ISSN: 2317-0840

A pandemia de COVID-19 nos municípios do Estado do Paraná: uma investigação via Análise de Agrupamento

Bruna Gabriela Wendpap[1], João Debastiani Neto[2†], Roney Peterson Pereira[1]

[1]*Programa de Pós-Graduação em Engenharia Agrícola - Universidade Estadual do Oeste do Paraná (Unioeste).*
[2]*Departamento de Ciências da Universidade Estadual de Maringá (UEM).*

Resumo: *A presente pesquisa aborda o contexto epidemiológico da pandemia de COVID-19, enfatizando o panorama específico do Brasil, com foco especial no Estado do Paraná, aproximadamente um ano após a implementação do primeiro lockdown. O objetivo geral desta pesquisa consiste em aplicar a técnica de análise de agrupamento conhecida como k-means para categorizar os municípios paranaenses com base em duas variáveis principais: o número diário de casos confirmados e o número de óbitos por COVID-19. Para alcançar esse propósito, foram utilizados dados fornecidos pela Secretaria de Saúde do Paraná, abrangendo o período de 1º de janeiro de 2021 a 15 de março de 2021. Os resultados obtidos revelaram a identificação de três clusters que se destacaram como ótimos, evidenciando divergências nos padrões de incidência de casos e óbitos entre os municípios. Notavelmente, observou-se uma correlação entre a densidade populacional e a frequência de casos e óbitos, com áreas mais densamente povoadas tendendo a registrar números mais elevados. Ademais, a avaliação da precisão dos algoritmos J48 e Naive Bayes na classificação dos clusters demonstrou resultados satisfatórios. Consequentemente, conclui-se que a técnica de agrupamento empregada revelou-se eficaz na identificação de similaridades nos padrões de propagação da COVID-19, oferecendo evidências relevantes para a formulação de estratégias direcionadas e eficientes no enfrentamento da pandemia, especialmente nas regiões mais impactadas.*

Palavras-chave: *Análise de agrupamento; Clusters; J48; Naive Bayes; Covid-19.*

COVID-19 pandemic in the municipalities of the State of Paraná: an investigation via Cluster Analysis

Abstract: *This research addresses the epidemiological context of the COVID-19 pandemic, emphasizing the specific panorama of Brazil, with a special focus on the State of Paraná, approximately one year after the implementation of the first lockdown. The general objective of this research is to apply the clustering technique known as k-means to categorize municipalities in Paraná based on two main variables: the daily number of confirmed cases and the number of deaths from COVID-19. To achieve this purpose, data provided by the Paraná Department of Health were used, covering the period from January 1, 2021 to March 15, 2021. The results obtained revealed the identification of three clusters that stood out as excellent , highlighting divergences in the incidence patterns of cases and deaths between municipalities. Notably, a correlation was observed between population density and the frequency of cases and deaths, with more densely populated areas tending to record higher numbers. Furthermore, the evaluation of the accuracy of the J48 and Naive Bayes algorithms in classifying clusters demonstrated satisfactory results. Consequently, it is concluded that the grouping technique used proved to be effective in identifying similarities in the spread patterns of COVID-19, offering relevant evidence for the formulation of targeted and efficient strategies to combat the pandemic, especially in the most impacted regions.*

Keywords: *Cluster analysis; Clusters; J48; Naive Bayes; Covid-19.*

[†]Autor correspondente: jdneto@uem.br

Introdução

O advento da pandemia de COVID-19, originada em Wuhan, na China, em dezembro de 2019, marcou um dos eventos mais significativos da história contemporânea da saúde pública (LU et al., 2020; EMAMI et al., 2020). O rápido espalhamento do vírus pelo território chinês transformou-se em uma crise de saúde global, desencadeando uma resposta mundial sem precedentes. O presente estudo contextualiza-se nesse cenário, emergindo em um momento em que o Brasil completava um ano desde o início do primeiro *lockdown*.

Durante esse período, observamos uma explosão de informações sobre a COVID-19, abrangendo sua epidemiologia, manifestações clínicas, tratamentos potenciais e, finalmente, a distribuição de vacinas. A proliferação de dados em várias plataformas possibilitou um acompanhamento quase em tempo real da evolução da pandemia, fornecendo uma visão abrangente e dinâmica da situação mundial.

De acordo com dados da Organização Mundial da Saúde (OMS), até 29 de março de 2021, os casos confirmados de COVID-19 ultrapassaram a marca de 126,6 milhões, com mais de 2,7 milhões de óbitos relatados em todo o mundo. No Brasil, nessa mesma data, o Ministério da Saúde registrou mais de 12,5 milhões de casos confirmados e um total de 313.866 óbitos, refletindo a gravidade da situação no país.

O Estado do Paraná, especificamente em março de 2021, enfrentou um colapso no sistema de saúde, evidenciado por uma taxa de ocupação das UTIs para adultos de 96% e uma lista de espera de 1320 pessoas aguardando por leitos, das quais 612 necessitavam de UTIs (SESA, 2021). Essa crise destacou a urgência de compreender e responder de forma eficaz aos desafios impostos pela pandemia.

Os esforços para mitigar os efeitos da COVID-19 têm sido orientados por uma análise constante dos dados epidemiológicos. Cientistas e analistas têm categorizado países, estados e municípios de acordo com seus índices de casos e mortalidade, identificando padrões e tendências que orientam a formulação de políticas públicas eficazes (JAMES & MENZIES, 2020). A abordagem de cluster tem se mostrado particularmente útil nesse contexto, permitindo a identificação de agrupamentos de regiões com características semelhantes e o estudo das dinâmicas locais da pandemia.

James & Menzies (2020) propuseram um método baseado em análise de cluster para analisar a evolução da COVID-19, enquanto estudos como o de Iritani et al. (2020) aplicaram essa técnica em prefeituras do Japão. No Brasil, Guimarães et al. (2020) e Alves et al. (2020) exploraram o método k-means para analisar a disseminação da doença e estratificar o risco de propagação e gravidade da COVID-19.

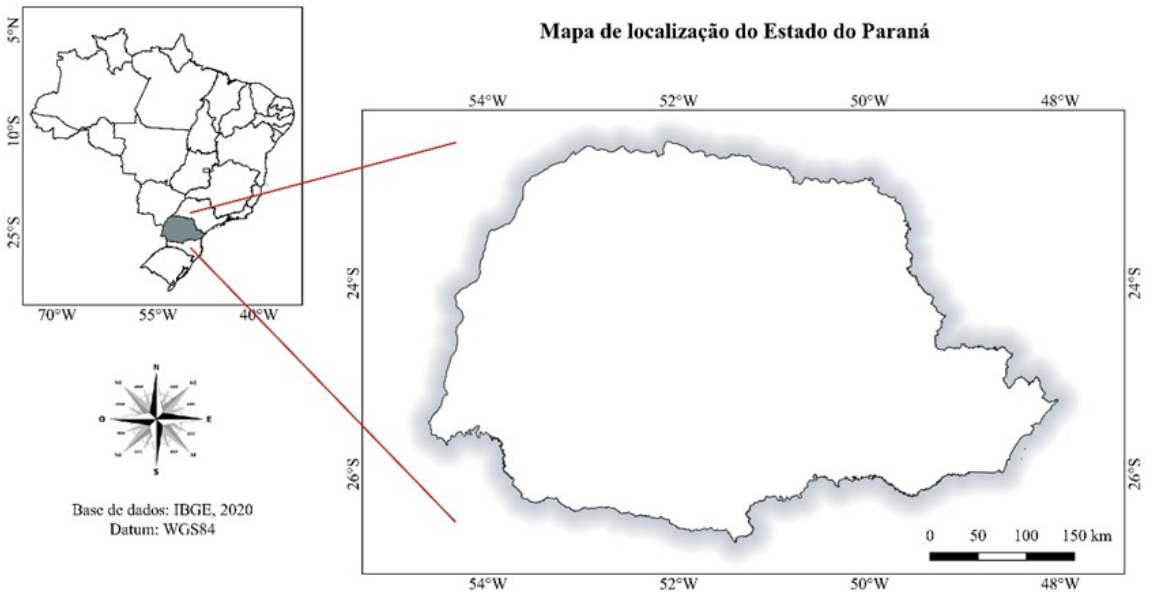
Nesse contexto, o presente estudo tem como objetivo agrupar os municípios do Estado do Paraná utilizando o método k-means, considerando o número diário de casos confirmados e óbitos por COVID-19 como variáveis. Essa abordagem permitirá uma análise mais detalhada dos padrões de disseminação da doença e identificará áreas com maior vulnerabilidade, fornecendo evidências para a formulação de políticas de saúde pública e a alocação eficiente de recursos.

Materiais e Métodos

A presente pesquisa propõe a aplicação do método k-means para a análise e agrupamento dos municípios do Estado do Paraná com base em dados relacionados à pandemia de COVID-19, como mencionado na seção introdutória. A Figura 1 apresenta o mapa de localização do estado.

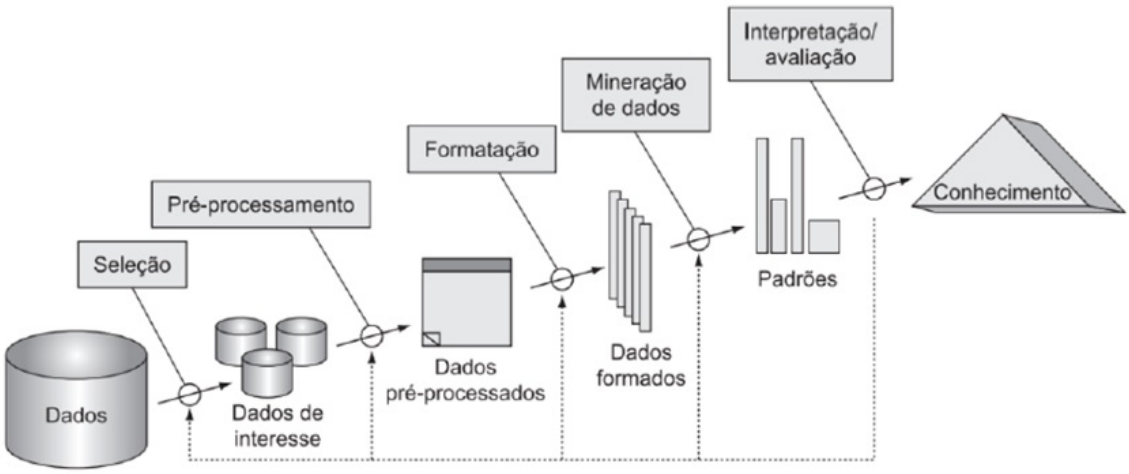
Visando atingir esse objetivo, adotou-se um processo de trabalho sistemático, resumido na Figura 2 e denominado Técnica de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Database - KDD), composto por cinco etapas: seleção, pré-processamento, transformação, mineração de dados e interpretação do resultado (FAYYAD, ET AL.,1996).

Figura 1: Localização da área de estudo



Fonte: Próprio autor.

Figura 2: Etapas do processo KDD



Fonte: FAYYAD et al., (1996).

Na fase inicial, de seleção de dados, foi essencial a escolha criteriosa do conjunto de dados que forneceria elementos relevantes para a análise. Para tal, foram utilizadas informações disponibilizadas pela Secretaria de Saúde do Paraná (SESA, 2021), que compreendiam os números acumulados de casos e óbitos por COVID-19 em cada um dos 399 municípios do estado.

Esses dados foram selecionados para o período de 1 de janeiro a 15 de março de 2021, justificado pelo crescimento da curva epidemiológica após um período de estabilização e queda no final de 2020. Posteriormente, realizou-se o cálculo dos números diários de casos e óbitos para o período considerado, visando uma análise mais detalhada da evolução da pandemia.

A fase subsequente, de pré-processamento, teve como objetivo primordial a limpeza dos dados, com o intuito de garantir a qualidade e consistência das informações utilizadas na análise.

Durante esta etapa, foram identificadas inconsistências nos dados, como valores negativos em algumas datas, indicando possíveis erros de digitação ou subnotificação. Tais registros foram removidos da análise para evitar distorções nos resultados. Os dados foram então organizados e salvos no formato CSV para posterior manipulação e análise nos softwares específicos, como Weka (Witten & Frank, 2005) e R (R Core Team, 2020).

A etapa de Mineração de Dados consistiu em procedimentos para a análise e exploração em amplos volumes de dados por meio de técnicas e algoritmos especializados que tem como objetivo a busca de padrões, previsões, associações e assim por diante.

A terceira etapa, de mineração de dados, constituiu-se na análise e exploração dos dados utilizando técnicas e algoritmos especializados, visando a identificação de padrões e associações relevantes. Para o agrupamento dos municípios com base nos perfis de casos e óbitos por COVID-19, optou-se pela utilização da técnica de análise de cluster, que tem por objetivo classificar observações de um dataset de forma que suas semelhanças sejam alocadas em um mesmo grupo, conforme abordado por Fávero & Belfiore (2019).

O algoritmo k-means foi selecionado para a geração dos clusters no software WEKA, utilizando o conjunto de dados de treinamento. Inicialmente, definiu-se o número de clusters (k) desejado, fundamentado em critérios estatísticos e objetivos da pesquisa. Posteriormente, foram selecionados k pontos arbitrários para representar os centroides dos grupos. Os elementos do conjunto de dados foram então particionados de maneira que cada um fosse atribuído ao grupo cujo centroide estivesse mais próximo, geralmente calculado pela distância euclidiana. Esse processo foi repetido iterativamente, recalculando-se os centroides em cada iteração, até que a convergência fosse alcançada. A determinação do número ideal de clusters foi baseada em critérios estatísticos e métodos como o proposto por Ratkowsky & Lance (1978), utilizando o pacote NbClust no software R (R Core Team, 2020).

Para a avaliação da eficácia do modelo proposto, empregamos a métrica de acurácia, reconhecida como o principal indicador de desempenho em problemas de classificação (MACIEL et al., 2020). Os algoritmos J48, uma implementação do método C4.5 de árvore de decisão, e Naive Bayes foram selecionados devido à sua ampla utilização e eficácia em tarefas de classificação. O J48 constrói uma árvore de decisão iterativa, visando melhorar a classificação. Por sua vez, o Naive Bayes, baseado no teorema de Bayes, presume independência condicional entre os atributos, e é reconhecido por seu desempenho, especialmente em conjuntos de dados menores.

Por meio dessas etapas, espera-se identificar padrões nos dados relacionados à propagação da COVID-19 nos municípios do Paraná, fornecendo informações valiosas para o planejamento e implementação de medidas de saúde pública. Além disso, os resultados obtidos poderão contribuir para a compreensão dos fatores que influenciam a disseminação da doença em diferentes regiões e auxiliar na tomada de decisões para o enfrentamento da pandemia.

Resultados e discussão

O critério para determinação do número de clusters adotado neste estudo foi fundamentado na proposta de Ratkowsky & Lance (1978), uma abordagem consolidada na literatura para a seleção do número ideal de agrupamentos. Tal critério foi implementado no ambiente estatístico R (Charrad et al., 2015; R Core Team, 2020) por meio do pacote NbClust, que é uma ferramenta robusta para análise de agrupamentos.

A função NbClust foi utilizada para avaliar os números de clusters variando de 2 a 10, com o intuito de identificar a configuração que melhor se adequava aos dados em estudo. Após a análise, verificou-se que o modelo com 3 clusters apresentou o melhor desempenho, conforme evidenciado pelo maior índice Ratkowsky Lance. Todos os índices obtidos durante o processo de avaliação estão detalhadamente registrados na Tabela 2, fornecendo uma visão abrangente da qualidade dos agrupamentos em questão.

Tabela 1: Índice Ratkowsky Lance

nº clusters	casos	óbitos
2	0.1649	0.3413
3	0.4295	0,3828
4	0,4128	0,3540
5	0,3731	0,3206
6	0,3418	0,2953
7	0,3166	0,2762
8	0,2972	0,2595
9	0,2807	0,2450
10	0,2666	0,2342

Fonte: Próprio autor.

Por meio da Tabela 2, observa-se que, para ambos os casos e óbitos, o índice de Ratkowsky Lance atinge seu valor máximo quando o número de clusters é igual a 3. Isso sugere que o modelo com 3 clusters apresenta a melhor divisão dos dados, com agrupamentos mais distintos e homogêneos em relação à variabilidade total dos dados. À medida que o número de clusters aumenta, o índice Ratkowsky Lance tende a diminuir, indicando uma menor qualidade na separação dos grupos. Isso pode ser interpretado como uma maior sobreposição entre os agrupamentos ou uma menor coesão dentro de cada cluster, o que pode dificultar a interpretação e aplicação dos resultados obtidos.

Portanto, com base na análise do Índice Ratkowsky Lance, pode-se concluir que o modelo com 3 clusters é o mais adequado para representar a estrutura dos dados de casos e óbitos relacionados à COVID-19 nos municípios do Estado do Paraná.

Uma vez estabelecido o número ideal de clusters, procedeu-se à construção do mapa do Estado do Paraná, no qual todos os seus municípios foram categorizados conforme os resultados advindos da aplicação do algoritmo K-Means, empregando $k = 3$, por meio do *software* WEKA. Esse procedimento considerou as variáveis relacionadas ao número de casos e óbitos por COVID-19 em cada município.

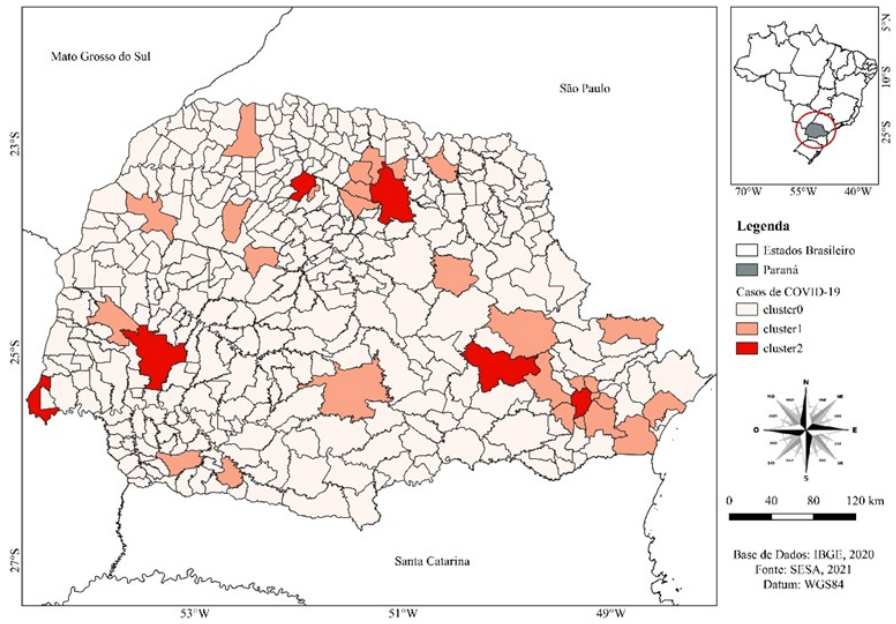
No que concerne à categorização dos municípios de acordo com o número de novos casos diários registrados durante o primeiro trimestre de 2021 (Figura 3), merecem destaque os municípios pertencentes ao cluster 2, a saber: Cascavel, Maringá, Londrina, Foz do Iguaçu, Ponta Grossa e a capital Curitiba, os quais apresentaram uma média de 280 novos casos diários. Em contrapartida, para o cluster 1, composto por 28 municípios, a média de novos casos diários foi de 42. Por fim, o cluster 0, compreendendo os 365 municípios restantes, registrou uma média de 4 novos casos diários.

No contexto da categorização dos municípios em relação ao número de óbitos diários durante o primeiro trimestre de 2021 (Figura 4), é notável a distinção observada nos agrupamentos. Os municípios reunidos no cluster 2, composto por Cascavel, Maringá, Londrina, Foz do Iguaçu, Ponta Grossa e a capital Curitiba, destacaram-se com uma média de 5 óbitos diários. Esses números evidenciam uma situação preocupante, indicando uma considerável incidência de fatalidades em áreas urbanas de maior densidade populacional e, possivelmente, com infraestrutura de saúde mais sobrecarregada.

Por outro lado, o cluster 0, compreendendo 39 municípios, apresentou uma média de 0,55 óbitos por dia. Esta taxa pode ser interpretada como significativa, representando, em média, 5 óbitos a cada 9 dias. Esta dinâmica de óbitos sugere uma situação de menor gravidade em comparação com o cluster 2, no entanto, ainda é importante monitorar de perto essas localidades para evitar uma escalada no número de fatalidades.

Já o cluster 1, composto pelos 354 municípios restantes, registrou uma média de 0,07 óbitos por dia, o que equivale a um óbito a cada 14 dias. Apesar de aparentemente apresentar uma

Figura 3: Agrupamento para o número de casos.

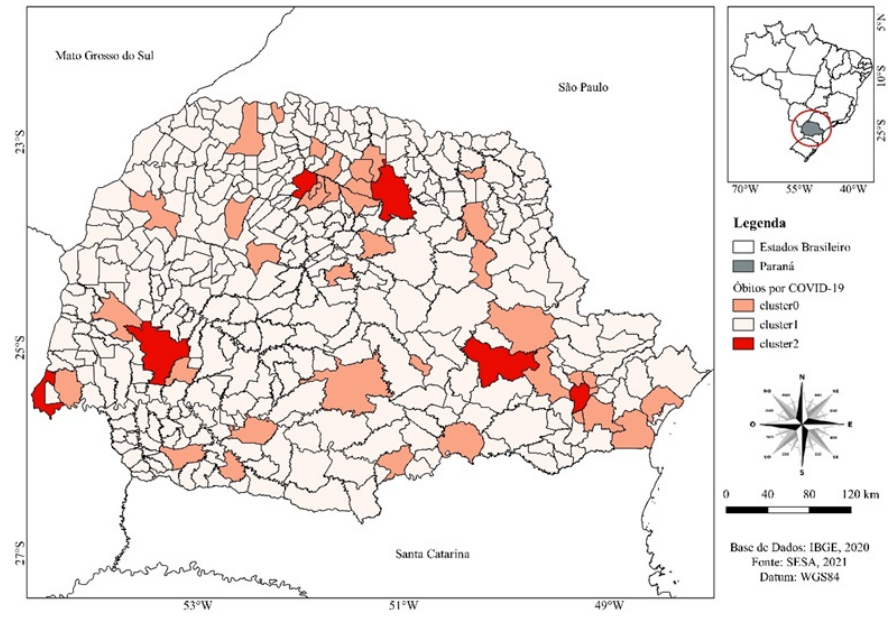


Fonte: Próprio autor.

situação menos crítica em comparação com os clusters 0 e 2, essa média não deve ser negligenciada, pois qualquer aumento repentino na taxa de mortalidade pode indicar uma mudança na dinâmica da pandemia nesses municípios.

Esses resultados ressaltam a importância da análise de agrupamentos na identificação de padrões e tendências relacionadas à disseminação e impacto da COVID-19 em diferentes regiões. Além disso, destacam a necessidade de estratégias diferenciadas de intervenção e controle da pandemia, adaptadas às características específicas de cada grupo de municípios.

Figura 4: Agrupamento para o número de óbitos



Fonte: Próprio autor.

Os resultados obtidos indicam uma correlação entre a densidade populacional e a incidência de casos e óbitos por COVID-19 nos municípios mais populosos do estado. A exceção notável de São José do Pinhais ressalta a complexidade dos fatores que influenciam na disseminação do vírus e na mortalidade associada a ele. A maior densidade populacional desses municípios pode estar contribuindo para uma disseminação mais intensa do vírus, exacerbando assim o número de casos e óbitos registrados.

Como métrica para avaliar a eficácia do modelo proposto, empregamos a acurácia, uma vez que é o principal indicador de desempenho em problemas de classificação (MACIEL et al., 2020). Os algoritmos J48 e Naive Bayes foram selecionados para essa análise devido à sua ampla utilização e eficácia comprovada em tarefas de classificação. Optamos por implementar ambos os algoritmos utilizando a técnica de validação cruzada com 10 partições, a fim de garantir uma avaliação robusta e confiável da classificação dos clusters em relação aos casos e óbitos de COVID-19.

A escolha desses algoritmos se baseou em sua capacidade de lidar com conjuntos de dados complexos e multidimensionais, como é o caso dos dados epidemiológicos analisados neste estudo. Assim, a Tabela 2 apresenta os resultados da acurácia da classificação dos clusters para casos e óbitos de COVID-19, obtidos por meio dos algoritmos J48 e Naive Bayes.

Tabela 2: Acurácia da classificação dos clusters

Algoritmo	**Casos**	**Óbitos**
J48	94,988%	94,236%
Naive Bayes	93,233%	91.98%

Fonte: Próprio autor.

Para os casos de COVID-19, o algoritmo J48 alcançou uma acurácia de 94,988%, enquanto o Naive Bayes obteve uma acurácia ligeiramente menor, com 93,233%. Isso indica que ambos os algoritmos são altamente eficazes na classificação dos clusters com base nos casos de COVID-19, com o J48 apresentando uma performance ligeiramente superior em relação ao Naive Bayes.

Quanto aos óbitos por COVID-19, o algoritmo J48 demonstrou uma acurácia de 94,236%, enquanto o Naive Bayes alcançou uma acurácia de 91,980%. Novamente, o algoritmo J48 superou o Naive Bayes em termos de acurácia na classificação dos clusters em relação aos óbitos de COVID-19.

Esses resultados indicam que tanto o algoritmo J48 quanto o Naive Bayes são eficientes na classificação dos clusters para casos e óbitos de COVID-19, com acurácias bastante elevadas. No entanto, é importante notar que o algoritmo J48 tende a apresentar uma performance ligeiramente superior em comparação ao Naive Bayes, especialmente no que diz respeito aos óbitos.

Essa análise reforça a robustez dos modelos de classificação implementados neste estudo e sugere que eles são capazes de identificar padrões epidemiológicos relevantes nos dados de casos e óbitos de COVID-19. Esses resultados são essenciais para orientar a tomada de decisão em políticas de saúde pública e contribuir para o controle eficaz da pandemia.

Conclusões

Esta pesquisa abordou o contexto epidemiológico da pandemia de COVID-19, com ênfase no Estado do Paraná, aproximadamente um ano após a implementação do primeiro *lockdown*. O estudo teve como objetivo principal aplicar a técnica de análise de agrupamento k-means para categorizar os municípios paranaenses com base no número diário de casos confirmados e óbitos por COVID-19. Os dados utilizados foram fornecidos pela Secretaria de Saúde do Paraná, abrangendo o período de 1º de janeiro de 2021 a 15 de março de 2021.

Os resultados revelaram a identificação de três clusters ótimos, evidenciando divergências nos padrões de incidência de casos e óbitos entre os municípios. Foi observada uma correlação entre a densidade populacional e a frequência de casos e óbitos, com áreas mais densamente povoadas tendendo a registrar números mais elevados.

A avaliação da precisão dos algoritmos J48 e Naive Bayes na classificação dos clusters demonstrou resultados satisfatórios. O algoritmo J48 apresentou uma performance ligeiramente superior ao Naive Bayes, especialmente na classificação dos óbitos por COVID-19.

Os resultados indicam a eficácia da técnica de agrupamento na identificação de similaridades nos padrões de propagação da COVID-19, oferecendo evidências relevantes para a formulação de estratégias direcionadas no enfrentamento da pandemia, especialmente nas regiões mais impactadas. Além disso, por meio da análise dos dados obteve-se uma correlação entre a densidade populacional e a incidência de casos e óbitos por COVID-19 nos municípios mais populosos do estado, com exceção de São José do Pinhais. Os resultados da acurácia dos algoritmos J48 e Naive Bayes indicaram uma alta eficácia na classificação dos clusters, destacando a robustez dos modelos implementados neste estudo.

Esses resultados são fundamentais para orientar a tomada de decisões em políticas de saúde pública e contribuir para o controle eficaz da pandemia, fornecendo informações valiosas sobre a dinâmica da COVID-19 nos municípios do Paraná.

Referências

ALVES, H. J. P.; FERNANDES, F. A.; LIMA, K. P.; BATISTA, B. D. O.; FERNANDES, T. J. *A pandemia da COVID-19 no Brasil: uma aplicação do método de clusterização k-means.* Research, Society and Development, v. 9, n. 10, 2020. DOI: http://dx.doi.org/10.33448/rsd-v9i10.9059.

CHARRAD, M., GHAZZALI, N., BOITEAU, V., NIKNAFS, A. *Determining the best number of clusters in a data set.* Package NbClust, 2015. Recuperado de http://cran.rediris.es/web/packages/NbClust/NbClust.pdf.

EMAMI, A.; JAVANMARDI, F.; PIRBONYEH, N.; AKBARI, A. *Prevalence of Underlying Diseases in Hospitalized Patients with COVID-19: a Systematic Review and Meta-Analysis.* Arch Acad Emerg Med. 8(1): e35, mar, 2020.

FÁVERO, L. P.; BELFIORE, P. *Data Science for Business and Decision Making.* Academic Press, Cambridge, MA, USA, 2019

FAYYAD, U.M. et al. *Advances in knowledge discovery and data mining.* Massachusetts: AAAI Press, 1996.

GUIMARÃES, R. M.; ELEUTERIO, T. D. A.; MONTEIRO-DA-SILVA, J. H. C. *Estratificação de risco para predição de disseminação e gravidade da Covid-19 no Brasil.* Revista Brasileira De Estudos De População, 37, 1-17, 2020. DOI: http://dx.doi.org/10.20947/s0102-3098a0122.

IRITANI, O.; OKUNO, T.; HAMA, D.; KANE, A.; KODERA, K.; MORIGAKI, K.; TERAI, T.; MAENO, N.; MORIMOTO, S. *Clusters of covid-19 in long-term care hospitals and facilities in japan from 16 january to 9 may 2020.* Geriatrics & gerontology international, 20(7), 715-719, 2020. DOI: 10.1111/ggi.13973.

JAMES, N.; MENZIES, M. *Cluster-based dual evolution for multivariate time series:*

Analyzing covid-19. Chaos: An Interdisciplinary Journal of Nonlinear Science, 30, 2020. DOI: https://doi.org/10.1063/5.0013156.

LU, R.; ZHAO, X.; LI, J.; NIU, P.; YANG, B.; WU, H.; WANG, W. *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding*. The Lancet. v.395, Feb 22, P. 565-574, 2020. https://doi.org/10.1016/S0140-6736(19)33096-X.

MACIEL, E. L.; JABOR, P.; GONÇALVES JÚNIOR, E.; TRISTÃO-SÁ, R.; LIMA, R.C.D.; REIS-SANTOS, B.; LIRA, P.; BUSSINGUER, E. C. A.; ZANDONADE, E. *Fatores associados ao óbito hospitalar por covid-19 no Espírito Santo*. Epidemiologia e Serviços de Saúde, 29(4), 1-11, 2020. DOI: 10.5123/S1679-49742020000400022.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical computing, Vienna, 2020. Disponível em: https://www.Rproject.org/.

RATKOWSKY, D.; LANCE, G. *Criterion for determining the number of groups in a classification*. Australian Computer Journal, 10(3), 115-117, 1978.

SESA, Secretaria da saúde: *Informe Epidemiológico Coronavírus (COVID-19)*. Boletim epidemiológico, Curitiba, Março, 2021. Disponível em: https://abrir.link/Sdiia

WITTEN IH, F. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edition. Morgan Kaufmann, San Francisco, 2005.