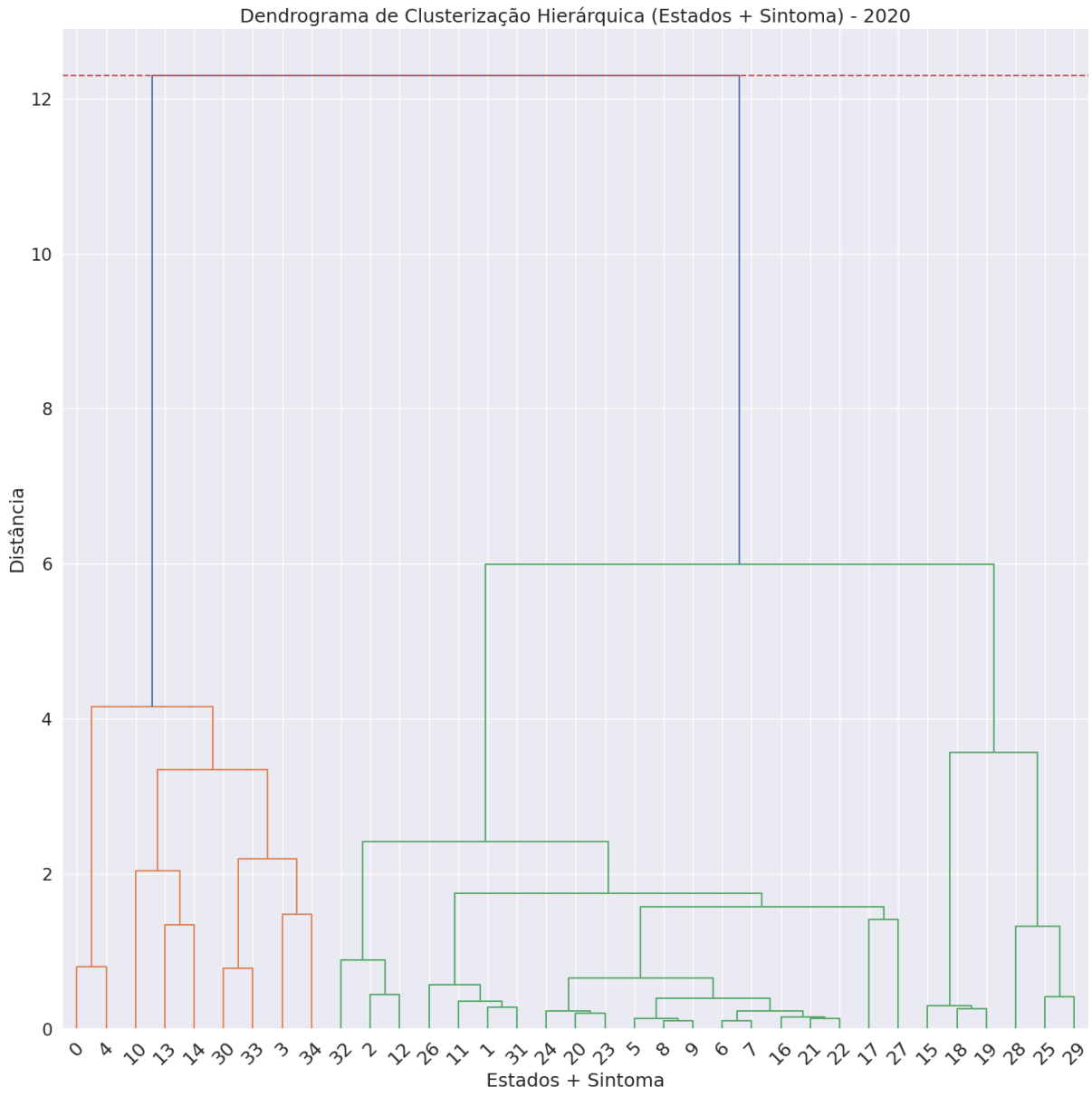
- (5) Goiás - Dispneia, (8) Goiás - Febre e (9) Goiás - Tosse;
- (20) Rio Grande do Sul - Dispneia, (23) Rio Grande do Sul - Febre e (24) Rio Grande do Sul - Tosse

No dendrograma resultante do DTW para o ano de 2020 (Figura 18), nota-se a formação de cinco grupos. O grupo 5, consistindo de apenas um registro, destaca-se como um outlier. Os grupos 2 e 4, por sua vez, exibem uma relação que remete aos achados na análise de distância Euclidiana, particularmente com os sintomas de dispneia, febre e tosse. Estes grupos parecem quase complementares; por exemplo, o grupo 4, que inclui (28) Rio de Janeiro - Febre e (34) São Paulo - Tosse, poderia teoricamente ser integrado ao grupo 2. A fusão desses dois grupos revela um padrão recorrente de sintomas entre os estados da Bahia, Minas Gerais, Rio de Janeiro e São Paulo, e até mesmo o estado do Pará, pertencente ao grupo 1, poderia ser considerado nessa união. Ao examinar o grupo 3, destaca-se o sintoma de dor de cabeça, que demonstra proximidade em sua ocorrência entre os diferentes estados.

- (1) Bahia - Dor de Cabeça e (11) Minas Gerais - Dor de Cabeça;
- (6) Goiás - Dor de Cabeça e (21) Rio Grande do Sul - Dor de Cabeça.

Estas relações entre febre, dispneia e tosse podem ser observadas como relações de primeira ordem nas Figuras 19 e 20.

Figura 17 – Dendrograma dos dados de estado + sintoma 2020 com medida de distância temporal sendo distância Euclidiana.

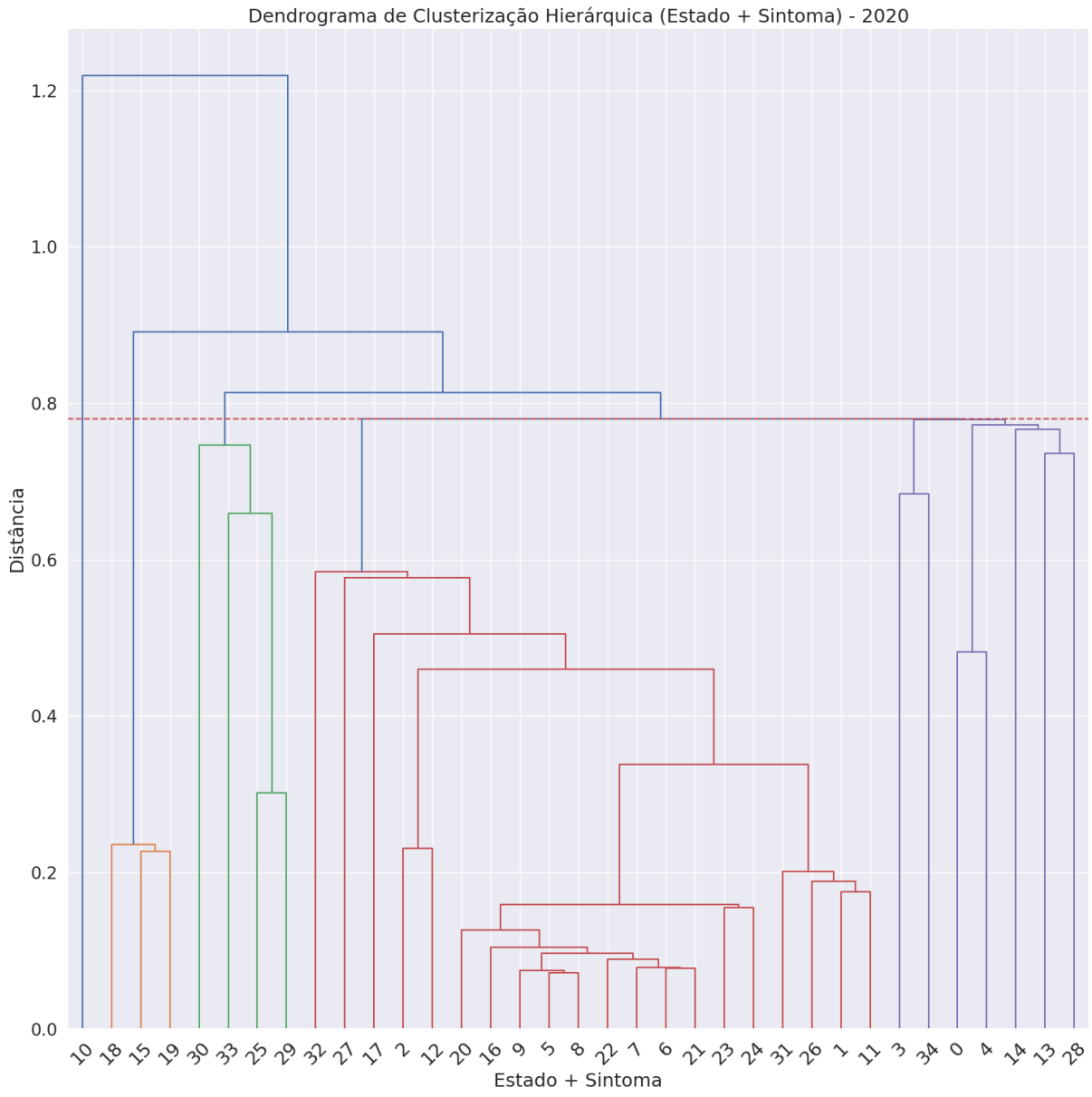


Fonte: Próprio autor.

Cluster	Registros
1	**0:** Bahia - Dispneia, **3:** Bahia - Febre, **4:** Bahia - Tosse, **10:** Minas Gerais - Dispneia, **13:** Minas Gerais - Febre, **14:** Minas Gerais - Tosse, **30:** São Paulo - Dispneia, **33:** São Paulo - Febre, **34:** São Paulo - Tosse
2	**1:** Bahia - Dor de Cabeça, **2:** Bahia - Dor de Garganta, **5:** Goiás - Dispneia, **6:** Goiás - Dor de Cabeça, **7:** Goiás - Dor de Garganta, **8:** Goiás - Febre, **9:** Goiás - Tosse, **11:** Minas Gerais - Dor de Cabeça, **12:** Minas Gerais - Dor de Garganta, **15:** Pará - Dispneia, **16:** Pará - Dor de Cabeça, **17:** Pará - Dor de Garganta, **18:** Pará - Febre, **19:** Pará - Tosse, **20:** Rio Grande do Sul - Dispneia, **21:** Rio Grande do Sul - Dor de Cabeça, **22:** Rio Grande do Sul - Dor de Garganta, **23:** Rio Grande do Sul - Febre, **24:** Rio Grande do Sul - Tosse, **25:** Rio de Janeiro - Dispneia, **26:** Rio de Janeiro - Dor de Cabeça, **27:** Rio de Janeiro - Dor de Garganta, **28:** Rio de Janeiro - Febre, **29:** Rio de Janeiro - Tosse, **31:** São Paulo - Dor de Cabeça, **32:** São Paulo - Dor de Garganta

Quadro 10 – Quadro com a separação dos registros da Figura 17.

Figura 18 – Dendrograma dos dados de estado + sintoma 2020 com medida de distância temporal sendo DTW.



Fonte: Próprio autor.

Cluster	Registros
1	**15:** Pará - Dispneia, **18:** Pará - Febre, **19:** Pará - Tosse
2	**25:** Rio de Janeiro - Dispneia, **29:** Rio de Janeiro - Tosse, **30:** São Paulo - Dispneia, **33:** São Paulo - Febre
3	**1:** Bahia - Dor de Cabeça, **2:** Bahia - Dor de Garganta, **5:** Goiás - Dispneia, **6:** Goiás - Dor de Cabeça, **7:** Goiás - Dor de Garganta, **8:** Goiás - Febre, **9:** Goiás - Tosse, **11:** Minas Gerais - Dor de Cabeça, **12:** Minas Gerais - Dor de Garganta, **16:** Pará - Dor de Cabeça, **17:** Pará - Dor de Garganta, **20:** Rio Grande do Sul - Dispneia, **21:** Rio Grande do Sul - Dor de Cabeça, **22:** Rio Grande do Sul - Dor de Garganta, **23:** Rio Grande do Sul - Febre, **24:** Rio Grande do Sul - Tosse, **26:** Rio de Janeiro - Dor de Cabeça, **27:** Rio de Janeiro - Dor de Garganta, **31:** São Paulo - Dor de Cabeça, **32:** São Paulo - Dor de Garganta
4	**0:** Bahia - Dispneia, **3:** Bahia - Febre, **4:** Bahia - Tosse, **13:** Minas Gerais - Febre, **14:** Minas Gerais - Tosse, **28:** Rio de Janeiro - Febre, **34:** São Paulo - Tosse
5	**10:** Minas Gerais - Dispneia

Quadro 11 – Quadro com a separação dos registros da Figura 18.

Figura 19 – Matriz de contagem dos relacionamentos de primeira ordem entre sintomas para distância Euclidiana em 2020.

	Tosse	Febre	Dispneia	Dor de cabeça	Dor de garganta
Tosse	0				
Febre	4	0			
Dispneia	2	2	0		
Dor de cabeça	0	0	0	1	
Dor de garganta	0	0	0	2	2

Fonte: Próprio autor.

Figura 20 – Matriz de contagem dos relacionamentos de primeira ordem entre sintomas para DTW em 2020.

	Tosse	Febre	Dispneia	Dor de cabeça	Dor de garganta
Tosse	0				
Febre	2	1			
Dispneia	3	1	0		
Dor de cabeça	0	0	0	2	
Dor de garganta	0	0	0	0	1

Fonte: Próprio autor.

4.2.2.2 Ano 2021

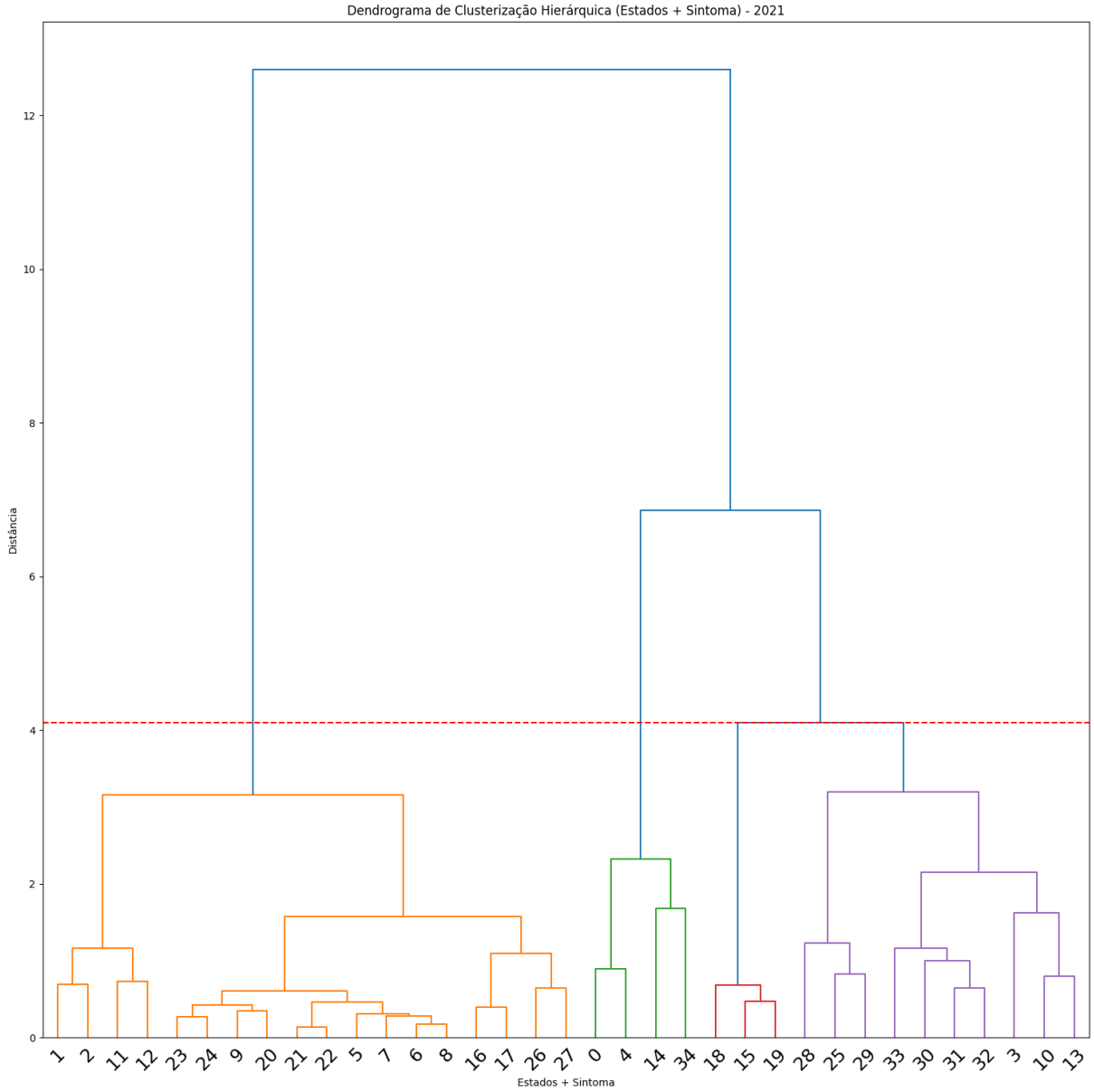
No ano de 2021, a análise utilizando a distância Euclidiana (Figura 21) resultou na formação de quatro agrupamentos. Notavelmente, o grupo 3 é exclusivamente composto pelo estado do Pará, apresentando uma associação concentrada nos sintomas de dispneia, febre e tosse. O grupo 2 revelou uma similaridade entre os estados da Bahia, Minas Gerais e São Paulo, particularmente no sintoma de tosse, além da inclusão do sintoma de dispneia associado à Bahia. Por outro lado, o grupo 1 caracterizou-se pela prevalência do sintoma de febre em estados como Bahia, Minas Gerais, Rio de Janeiro e São Paulo, com a inclusão ocasional dos sintomas de tosse e dispneia. Curiosamente, os sintomas de dor de cabeça e dor de garganta mostraram uma tendência a agrupar-se juntos para cada estado.

- (31) São Paulo - Dor de Cabeça e (32) São Paulo - Dor de Garganta;
- (16) Pará - Dor de Cabeça e (17) Pará - Dor de Garganta;
- (1) Bahia - Dor de Cabeça e (2) Bahia - Dor de Garganta;
- (11) Minas Gerais - Dor de Cabeça e (12) Minas Gerais - Dor de Garganta.

No dendrograma gerado pelo DTW para o ano de 2021 (Figura 22), teve a formação de dois grupos principais. Interessante notar que um desses grupos corresponde exatamente ao grupo 2 identificado na análise da distância Euclidiana, reforçando a existência de uma relação significativa do sintoma de tosse entre os estados de Bahia, Minas Gerais e São Paulo. Quanto ao grupo maior, apesar de sua extensão, não se identificaram padrões claros que permitissem inferências significativas. Entretanto, dois subgrupos menores, localizados à esquerda, chamam atenção: estes incluem os sintomas de dispneia, febre e tosse, especificamente nos estados do Pará (registros 15, 18 e 19) e Rio de Janeiro (registros 25, 28 e 29).

Para 2021, as relações entre os sintomas se mantiveram basicamente iguais se tratando de grupos, mas com menos relações diretas entre febre, dispneia e tosse, e um resultado bem expressivo para representar a relação direta entre dor de garganta e dor de cabeça, conforme pode ser visto nas matrizes de contagem nas Figuras 23 e 24.

Figura 21 – Dendrograma dos dados de estado + sintoma 2021 com medida de distância temporal sendo distância Euclidiana.

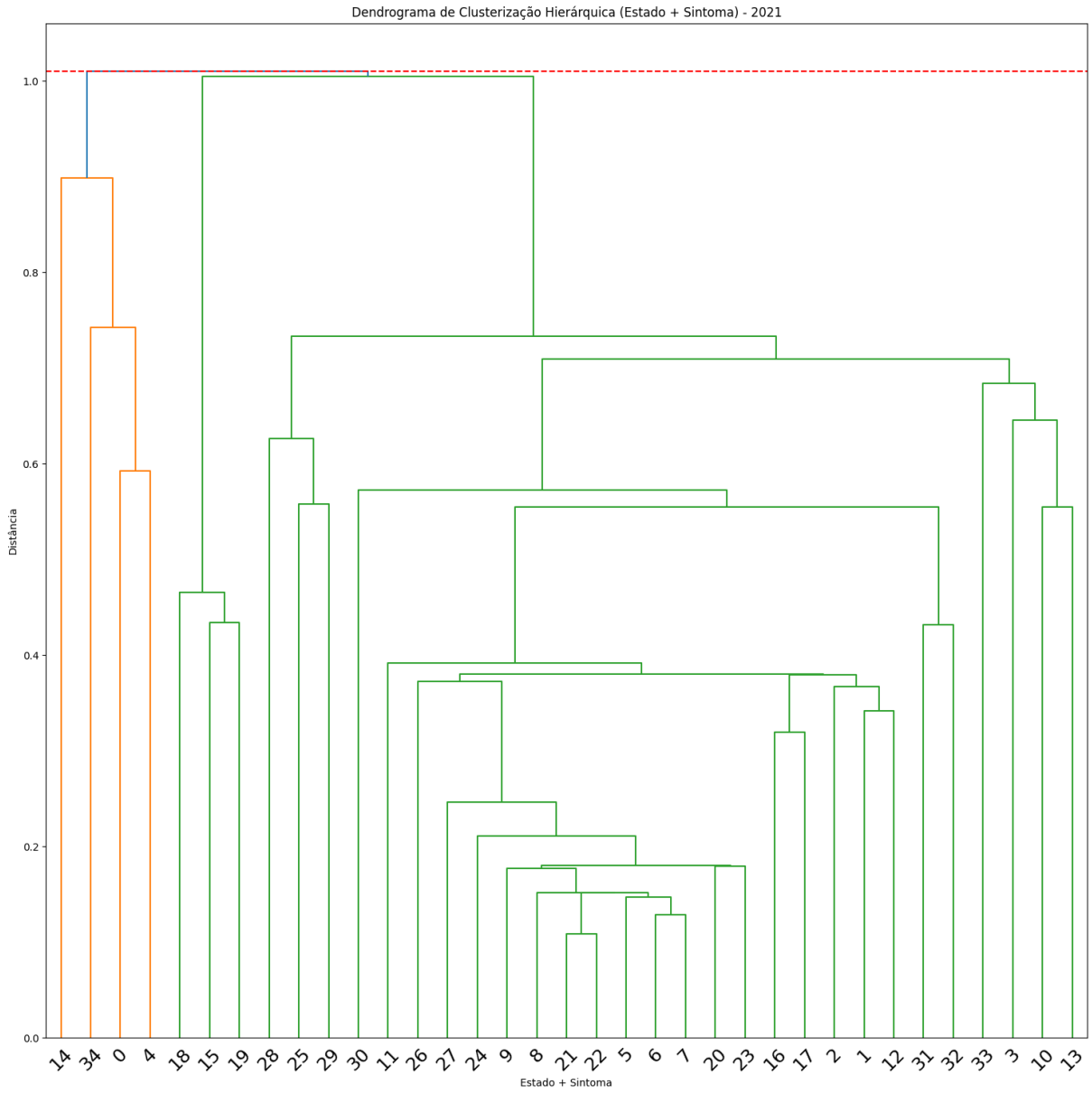


Fonte: Próprio autor.

Cluster	Registros
1	**1:** Bahia - Dor de Cabeça, **2:** Bahia - Dor de Garganta, **5:** Goiás - Dispneia, **6:** Goiás - Dor de Cabeça, **7:** Goiás - Dor de Garganta, **8:** Goiás - Febre, **9:** Goiás - Tosse, **11:** Minas Gerais - Dor de Cabeça, **12:** Minas Gerais - Dor de Garganta, **16:** Pará - Dor de Cabeça, **17:** Pará - Dor de Garganta, **20:** Rio Grande do Sul - Dispneia, **21:** Rio Grande do Sul - Dor de Cabeça, **22:** Rio Grande do Sul - Dor de Garganta, **23:** Rio Grande do Sul - Febre, **24:** Rio Grande do Sul - Tosse, **26:** Rio de Janeiro - Dor de Cabeça, **27:** Rio de Janeiro - Dor de Garganta
2	**0:** Bahia - Dispneia, **4:** Bahia - Tosse, **14:** Minas Gerais - Tosse, **34:** São Paulo - Tosse
3	**15:** Pará - Dispneia, **18:** Pará - Febre, **19:** Pará - Tosse
4	**3:** Bahia - Febre, **10:** Minas Gerais - Dispneia, **13:** Minas Gerais - Febre, **25:** Rio de Janeiro - Dispneia, **28:** Rio de Janeiro - Febre, **29:** Rio de Janeiro - Tosse, **30:** São Paulo - Dispneia, **31:** São Paulo - Dor de Cabeça, **32:** São Paulo - Dor de Garganta, **33:** São Paulo - Febre

Quadro 12 – Quadro com a separação dos registros da Figura 21.

Figura 22 – Dendrograma dos dados de estado + sintoma 2021 com medida de distância temporal sendo DTW.



Fonte: Próprio autor.

Cluster	Registros
1	**0:** Bahia - Dispneia, **4:** Bahia - Tosse, **14:** Minas Gerais - Tosse, **34:** São Paulo - Tosse
2	**1:** Bahia - Dor de Cabeça, **2:** Bahia - Dor de Garganta, **3:** Bahia - Febre, **5:** Goiás - Dispneia, **6:** Goiás - Dor de Cabeça, **7:** Goiás - Dor de Garganta, **8:** Goiás - Febre, **9:** Goiás - Tosse, **10:** Minas Gerais - Dispneia, **11:** Minas Gerais - Dor de Cabeça, **12:** Minas Gerais - Dor de Garganta, **13:** Minas Gerais - Febre, **15:** Pará - Dispneia, **16:** Pará - Dor de Cabeça, **17:** Pará - Dor de Garganta, **18:** Pará - Febre, **19:** Pará - Tosse, **20:** Rio Grande do Sul - Dispneia, **21:** Rio Grande do Sul - Dor de Cabeça, **22:** Rio Grande do Sul - Dor de Garganta, **23:** Rio Grande do Sul - Febre, **24:** Rio Grande do Sul - Tosse, **25:** Rio de Janeiro - Dispneia, **26:** Rio de Janeiro - Dor de Cabeça, **27:** Rio de Janeiro - Dor de Garganta, **28:** Rio de Janeiro - Febre, **29:** Rio de Janeiro - Tosse, **30:** São Paulo - Dispneia, **31:** São Paulo - Dor de Cabeça, **32:** São Paulo - Dor de Garganta, **33:** São Paulo - Febre

Quadro 13 – Quadro com a separação dos registros da Figura 22.

Figura 23 – Matriz de contagem dos relacionamentos de primeira ordem entre sintomas para distância Euclidiana em 2021.

	Tosse	Febre	Dispneia	Dor de cabeça	Dor de garganta
Tosse	1				
Febre	1	0			
Dispneia	4	1	0		
Dor de cabeça	0	1	0	0	
Dor de garganta	0	0	0	6	0

Fonte: Próprio autor.

Figura 24 – Matriz de contagem dos relacionamentos de primeira ordem entre sintomas para DTW em 2021.

	Tosse	Febre	Dispneia	Dor de cabeça	Dor de garganta
Tosse	0				
Febre	0	0			
Dispneia	3	2	0		
Dor de cabeça	0	0	0	0	
Dor de garganta	0	0	0	5	0

Fonte: Próprio autor.

4.2.3 Agrupamento estado + condição

Na análise de agrupamento que integra os estados brasileiros às condições pré-existentes dos pacientes, os dendrogramas revelam uma segmentação mais nítida e equitativa. Observa-se a formação de múltiplas relações de primeira ordem, sugerindo uma distinção clara entre os pares de registros.

Semelhante à abordagem adotada para os sintomas, a presente análise centrou-se nas cinco principais condições prevalentes em todos os estados para cada ano. Quatro condições se mantiveram comuns entre os dois anos, porém, imunossupressão (2020) foi substituída por obesidade (2021). Selecionadas essas condições, procedeu-se ao agrupamento hierárquico, correlacionando-as com os estados correspondentes.

Os dendrogramas com os quadros dos respectivos valores estão separados em duas subseções para cada um dos anos ser detalhados. O conjunto de estado + condição é referido como registro, e grupo refere-se aos grandes grupos montados nos dendrogramas.

4.2.3.1 Ano 2020

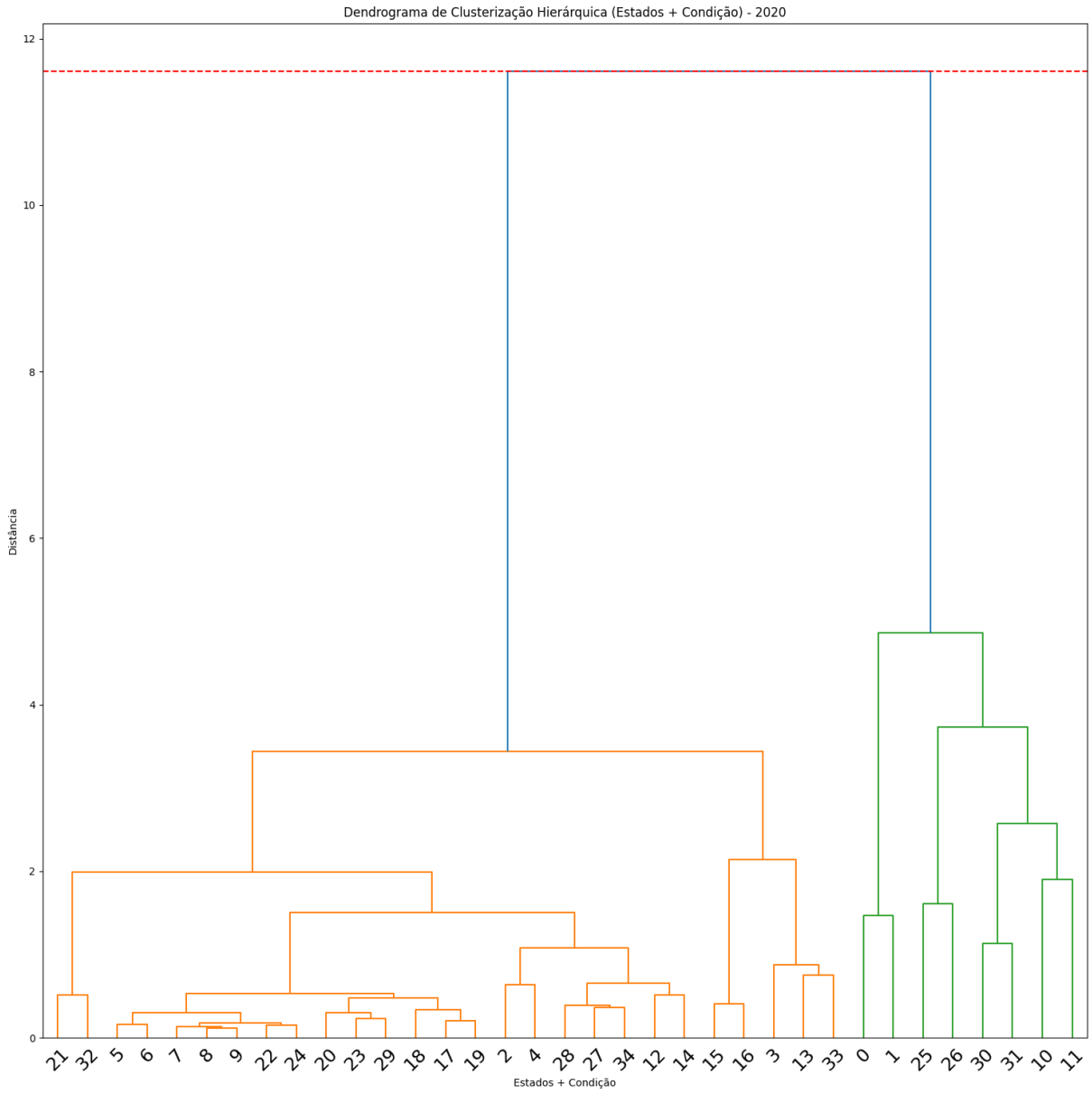
No que se refere à aplicação da medida de distância Euclidiana para o ano de 2020, conforme ilustrado na Figura 25, identificaram-se dois agrupamentos. O grupo 2 consiste em estados como Bahia, Minas Gerais, Rio de Janeiro e São Paulo, sendo que a associação se deu exclusivamente em relação às condições de diabetes e doenças cardíacas crônicas. Observa-se que dentro desse conjunto, as condições clínicas se alinham em uma ordem direta por estados. Estes estados agrupados novamente sugere uma potencial influência do volume significativo de notificações nesses estados sobre a configuração do agrupamento. Tal hipótese é reforçada pela análise dos registros 5 e 6 ou 15 e 16 do grupo 1, onde se evidencia a proximidade entre diabetes e doenças cardíacas crônicas, evidenciada pela Figura 27 que representa as contagens de relações diretas entre condições.

Além disso, no grupo 1, as demais condições apresentam uma tendência de agrupamento entre elas em cada estado, mais do que entre diferentes estados.

A análise do DTW, representada na Figura 26, revelou uma configuração distributiva semelhante à da distância Euclidiana, com dois grupos distintos que mantêm os mesmos estados e condições pré-existentes. O grupo 2 é formado por estados como Bahia, Minas Gerais, Rio de Janeiro e São Paulo, associados especificamente às condições de diabetes e doenças cardíacas crônicas. Contudo, neste modelo, observa-se uma dinâmica de inter-relação escalonada entre os estados, diferentemente da estrutura observada com a distância Euclidiana. Essas relações escalonadas acabam impactando no número de relações diretas, ou seja, acaba tendo menos, como pode-se ver na Figura 28. O grupo 1 apresenta algumas inter-relações de primeira ordem, bem como algumas proximidades notáveis entre estados e condições específicas, exemplificadas por:

- (21) Rio Grande do Sul - Doenças cardíacas crônicas e (32) São Paulo - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5);
- (8) Goiás - Doenças respiratórias crônicas descompensadas e (22) Rio Grande do Sul - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5);
- (6) Goiás - Doenças cardíacas crônicas (23) Rio Grande do Sul - Doenças respiratórias crônicas descompensadas.

Figura 25 – Dendrograma dos dados de estado + condição 2020 com medida de distância temporal sendo distância Euclidiana.

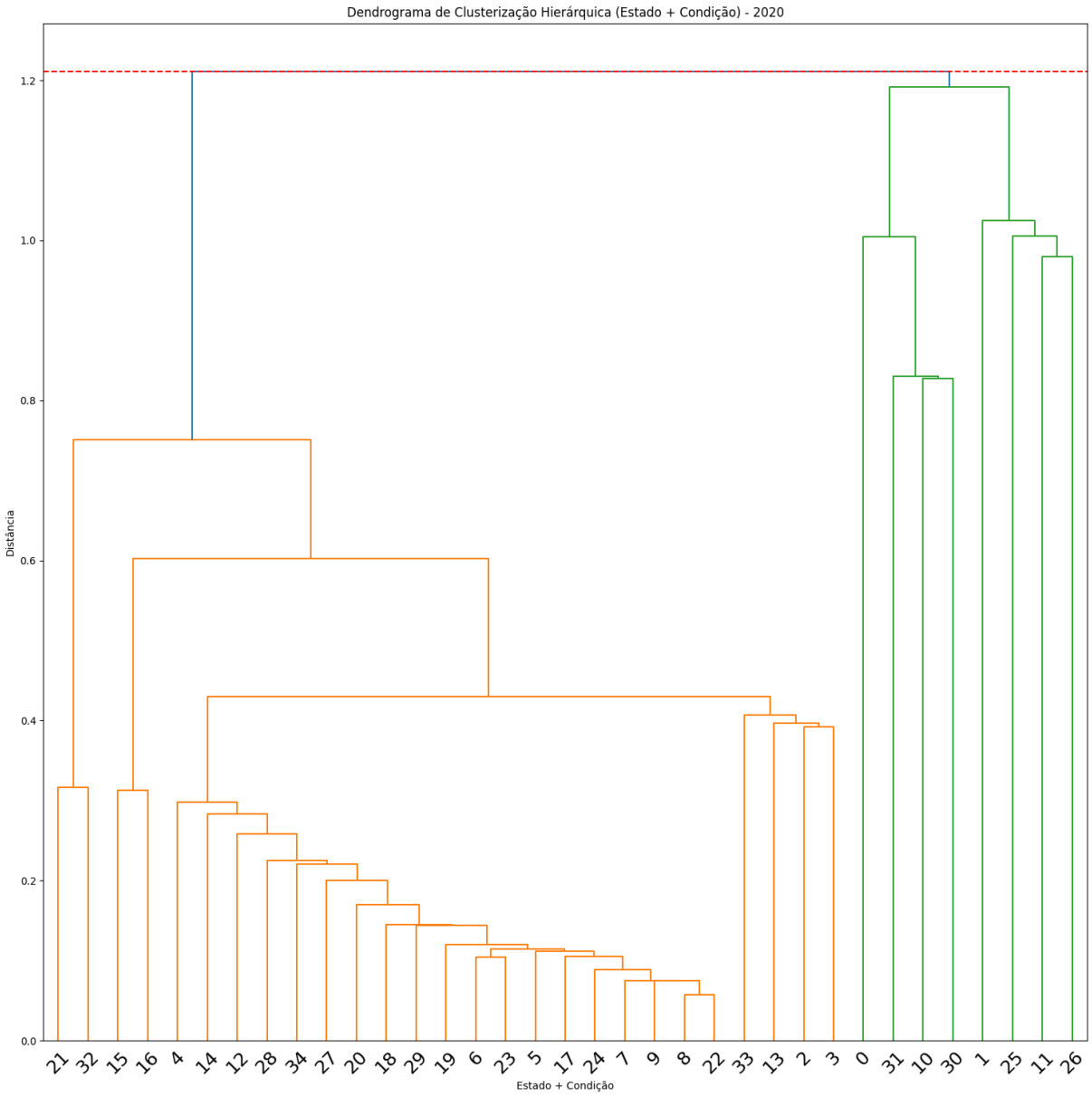


Fonte: Próprio autor.

Cluster	Registros
1	**2:** Bahia - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **3:** Bahia - Doenças respiratórias crônicas descompensadas, **4:** Bahia - Imunossupressão, **5:** Goiás - Diabetes, **6:** Goiás - Doenças cardíacas crônicas, **7:** Goiás - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **8:** Goiás - Doenças respiratórias crônicas descompensadas, **9:** Goiás - Imunossupressão, **12:** Minas Gerais - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **13:** Minas Gerais - Doenças respiratórias crônicas descompensadas, **14:** Minas Gerais - Imunossupressão, **15:** Pará - Diabetes, **16:** Pará - Doenças cardíacas crônicas, **17:** Pará - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **18:** Pará - Doenças respiratórias crônicas descompensadas, **19:** Pará - Imunossupressão, **20:** Rio Grande do Sul - Diabetes, **21:** Rio Grande do Sul - Doenças cardíacas crônicas, **22:** Rio Grande do Sul - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **23:** Rio Grande do Sul - Doenças respiratórias crônicas descompensadas, **24:** Rio Grande do Sul - Imunossupressão, **27:** Rio de Janeiro - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **28:** Rio de Janeiro - Doenças respiratórias crônicas descompensadas, **29:** Rio de Janeiro - Imunossupressão, **32:** São Paulo - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **33:** São Paulo - Doenças respiratórias crônicas descompensadas, **34:** São Paulo - Imunossupressão
2	**0:** Bahia - Diabetes, **1:** Bahia - Doenças cardíacas crônicas, **10:** Minas Gerais - Diabetes, **11:** Minas Gerais - Doenças cardíacas crônicas, **25:** Rio de Janeiro - Diabetes, **26:** Rio de Janeiro - Doenças cardíacas crônicas, **30:** São Paulo - Diabetes, **31:** São Paulo - Doenças cardíacas crônicas

Quadro 14 – Quadro com a separação dos registros da Figura 25.

Figura 26 – Dendrograma dos dados de estado + condição 2020 com medida de distância temporal sendo DTW.



Fonte: Próprio autor.

Cluster	Registros
1	**2:** Bahia - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **3:** Bahia - Doenças respiratórias crônicas descompensadas, **4:** Bahia - Imunossupressão, **5:** Goiás - Diabetes, **6:** Goiás - Doenças cardíacas crônicas, **7:** Goiás - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **8:** Goiás - Doenças respiratórias crônicas descompensadas, **9:** Goiás - Imunossupressão, **12:** Minas Gerais - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **13:** Minas Gerais - Doenças respiratórias crônicas descompensadas, **14:** Minas Gerais - Imunossupressão, **15:** Pará - Diabetes, **16:** Pará - Doenças cardíacas crônicas, **17:** Pará - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **18:** Pará - Doenças respiratórias crônicas descompensadas, **19:** Pará - Imunossupressão, **20:** Rio Grande do Sul - Diabetes, **21:** Rio Grande do Sul - Doenças cardíacas crônicas, **22:** Rio Grande do Sul - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **23:** Rio Grande do Sul - Doenças respiratórias crônicas descompensadas, **24:** Rio Grande do Sul - Imunossupressão, **27:** Rio de Janeiro - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **28:** Rio de Janeiro - Doenças respiratórias crônicas descompensadas, **29:** Rio de Janeiro - Imunossupressão, **32:** São Paulo - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **33:** São Paulo - Doenças respiratórias crônicas descompensadas, **34:** São Paulo - Imunossupressão
2	**0:** Bahia - Diabetes, **1:** Bahia - Doenças cardíacas crônicas, **10:** Minas Gerais - Diabetes, **11:** Minas Gerais - Doenças cardíacas crônicas, **25:** Rio de Janeiro - Diabetes, **26:** Rio de Janeiro - Doenças cardíacas crônicas, **30:** São Paulo - Diabetes, **31:** São Paulo - Doenças cardíacas crônicas

Quadro 15 – Quadro com a separação dos registros da Figura 26.

Figura 27 – Matriz de contagem dos relacionamentos de primeira ordem entre condições pré-existentes para distância Euclidiana em 2020.

	Diabetes	Imunossupressão	Doenças card. crôn.	Doenças renais crôn. est. avan.	Doenças resp. crôn. desc.
Diabetes	0				
Imunossupressão	0	0			
Doenças card. crôn.	6	0	0		
Doenças renais crôn. est. avan.	0	5	1	0	
Doenças resp. crôn. desc.	0	2	0	0	1

Fonte: Próprio autor.

Figura 28 – Matriz de contagem dos relacionamentos de primeira ordem entre condições pré-existentes para DTW em 2020.

	Diabetes	Imunossupressão	Doenças card. crôn.	Doenças renais crôn. est. avan.	Doenças resp. crôn. desc.
Diabetes	1				
Imunossupressão	0	0			
Doenças card. crôn.	1	0	1		
Doenças renais crôn. est. avan.	0	0	1	0	
Doenças resp. crôn. desc.	0	0	1	2	0

Fonte: Próprio autor.

4.2.3.2 Ano 2021

A primeira observação a ser feita para 2021 é o fato de uma nova condição ter sido uma das cinco mais comuns no lugar da imunossupressão, que é a obesidade. Ao analisar utilizando a distância Euclidiana, conforme ilustrado na Figura 29, identificou-se a formação de dois grupos principais. Notavelmente, o grupo 2 continuou a evidenciar a presença das condições de diabetes e doenças cardíacas crônicas (Figura 31), com os mesmos estados do ano anterior e a inclusão do Pará. Uma revisão dos dados da Tabela 3 indica um incremento no número de óbitos para o Pará em comparação ao ano de 2020, como mostrado na Tabela 3. Contudo, a proporção de notificações de óbitos entre o Pará e os outros estados alterou-se em 2021, sugerindo que uma parcela significativa dos óbitos nesse estado pode estar associada às referidas condições pré-existentes. Por outro lado, o grupo 1 apresentou uma diversidade maior, incluindo as mesmas condições para Goiás (registros 5 e 6), além de várias inter-relações de primeira ordem entre estados e condições específicas.

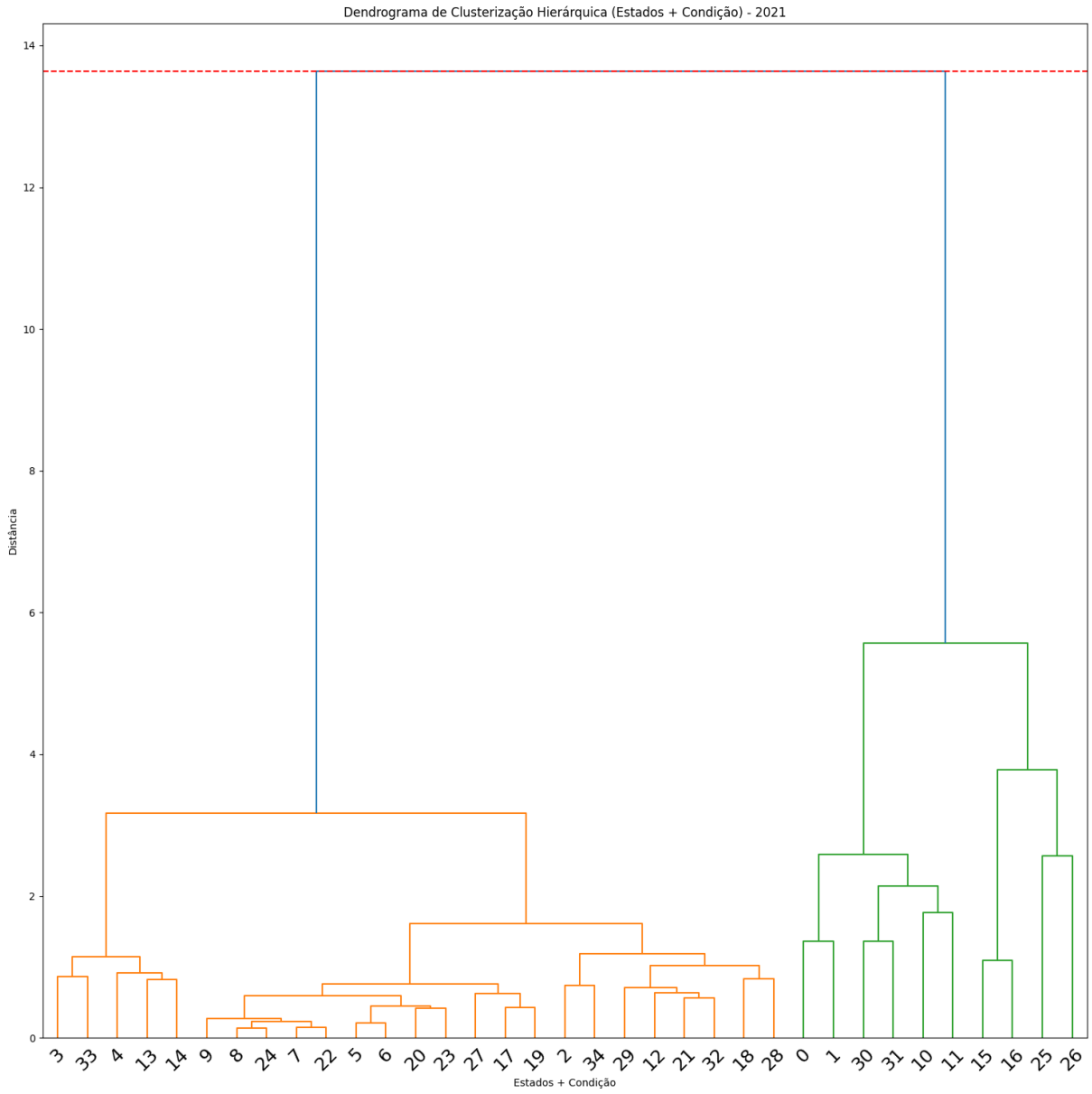
- (3) Bahia - Doenças respiratórias crônicas descompensadas e (33) São Paulo - Doenças respiratórias crônicas descompensadas;
- (8) Goiás - Doenças cardíacas crônicas e (24) Rio Grande do Sul - Obesidade;
- (7) Goiás - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5) e (22) Rio Grande do Sul - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5);
- (21) Rio Grande do Sul - Doenças cardíacas crônicas e (32) São Paulo - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5).

A análise de DTW para o ano de 2021, representada na Figura 30, revelou uma configuração distinta, com a formação de seis grupos. Desconsidera-se os dois outliers que constituem grupos isolados (Minas Gerais e Rio de Janeiro, ambos associados a doenças cardíacas crônicas), efetivamente observam-se quatro grupos distintos. O grupo

1 é singularmente composto pelo estado do Pará, enfatizando as condições de diabetes e doenças cardíacas crônicas. O grupo 2 delineia uma associação entre dois diferentes estados e condições de saúde. No grupo 3, identifica-se uma conexão entre dois estados com a condição de diabetes. O grupo 4, o mais numeroso, apresenta uma estrutura escalonada com exceção de dois subgrupos mais definidos. Uma análise detalhada desse grupo maior indica que várias das relações de primeira ordem compartilham uma condição de saúde comum: a obesidade, isso pode ser observado na Figura 32.

- (4) Bahia - Obesidade e (14) Minas Gerais - Obesidade;
- (8) Goiás - Doenças respiratórias crônicas descompensadas e (24) Rio Grande do Sul - Obesidade;
- (9) Goiás - Obesidade e (22) Rio Grande do Sul - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5).

Figura 29 – Dendrograma dos dados de estado + condição 2021 com medida de distância temporal sendo distância Euclidiana.

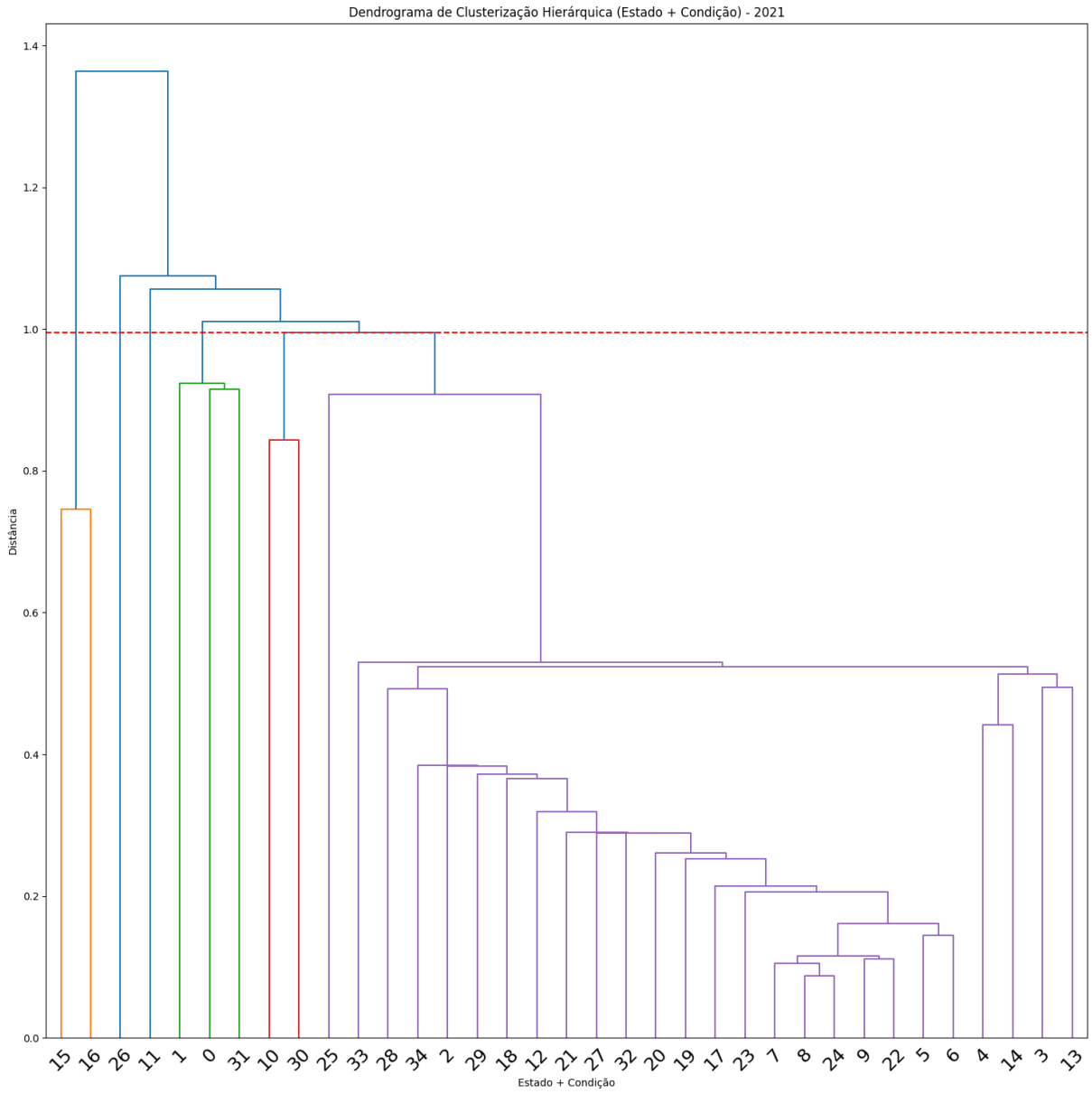


Fonte: Próprio autor.

Cluster	Registros
1	**2:** Bahia - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **3:** Bahia - Doenças respiratórias crônicas descompensadas, **4:** Bahia - Obesidade, **5:** Goiás - Diabetes, **6:** Goiás - Doenças cardíacas crônicas, **7:** Goiás - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **8:** Goiás - Doenças respiratórias crônicas descompensadas, **9:** Goiás - Obesidade, **12:** Minas Gerais - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **13:** Minas Gerais - Doenças respiratórias crônicas descompensadas, **14:** Minas Gerais - Obesidade, **17:** Pará - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **18:** Pará - Doenças respiratórias crônicas descompensadas, **19:** Pará - Obesidade, **20:** Rio Grande do Sul - Diabetes, **21:** Rio Grande do Sul - Doenças cardíacas crônicas, **22:** Rio Grande do Sul - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **23:** Rio Grande do Sul - Doenças respiratórias crônicas descompensadas, **24:** Rio Grande do Sul - Obesidade, **27:** Rio de Janeiro - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **28:** Rio de Janeiro - Doenças respiratórias crônicas descompensadas, **29:** Rio de Janeiro - Obesidade, **32:** São Paulo - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **33:** São Paulo - Doenças respiratórias crônicas descompensadas, **34:** São Paulo - Obesidade
2	**0:** Bahia - Diabetes, **1:** Bahia - Doenças cardíacas crônicas, **10:** Minas Gerais - Diabetes, **11:** Minas Gerais - Doenças cardíacas crônicas, **15:** Pará - Diabetes, **16:** Pará - Doenças cardíacas crônicas, **25:** Rio de Janeiro - Diabetes, **26:** Rio de Janeiro - Doenças cardíacas crônicas, **30:** São Paulo - Diabetes, **31:** São Paulo - Doenças cardíacas crônicas

Quadro 16 – Quadro com a separação dos registros da Figura 29.

Figura 30 – Dendrograma dos dados de estado + condição 2021 com medida de distância temporal sendo DTW.



Fonte: Próprio autor.

Cluster	Registros
1	**15:** Pará - Diabetes, **16:** Pará - Doenças cardíacas crônicas
2	**0:** Bahia - Diabetes, **1:** Bahia - Doenças cardíacas crônicas, **31:** São Paulo - Doenças cardíacas crônicas
3	**10:** Minas Gerais - Diabetes, **30:** São Paulo - Diabetes
4	**2:** Bahia - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **3:** Bahia - Doenças respiratórias crônicas descompensadas, **4:** Bahia - Obesidade, **5:** Goiás - Diabetes, **6:** Goiás - Doenças cardíacas crônicas, **7:** Goiás - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **8:** Goiás - Doenças respiratórias crônicas descompensadas, **9:** Goiás - Obesidade, **12:** Minas Gerais - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **13:** Minas Gerais - Doenças respiratórias crônicas descompensadas, **14:** Minas Gerais - Obesidade, **17:** Pará - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **18:** Pará - Doenças respiratórias crônicas descompensadas, **19:** Pará - Obesidade, **20:** Rio Grande do Sul - Diabetes, **21:** Rio Grande do Sul - Doenças cardíacas crônicas, **22:** Rio Grande do Sul - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **23:** Rio Grande do Sul - Doenças respiratórias crônicas descompensadas, **24:** Rio Grande do Sul - Obesidade, **25:** Rio de Janeiro - Diabetes, **27:** Rio de Janeiro - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **28:** Rio de Janeiro - Doenças respiratórias crônicas descompensadas, **29:** Rio de Janeiro - Obesidade, **32:** São Paulo - Doenças renais crônicas em estágio avançado (graus 3, 4 e 5), **33:** São Paulo - Doenças respiratórias crônicas descompensadas, **34:** São Paulo - Obesidade
5	**11:** Minas Gerais - Doenças cardíacas crônicas
6	**26:** Rio de Janeiro - Doenças cardíacas crônicas

Quadro 17 – Quadro com a separação dos registros da Figura 30.

Figura 31 – Matriz de contagem dos relacionamentos de primeira ordem entre condições pré-existentes para distância Euclidiana em 2021.

	Diabetes	Obesidade	Doenças card. crôn.	Doenças renais crôn. est. avan.	Doenças resp. crôn. desc.
Diabetes	0				
Obesidade	0	0			
Doenças card. crôn.	6	0	0		
Doenças renais crôn. est. avan.	0	2	1	1	
Doenças resp. crôn. desc.	1	2	0	0	2

Fonte: Próprio autor.

Figura 32 – Matriz de contagem dos relacionamentos de primeira ordem entre condições pré-existentes para DTW em 2021.

	Diabetes	Obesidade	Doenças card. crôn.	Doenças renais crôn. est. avan.	Doenças resp. crôn. desc.
Diabetes	1				
Obesidade	0	1			
Doenças card. crôn.	3	0	0		
Doenças renais crôn. est. avan.	0	1	0	0	
Doenças resp. crôn. desc.	0	1	0	0	1

Fonte: Próprio autor.

4.2.4 Agrupamento estado + sexo

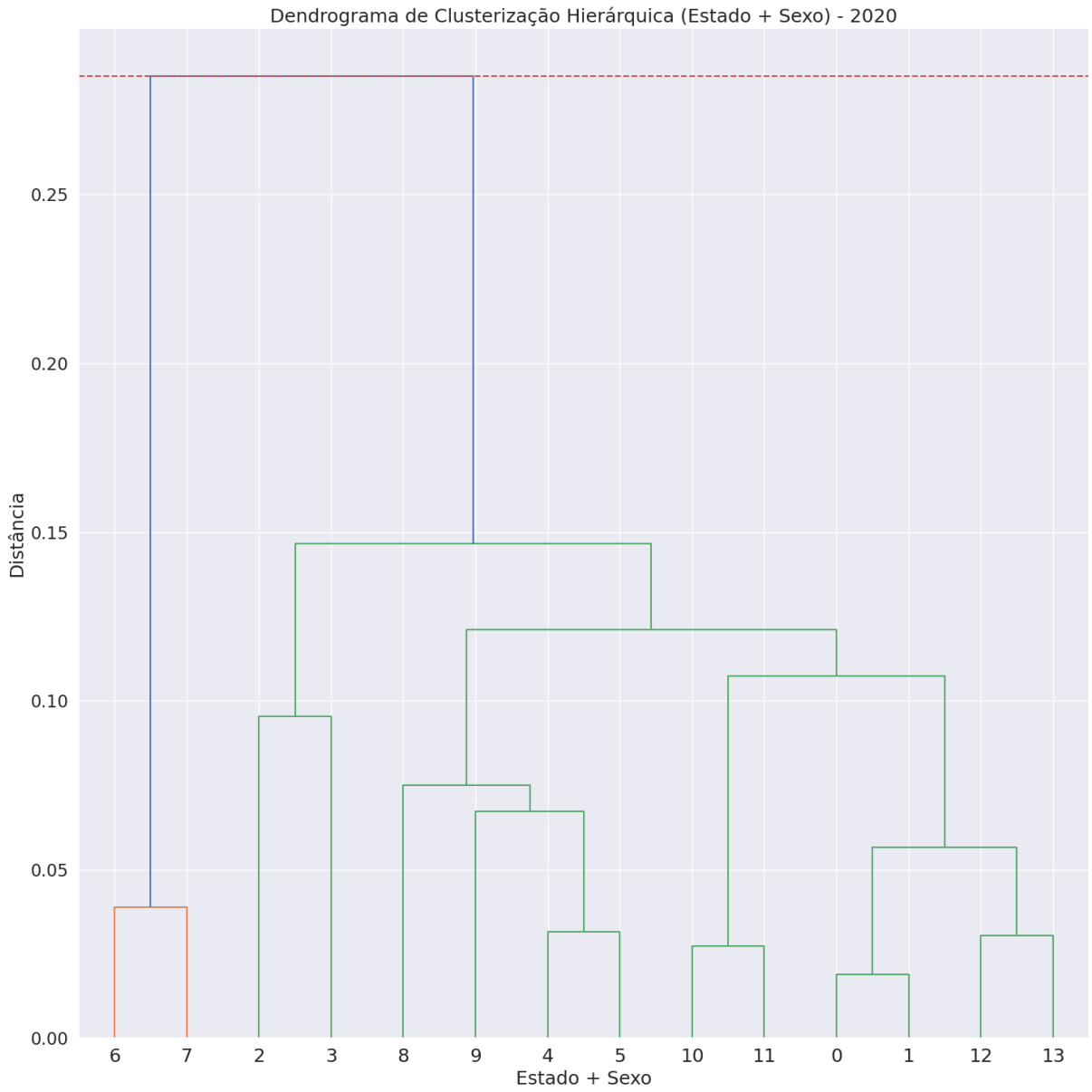
Os agrupamentos hierárquicos combinando os estados brasileiros e os dados de sexo não foram analisados anualmente, diferentemente dos anteriores. Esta análise tem por objetivo funcionar mais como um validador dos dados coletados, visto que a variável sexo não é esperada como um fator influente na evolução clínica de casos fatais.

Importante ressaltar que observou-se a presença de uma categoria de gênero 'indefinido' nas notificações do estado de São Paulo referentes ao ano de 2020. A quantidade substancial de registros sob essa designação exigiu a sua exclusão para assegurar a precisão da análise. Assim, todas as notificações associadas ao gênero 'indefinido' foram removidas do conjunto de dados utilizados neste agrupamento.

Na análise inicial dos quatro dendrogramas, nota-se que a distância Euclidiana, para ambos os anos de 2020 e 2021 (Figuras 33 e 35), revelou resultados mais alinhados com as expectativas em comparação ao DTW, principalmente pela forma como os sexos foram associados aos estados em relações de primeira ordem. Observa-se também que, em três dos dendrogramas (Figuras 33, 34 e 36), foram formados dois grandes grupos. Contudo, no caso da distância Euclidiana de 2021 (Figura 35), identificaram-se cinco grupos distintos, sugerindo uma separação ainda mais acentuada entre os estados. Este resultado corrobora a intenção inicial deste agrupamento de estados com sexos. Neste

mesmo dendrograma da Figura 35, três estados se mostraram particularmente próximos dentro do grupo 3: Bahia, Minas Gerais e São Paulo.

Figura 33 – Dendrograma dos dados de estado + sexo 2020 com medida de distância temporal sendo distância Euclidiana.

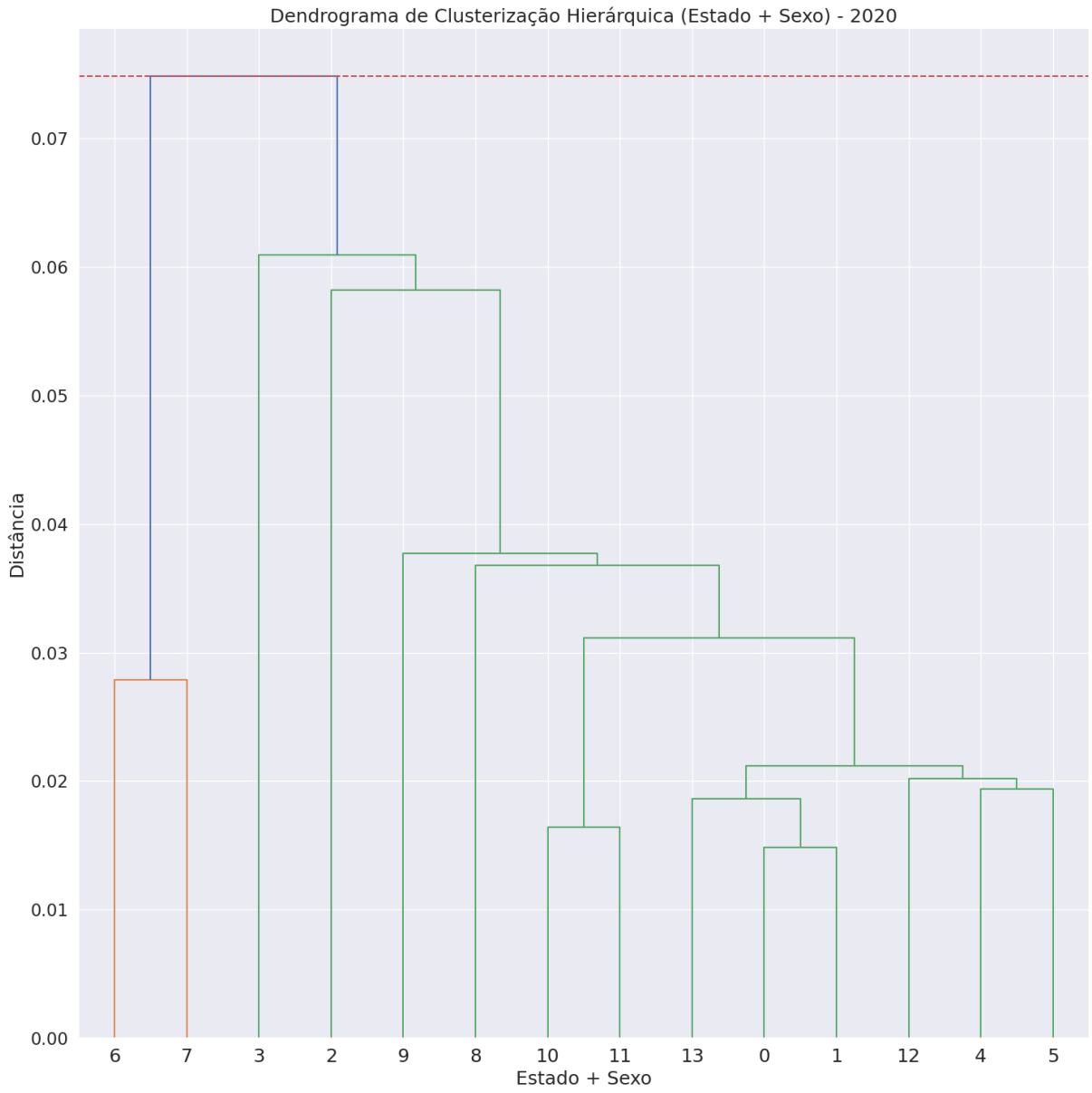


Fonte: Próprio autor.

Cluster	Registros
1	**6:** (’Pará’, ’Feminino’), **7:** (’Pará’, ’Masculino’)
2	**0:** (’Bahia’, ’Feminino’), **1:** (’Bahia’, ’Masculino’), **2:** (’Goiás’, ’Feminino’), **3:** (’Goiás’, ’Masculino’), **4:** (’Minas Gerais’, ’Feminino’), **5:** (’Minas Gerais’, ’Masculino’), **8:** (’Rio Grande do Sul’, ’Feminino’), **9:** (’Rio Grande do Sul’, ’Masculino’), **10:** (’Rio de Janeiro’, ’Feminino’), **11:** (’Rio de Janeiro’, ’Masculino’), **12:** (’São Paulo’, ’Feminino’), **13:** (’São Paulo’, ’Masculino’)

Quadro 18 – Quadro com a separação dos registros da Figura 33.

Figura 34 – Dendrograma dos dados de estado + sexo 2020 com medida de distância temporal sendo DTW.

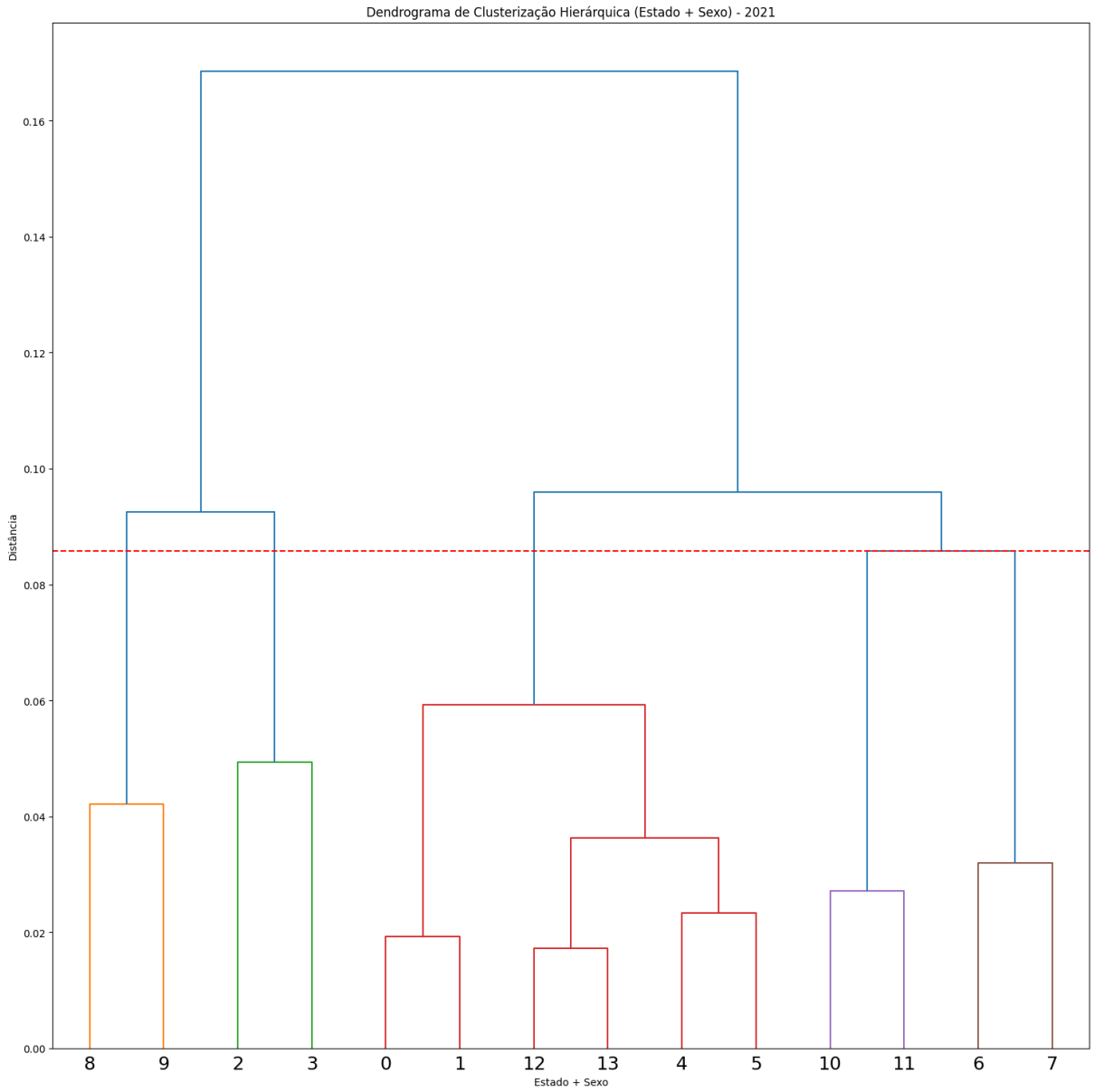


Fonte: Próprio autor.

Cluster	Registros
1	**6:** (’Pará’, ’Feminino’), **7:** (’Pará’, ’Masculino’)
2	**0:** (’Bahia’, ’Feminino’), **1:** (’Bahia’, ’Masculino’), **2:** (’Goiás’, ’Feminino’), **3:** (’Goiás’, ’Masculino’), **4:** (’Minas Gerais’, ’Feminino’), **5:** (’Minas Gerais’, ’Masculino’), **8:** (’Rio Grande do Sul’, ’Feminino’), **9:** (’Rio Grande do Sul’, ’Masculino’), **10:** (’Rio de Janeiro’, ’Feminino’), **11:** (’Rio de Janeiro’, ’Masculino’), **12:** (’São Paulo’, ’Feminino’), **13:** (’São Paulo’, ’Masculino’)

Quadro 19 – Quadro com a separação dos registros da Figura 34.

Figura 35 – Dendrograma dos dados de estado + sexo 2021 com medida de distância temporal sendo distância Euclidiana.

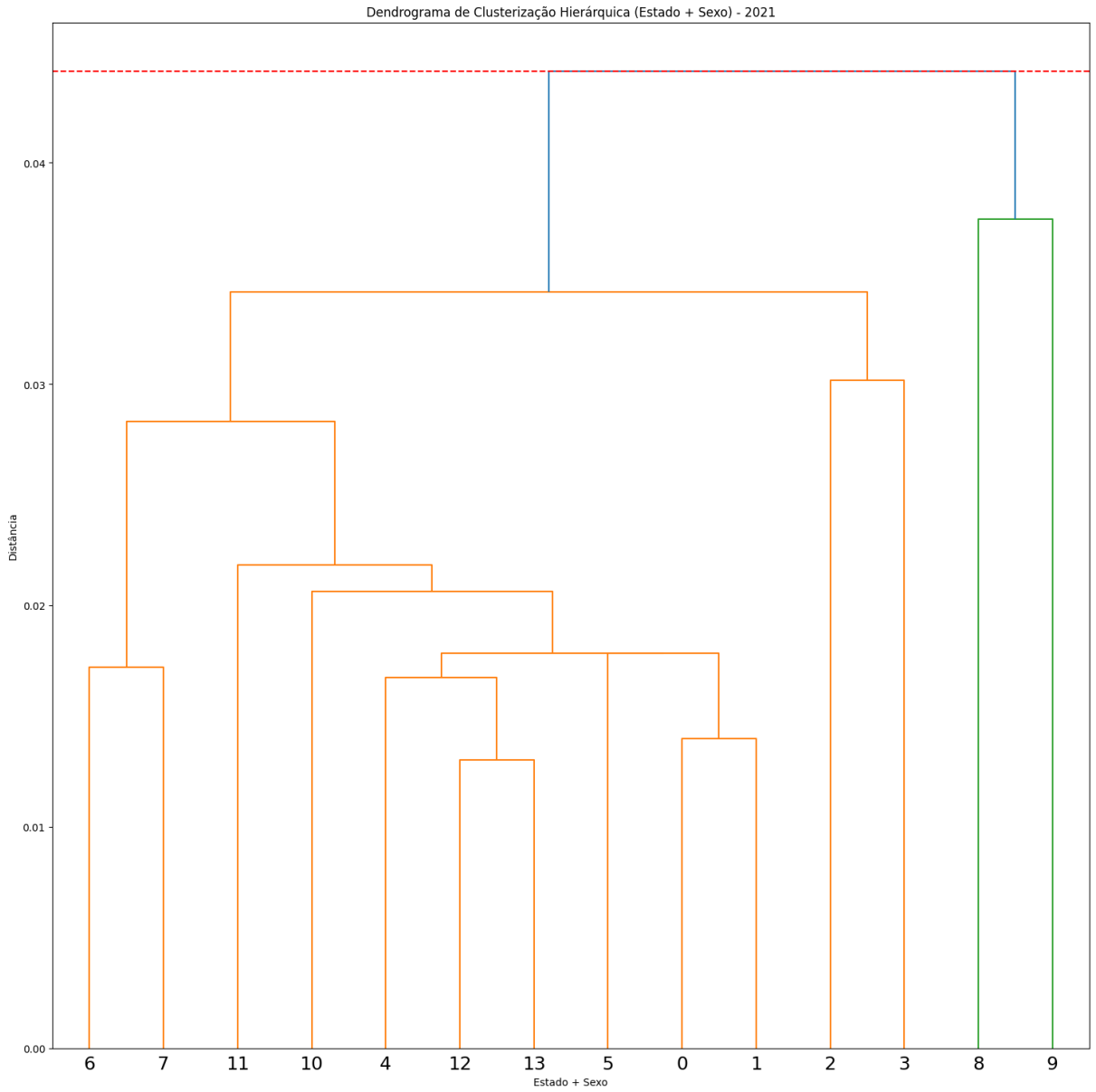


Fonte: Próprio autor.

Cluster	Registros
1	**8:** Rio Grande do Sul - Feminino, **9:** Rio Grande do Sul - Masculino
2	**2:** Goiás - Feminino, **3:** Goiás - Masculino
3	**0:** Bahia - Feminino, **1:** Bahia - Masculino, **4:** Minas Gerais - Feminino, **5:** Minas Gerais - Masculino, **12:** São Paulo - Feminino, **13:** São Paulo - Masculino
4	**10:** Rio de Janeiro - Feminino, **11:** Rio de Janeiro - Masculino
5	**6:** Pará - Feminino, **7:** Pará - Masculino

Quadro 20 – Quadro com a separação dos registros da Figura 35.

Figura 36 – Dendrograma dos dados de estado + sexo 2021 com medida de distância temporal sendo DTW.



Fonte: Próprio autor.

Cluster	Registros
1	**0:** Bahia - Feminino, **1:** Bahia - Masculino, **2:** Goiás - Feminino, **3:** Goiás - Masculino, **4:** Minas Gerais - Feminino, **5:** Minas Gerais - Masculino, **6:** Pará - Feminino, **7:** Pará - Masculino, **10:** Rio de Janeiro - Feminino, **11:** Rio de Janeiro - Masculino, **12:** São Paulo - Feminino, **13:** São Paulo - Masculino
2	**8:** Rio Grande do Sul - Feminino, **9:** Rio Grande do Sul - Masculino

Quadro 21 – Quadro com a separação dos registros da Figura 36.

5 CONCLUSÃO

O presente trabalho teve como objetivo investigar padrões em algumas características específicas entre os estados brasileiros, com a intenção de descobrir possíveis similaridades entre os estados. Ao analisar os resultados, observa-se que o ano de 2021 revelou relações mais interessantes em todas características, como as faixas etárias estarem inter-relacionadas entre estados, uma similaridade entre três sintomas (tosse, dispneia e febre) em diferentes estados, ou também duas condições pré-existentes (diabetes e doenças cardíacas crônicas) estarem agrupadas entre os estados. Mesmo assim, um aspecto influenciou negativamente na formação dos clusters para ambos os anos, principalmente o ano de 2020, que foi a discrepância no número de notificações de óbitos entre os estados. No que tange à característica de faixa etária, por exemplo, o ano de 2020 apresentou dificuldades em definir padrões claros, em grande parte devido à ausência de dados de idade em muitos registros de óbitos. Por outro lado, as características de sintomas e condições mostraram maior consistência nos resultados entre 2020 e 2021, com o ano de 2021 contribuindo com observações adicionais.

Certamente que, com acesso a dados mais completos e representativos da realidade, as inferências deste estudo poderiam ser substancialmente verificadas ou até desfeitas, levando à identificação de novos padrões. Porém, este trabalho reflete a realidade dos registros (dados) até o presente momento, visto que os dados foram massivamente coletados de uma fonte oficial, o OpenDataSUS[1].

Importa destacar que, mesmo considerando a incompletude dos dados utilizados neste estudo, tanto em termos quantitativos quanto qualitativos, as análises realizadas abrem a possibilidade de comparações futuras com informações oficiais providas pelas secretarias de saúde estaduais. Esta comparação entre os dados pode revelar discrepâncias notáveis. Ademais, empregar abordagens analíticas alternativas, como a análise de média móvel, constitui uma via complementar de investigação. Esta técnica se mostra particularmente valiosa em situações onde os dados exibem distribuição temporal inconsistente ou estão fortemente concentrados em certos períodos, como observou-se para certos casos de faixas etárias.

A ciência de dados, um campo interdisciplinar, exige um conhecimento profundo ou a colaboração de especialistas para validar hipóteses, especialmente ao transitar por áreas específicas como a saúde. Este estudo, embora não se aprofunde em análises clínicas detalhadas, focou-se na extração, unificação dos dados disponíveis para representação de similaridades e demonstração dos padrões entre estados através da clusterização hierárquica. Apesar desta abordagem limitada, o trabalho demonstrou ser útil para a avaliação de um conjunto de dados promissor, cuja finalidade é centralizar e padronizar as notifica-

[1] https://opendatasus.saude.gov.br/dataset?tags=covid

ções de saúde, mas que ainda está em fase de atualização e desenvolvimento.

REFERÊNCIAS

Aggarwal, Charu C and Reddy, Chandan K. **Data Clustering: Algorithms and Applications**. [S.l.]: Chapman and Hall/CRC, 2013.

ALBERTO-OLIVARES, M. et al. Remaining useful life prediction for turbofan based on a multilayer perceptron and kalman filter*. In: . [S.l.: s.n.], 2019. p. 1–6.

BERTHOLD, M. R.; HöPPNER, F. **On Clustering Time Series Using Euclidean Distance and Pearson Correlation**. 2016.

CASSãO, V. et al. Unsupervised analysis of covid-19 pandemic evolution in brazilian states. **Procedia Computer Science**, v. 196, p. 655–662, 2022.

Evandro Furoni and Giulia Alecrim and André Luiz Rosada. **Brasil ultrapassa a marca de 600 mil mortes pela Covid-19, segundo dados da CNN**: De acordo com o levantamento da agência cnn, o brasil já soma 600.067 vítimas na pandemia. CNN, 2021. Acesso em 27 set. 2023. Disponível em: <https://www.cnnbrasil.com.br/saude/brasil-ultrapassa-a-marca-de-600-mil-mortes-pela-covid-19-segundo-dados-da-cnn>.

HEIDARI, E.; SOBATI, M.; MOVAHEDIRAD, S. Accurate prediction of nanofluid viscosity using a multilayer perceptron artificial neural network (mlp-ann). **Chemometrics and Intelligent Laboratory Systems**, v. 155, 07 2016.

Heloisa Cristaldo and Marcelo Brandão. **Vacinação contra a covid-19 começa em todo o país**: Ministério da saúde distribuiu 6 milhões de doses do imunizante. agencia Brasil, 2021. Acesso em 04 out. 2023. Disponível em: <https://agenciabrasil.ebc.com.br/saude-/noticia/2021-01/vacinacao-contra-covid-19-come%C3%A7a-em-todo-o-pais>.

Jain, A. K. and Murty, M. N. and Flynn, P. J. Data clustering: A review. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 31, n. 3, p. 264323, sep 1999. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/331499.331504>.

JAMES, N.; MENZIES, M.; BONDELL, H. Comparing the dynamics of covid-19 infection and mortality in the united states, india, and brazil. **Physica D: Nonlinear Phenomena**, v. 432, p. 133–158, 2022.

JOHNSON, S. C. Hierarchical clustering schemes. **Psychometrika**, v. 32, p. 241–254, 1967.

Jonas Valente. **Covid-19: Brasil bate recorde com 4.249 mortes registradas em 24 horas**: Número de pessoas recuperadas subiu para 11.732.193. agencia Brasil, 2021. Acesso em 27 set. 2023. Disponível em: <https://agenciabrasil.ebc.com.br/saude/noticia-/2021-04/covid-19-brasil-bate-recorde-com-4249-mortes-registradas-em-24-horas>.

KEOGH, E.; RATANAMAHATANA, C. A. Exact indexing of dynamic time warping. **Knowl Inf Syst 7**, p. 358–386, 2005.

MCKINNEY, W. Data structures for statistical computing in python. In: WALT, S. van der; MILLMAN, J. (Ed.). **Proceedings of the 9th Python in Science Conference**. [S.l.: s.n.], 2010. p. 56 – 61.

MURTAGH, F.; LEGENDRE, P. Wards hierarchical agglomerative clustering method: Which algorithms implement wards criterion? **J Classif**, v. 31, p. 274–295, 2014.

MüLLER, M. **Information Retrieval for Music and Motion**. [S.l.]: Springer, Berlin, Heidelberg, 2007.

PEARSON, K. Notes on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London**, v. 58, p. 240–242, 1895.

ROJAS-VALENZUELA, I. et al. Estimation of covid-19 dynamics in the different states of the united states during the first months of the pandemic. **Engineering Proceedings**, v. 5, 2021.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii-/0377042787901257>.

SAARELA, M.; JAUHIAINEN, S. Comparison of feature importance measures as explanations for classification models. **SN Applied Sciences**, v. 3, n. 272, 2021.

SAKOE, H.; CHIBA, S. Dynamic programming algorithm optimization for spoken word recognition. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, v. 26, n. 1, p. 43–49, 1978.

XANTACROSS. **Diffference in matching between Euclidean and Dynamic Time Warping.** 2011. Disponível em: <https://commons.wikimedia.org/wiki/File:Euclidean_vs_DTW-.jpg>.

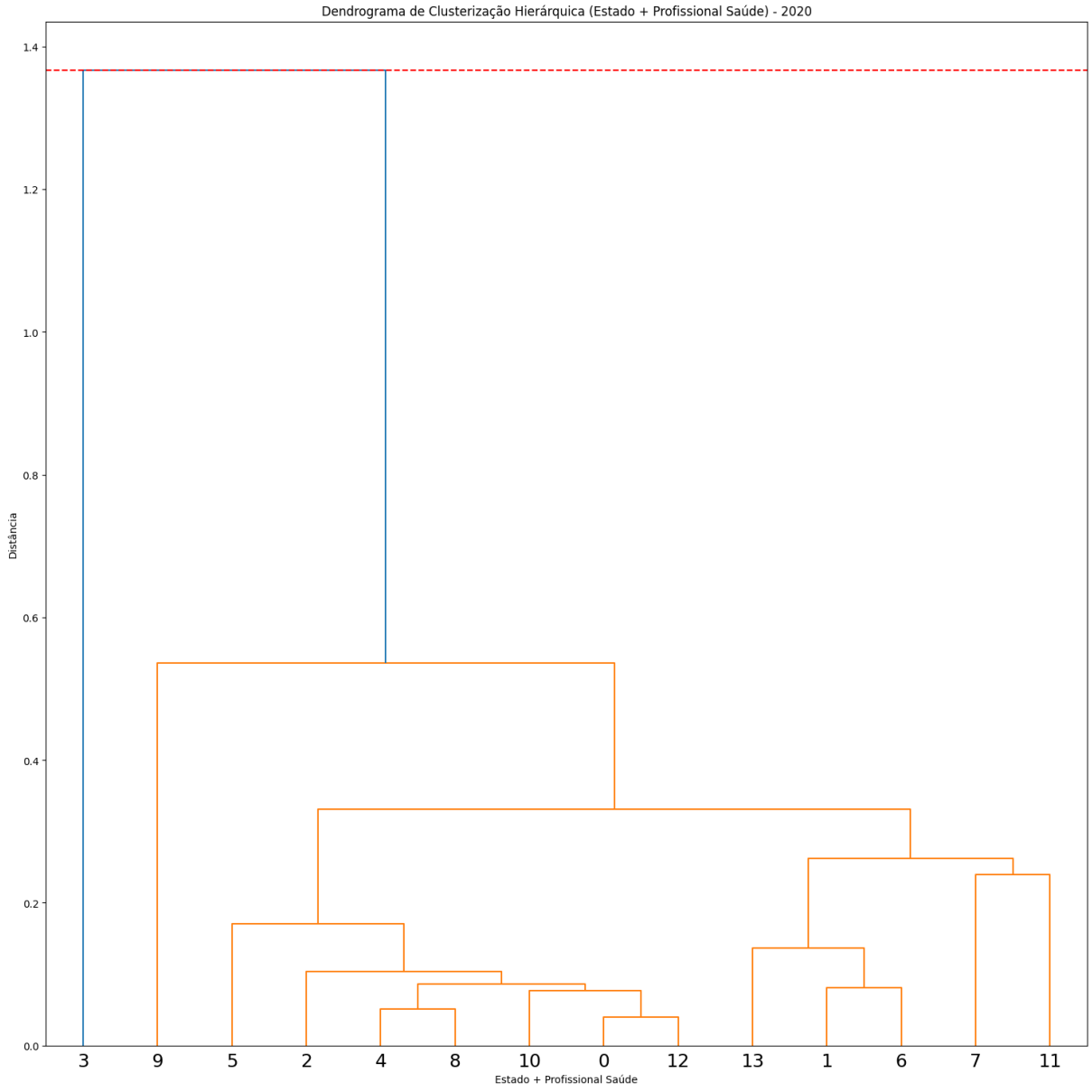
ZAHARIA, M. et al. Apache spark: A unified engine for big data processing. **Communications of the ACM**, v. 59, p. 56–65, 2016.

APÊNDICE A – AGRUPAMENTO ESTADO + PROFISSIONAL DE SAÚDE

Este é uma característica booleana que mostra se a notificação é de uma pessoa que é profissional da saúde ou não.

A.1 – ANO 2020

Figura 37 – Dendrograma dos dados de estado + profissional da saúde 2020 com medida de distância temporal sendo distância euclidiana.

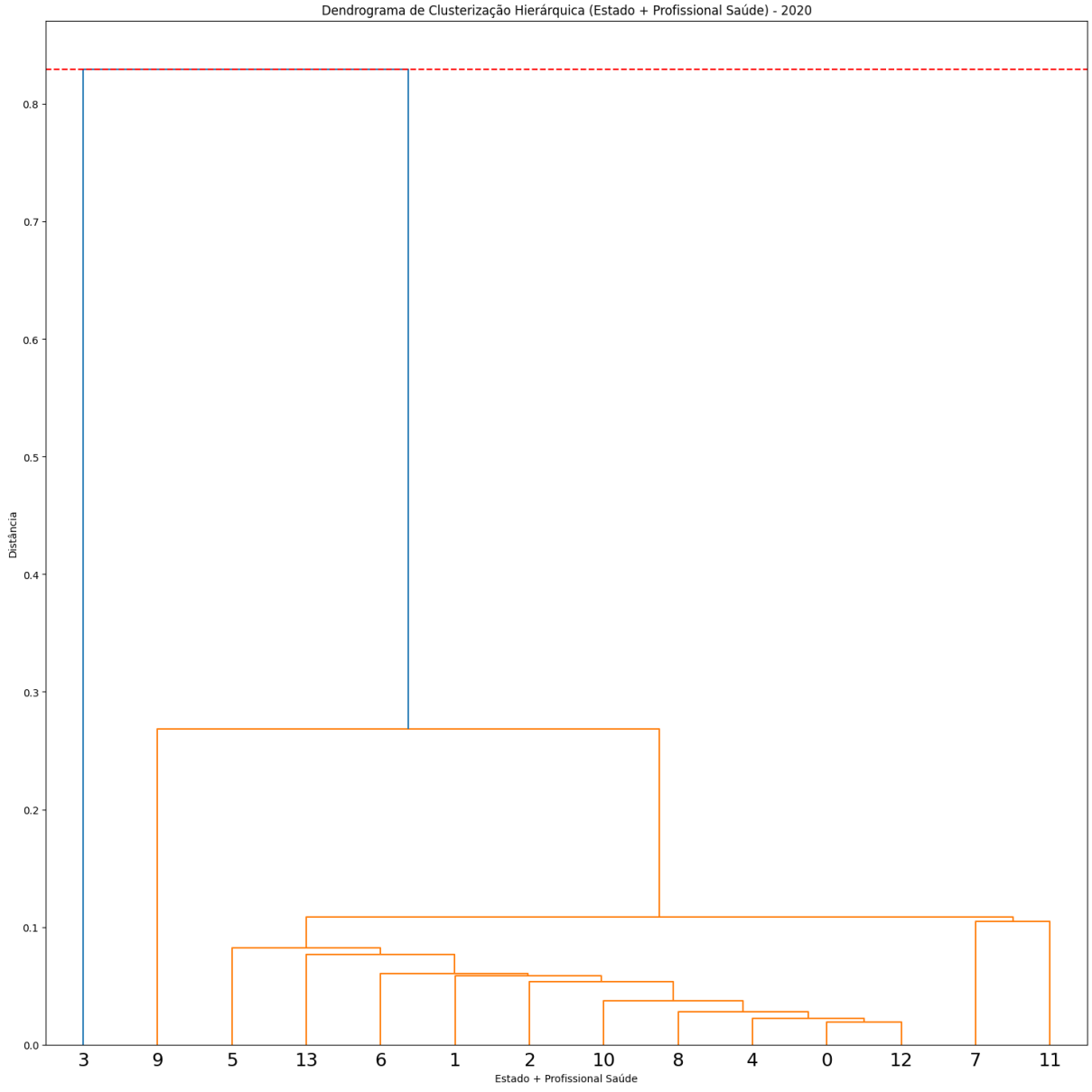


Fonte: Próprio autor.

Cluster	Registros
1	**0:** Bahia (Não), **1:** Bahia (Sim), **2:** Goiás (Não), **4:** Minas Gerais (Não), **5:** Minas Gerais (Sim), **6:** Pará (Não), **7:** Pará (Sim), **8:** Rio Grande do Sul (Não), **9:** Rio Grande do Sul (Sim), **10:** Rio de Janeiro (Não), **11:** Rio de Janeiro (Sim), **12:** São Paulo (Não), **13:** São Paulo (Sim)
2	**3:** Goiás (Sim)

Quadro 22 – Quadro com a separação dos registros da Figura 37.

Figura 38 – Dendrograma dos dados de estado + profissional da saúde 2020 com medida de distância temporal sendo DTW.



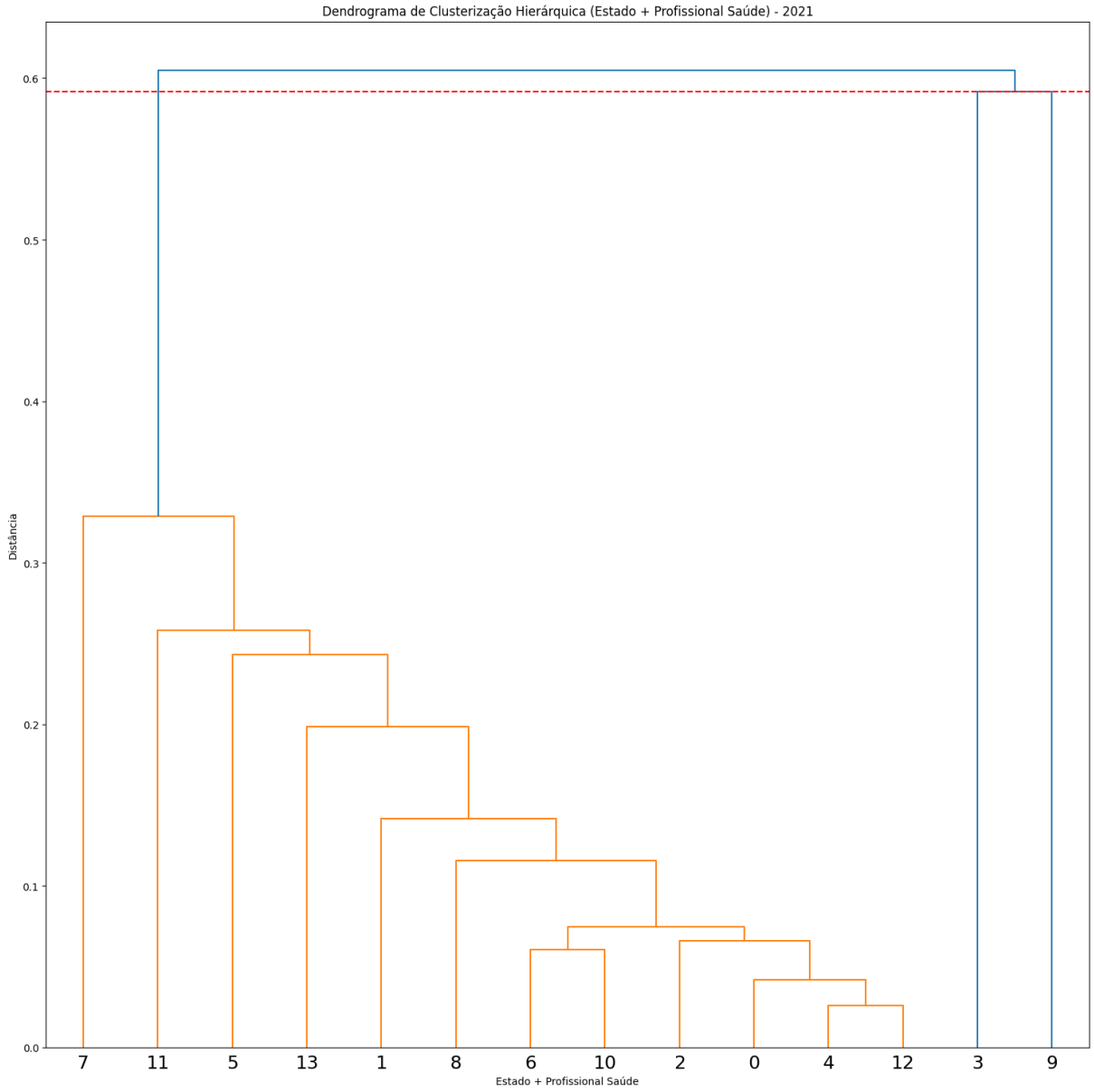
Fonte: Próprio autor.

Cluster	Registros
1	**0:** Bahia (Não), **1:** Bahia (Sim), **2:** Goiás (Não), **4:** Minas Gerais (Não), **5:** Minas Gerais (Sim), **6:** Pará (Não), **7:** Pará (Sim), **8:** Rio Grande do Sul (Não), **9:** Rio Grande do Sul (Sim), **10:** Rio de Janeiro (Não), **11:** Rio de Janeiro (Sim), **12:** São Paulo (Não), **13:** São Paulo (Sim)
2	**3:** Goiás (Sim)

Quadro 23 – Quadro com a separação dos registros da Figura 38.

A.2 – 2021

Figura 39 – Dendrograma dos dados de estado + profissional da saúde 2021 com medida de distância temporal sendo distância euclidiana.

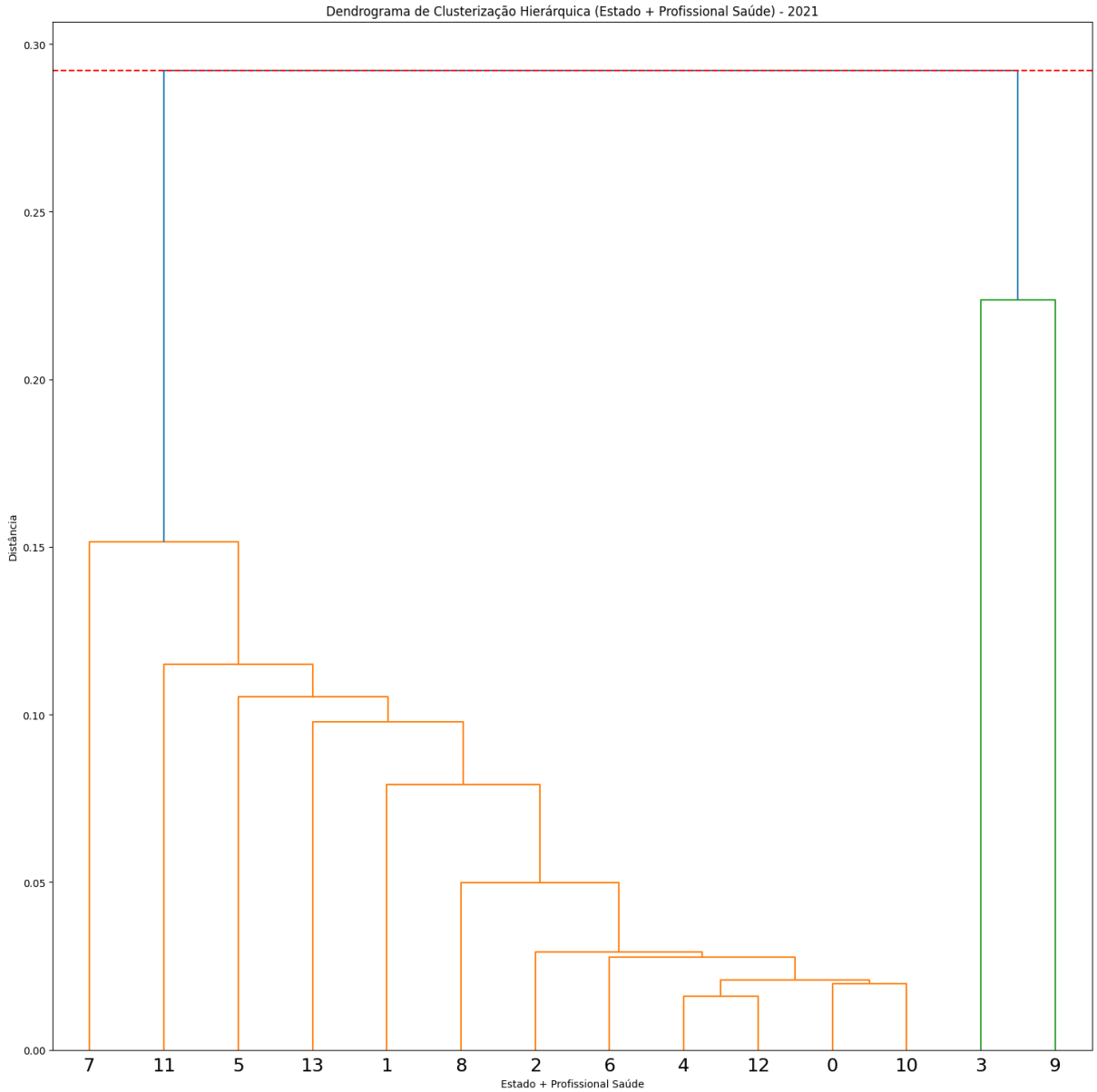


Fonte: Próprio autor.

Cluster	Registros
1	**0:** Bahia (Não), **1:** Bahia (Sim), **2:** Goiás (Não), **4:** Minas Gerais (Não), **5:** Minas Gerais (Sim), **6:** Pará (Não), **7:** Pará (Sim), **8:** Rio Grande do Sul (Não), **10:** Rio de Janeiro (Não), **11:** Rio de Janeiro (Sim), **12:** São Paulo (Não), **13:** São Paulo (Sim)
2	**3:** Goiás (Sim)
3	**9:** Rio Grande do Sul (Sim)

Quadro 24 – Quadro com a separação dos registros da Figura 39.

Figura 40 – Dendrograma dos dados de estado + profissional da saúde 2021 com medida de distância temporal sendo DTW.



Fonte: Próprio autor.

Cluster	Registros
1	**0:** Bahia (Não), **1:** Bahia (Sim), **2:** Goiás (Não), **4:** Minas Gerais (Não), **5:** Minas Gerais (Sim), **6:** Pará (Não), **7:** Pará (Sim), **8:** Rio Grande do Sul (Não), **10:** Rio de Janeiro (Não), **11:** Rio de Janeiro (Sim), **12:** São Paulo (Não), **13:** São Paulo (Sim)
2	**3:** Goiás (Sim), **9:** Rio Grande do Sul (Sim)

Quadro 25 – Quadro com a separação dos registros da Figura 40.