A Aplicação do Processo de KDD aos Dados da COVID-19: Um Estudo de Caso no Rio Grande do Sul, Brasil

Gabriel V. Heisler[1], **Joaquim V. C. Assunção**[1]

[1]Universidade Federal de Santa Maria (UFSM)
Santa Maria – RS – Brasil

{gvheisler,joaquim}@inf.ufsm.br

Abstract. *Given the increasing amount of data linked to a complex healthcare system, challenges arise in enhancing decision-making based on data patterns. To address this issue, it is crucial to explore data mining as a tool to extract valuable insights. This study focuses on applying the Knowledge Discovery in Databases (KDD) process, particularly in the preliminary and data mining stages, to identify patterns in the COVID-19 pandemic data in Rio Grande do Sul, Brazil. Our analyses revealed interesting patterns, such as the association between specific symptoms and patient outcomes. While the results offer valuable insights, it is important to note that this study does not aim to provide definitive conclusions regarding the causal relationship between symptoms and patient outcomes. Instead, the goal is to present patterns identified in the data without interpreting their clinical significance. These findings have the potential to inform future research and provide a solid foundation for proactive decision-making in public health.*

Resumo. *Diante da crescente quantidade de dados vinculados a um sistema de saúde complexo, surgem desafios para aprimorar a tomada de decisões com base em padrões de dados. Para enfrentar essa questão, é fundamental explorar a mineração de dados como uma ferramenta para extrair insights valiosos. Este estudo focaliza a aplicação do processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases – KDD), especialmente nas fases preliminares e de mineração de dados, para identificar padrões nos dados da pandemia de COVID-19 no Rio Grande do Sul, Brasil. Nossas análises revelaram padrões interessantes, como a associação entre sintomas específicos e desfechos dos pacientes. Embora os resultados ofereçam insights valiosos, é importante ressaltar que este estudo não tem a intenção de fornecer conclusões definitivas sobre a relação causal entre os sintomas e os resultados dos pacientes. Em vez disso, busca-se apresentar padrões identificados nos dados, sem interpretar seu significado clínico. Essas descobertas têm o potencial de informar futuras investigações e fornecer uma base sólida para a tomada de decisões proativas em saúde pública.*

1. Introdução

O processo de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases – KDD*) encontra aplicação em diversos domínios científicos, incluindo o campo crítico da saúde. Uma aplicação proeminente reside na epidemiologia, onde

os pesquisadores se esforçam para prever surtos de doenças, bem como identificar correlações entre sintomas e mortalidade. Notavelmente, a pandemia de COVID-19, causada pelo vírus *SARS-CoV-2*, foi declarada uma pandemia global em 11 de março de 2020 [Cucinotta and Vanelli 2020]. Aproveitando o poder da mineração de dados, esforços podem ser feitos para detectar padrões úteis e entender melhor o impacto da doença, levando a uma possível melhor preparação das organizações de saúde pública no que diz respeito à tomada de decisões baseadas em dados.

Embora haja uma miríade de esforços em relação ao COVID-19, apenas um número limitado de estudos tem se concentrado especificamente no estado do Rio Grande do Sul, Brasil [Dagnino et al. 2020][Hallal et al. 2020]. Além disso, a maioria dessas investigações baseia-se em métodos estatísticos simples ou utiliza conjuntos de dados em pequena escala [Silveira et al. 2020]. Em contraste, este estudo preliminar emprega o processo de Descoberta de Conhecimento em Bases de Dados para fornecer uma análise abrangente dos dados atuais, oferecendo padrões apoiados por métricas robustas, para o estado do Rio Grande do Sul. Ao utilizar o processo de KDD, o objetivo deste trabalho[1] é preencher a lacuna na pesquisa e contribuir com *insights* valiosos sobre a situação do COVID-19 na região. A escolha desse estado como foco do estudo é motivada também pela disponibilidade dos dados, os quais já foram compilados pelo governo estadual e são mostrados em um painel com *dashboards*, e também disponibilizados para download.

2. Metodologia

O processo de Descoberta de Conhecimento em Bases de Dados (frequentemente referido pela sigla em inglês, "KDD") é amplamente utilizado e possui várias versões propostas. Neste trabalho a proposta aderida é a de [Fayyad et al. 1996]. Este processo visa não apenas utilizar algoritmos de mineração de dados, mas sim realizar etapas prévias, com o intuito de "moldar" os dados para a mineração. Em linha com este processo (Figura. 1) proposto por [Fayyad et al. 1996], primeiramente foram reunidos dados abrangentes sobre COVID-19 para o estado do Rio Grande do Sul, Brasil. O foco foi obter o conjunto de dados mais completo disponível (o qual é atualizado constantemente). Posteriormente, foi feita a seleção cuidadosa de atributos pertinentes para extrair regras relacionadas aos sintomas de casos fatais. Para facilitar a análise, os dados foram divididos em dois formatos: um para regras de associação (Seção 5.1) e outro para fins de classificação (Seção 5.2). As descobertas também são discutidas nessas seções.

3. Dados

Durante a pandemia de COVID-19, o governo do Rio Grande do Sul desenvolveu um painel[2] abrangente de dados sobre esta doença para facilitar o acesso a informações cruciais dentro do estado. Este conjunto de dados, que inclui dados sobre casos da doença, vacinação, leitos hospitalares e mais, permanece publicamente disponível por meio da interface do painel e também pode ser baixado no formato CSV. Neste estudo, o conjunto de dados de casos de COVID-19 utilizado foi o de casos confirmados, abrangendo o período de 2020 a 2024. A coleta dos dados foi feita diretamente do site oficial. No entanto, para

[1]Este trabalho tem o objetivo específico de mostrar informações ocultas nos dados. Quaisquer interpretações relativas a diagnósticos não fazem parte da área dos autores e, portanto, estão fora do escopo desse trabalho

[2]*ti.saude.rs.gov.br/covid19/*

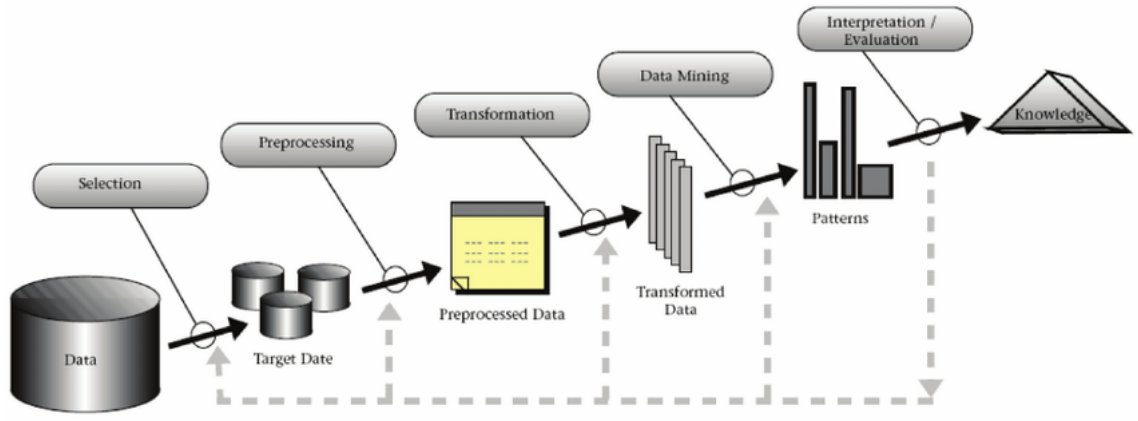


Figura 1. Passos do KDD, ilustrados por [Fayyad et al. 1996]

manter o foco na relevância, foram utilizados apenas os dados de 25 de fevereiro de 2020 a 31 de dezembro de 2022, principal época da pandemia. A Figura 2 mostra o dicionário do conjunto de dados.

COD_IBGE	Código IBGE do Município	**GARGANTA**	Sintomas de dor de garganta
MUNICIPIO	Nome do município	**DISPNEIA**	Sintomas de dor de dispnéia/falta de ar
COD_REGIAO_COVID	Código da região de saúde COVID	**OUTROS**	Outros sintomas
REGIAO_COVID	Nome da região de saúde COVID	**CONDICOES**	Alguma condição de saúde que necessite atenção
SEXO	Sexo	**GESTANTE**	Paciente é gestante
FAIXAETARIA	Faixa Etária	**DATA_INCLUSAO_OBITO**	Data de inclusão da informação de óbito
CRITERIO	Tipo de teste	**DATA_EVOLUCAO_ESTIMADA**	Data da evolução estimada (casos não hospitalizados)
DATA_CONFIRMACAO	Data de confirmação	**RACA_COR**	Raça/Cor
DATA_SINTOMAS	Data de início dos sintomas	**ETNIA_INDIGENA**	Etnia indígena
DATA_INCLUSAO	Data de inclusão no dashboard do RS	**PROFISSIONAL_SAUDE**	Profissional de saúde
DATA_EVOLUCAO	Data da evolução	**BAIRRO**	Bairro (apenas municípios com mais de 100.000 hab)
EVOLUCAO	Descrição da evolução	**SRAG**	Paciente apresentou síndrome respiratória aguda grave
HOSPITALIZADO	Paciente foi hospitalizado	**FONTE_INFORMACAO**	Fonte da informação
FEBRE	Sintomas de febre	**PAIS_NASCIMENTO**	País de nascimento
TOSSE	Sintomas de tosse	**PES_PRIV_LIBERDADE**	Pessoa privada de liberdade

Figura 2. Dicionário de Dados

4. Primeiras etapas do KDD

Nesta pesquisa, todas as etapas do KDD foram utilizadas. Conforme mostrado na Figura 1, existem cinco etapas neste processo, no entanto, considerando o escopo deste trabalho, apenas as mais relevantes são mostradas.

4.1. Extração/Seleção

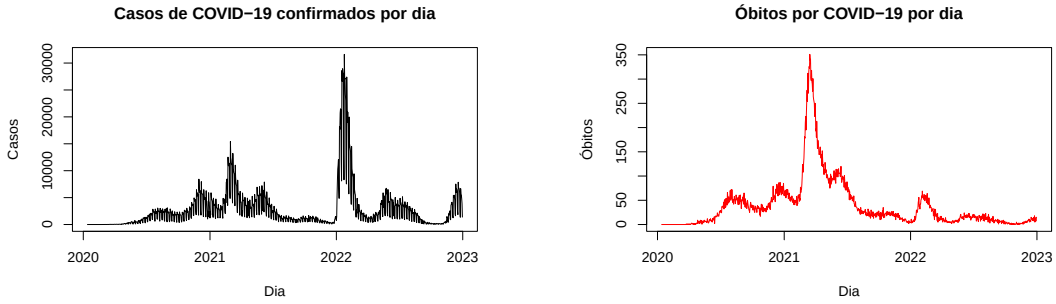
Ao analisar o conjunto de dados, torna-se aparente o volume substancial de dados que abrange todo o período da pandemia. Seguindo o processo de KDD, o primeiro passo é selecionar cuidadosamente os dados relevantes para a análise. Para este trabalho[3] foi utilizada a linguagem R. Primeiramente, os dados foram extraídos e inseridos em um

[3]Código fonte disponível em *github.com/gvheisler/KDD-COVID-19*

data.frame apropriado. Após esta etapa inicial, foi observado que o conjunto de dados consiste em cerca de 3 milhões de linhas, ocupando mais de 600MB de espaço, e abrange 32 colunas. Cada linha representa um caso confirmado de COVID-19 no estado, com detalhes resumidos na Figura 2. Antes de aplicar algoritmos de mineração de dados a este conjunto de dados, é essencial realizar os estágios cruciais de limpeza e pré-processamento de dados, conforme recomendado por [Fayyad et al. 1996]. Além disso, foram geradas ilustrações gráficas que mostram características-chave do conjunto de dados, apresentadas na Figura 3.

4.2. Visualização

A etapa de visualização dos dados é um componente importante do processo de KDD, podendo ser repetida várias vezes ao longo do ciclo. Esta etapa desempenha um papel essencial tanto na identificação de padrões visuais quanto na análise dos resultados produzidos pelos algoritmos empregados. Neste estudo, inicialmente foram gerados gráficos para proporcionar uma visão geral da situação da COVID-19 no estado. A Figura 3 apresenta dois gráficos destacados. Para criar esses gráficos, novos conjuntos de dados foram derivados do conjunto original. Esta abordagem de criação de novos conjuntos de dados possibilita uma exploração mais detalhada das relações e padrões presentes nos dados originais, promovendo uma análise mais abrangente e esclarecedora. A Figura 3a mostra



(a) Casos confirmados diários

(b) Óbitos diários

Figura 3. Casos e óbitos por COVID-19 diários

um gráfico com algumas informações interessantes e importantes, como o fato de que houve alguns picos dispersos nos casos confirmados. O gráfico da Figura 3b também exibe alguns picos que, logicamente, correlacionam-se com o gráfico 3a.

5. Mineração de dados

Esta seção está dividida em duas partes, de acordo com o tipo de algoritmo de aprendizado. Primeiro, a Seção 5.1 é dedicada às regras de associação obtidas. Em segundo lugar, a Seção 5.2 foca na classificação e seus resultados, principalmente ilustrados pela árvore de classificação. Todas as descobertas descritas nas próximas seções revelam padrões e informações ocultas nos dados.

5.1. Associação

Algumas colunas no conjunto de dados representam informações sobre os sintomas que cada paciente teve. Essas colunas têm apenas dois valores possíveis, “SIM” e “NAO”, que são adequados quando se trata de aplicar o algoritmo de associação, Apriori.

Quando usado com dados que têm apenas dois valores possíveis (binários), o algoritmo de associação Apriori verifica os conjuntos de itens mais fortemente associados. Esse processo é feito usando o Princípio do Apriori, que afirma que se um conjunto de itens for infrequente, seus subconjuntos também serão infrequentes. Isso torna esse algoritmo mais rápido e eficiente do que uma busca exaustiva por associações[Agrawal et al. 1996].

Neste estudo, para utilizar o algoritmo Apriori, foi utilizada a biblioteca Arules [Hahsler et al. 2005]. O algoritmo foi aplicado aos sintomas e ao estado do paciente, para encontrar quais sintomas são mais comumente encontrados em pacientes que faleceram. No entanto, antes de aplicá-lo, foram realizadas as outras etapas do processo de KDD.

Primeiramente, foi feita a seleção dos dados a serem utilizados. Após verificar o conjunto de dados, fica claro que, em geral, os dados estão completos, sem nenhuma informação importante faltando. As colunas a serem utilizadas foram então selecionadas, como mostrado na Figura 4.

EVOLUCAO	FEBRE	TOSSE	GARGANTA	DISPNEIA	OUTROS
RECUPERADO	NAO	NAO	NAO	NAO	SIM
RECUPERADO	SIM	SIM	SIM	NAO	NAO
RECUPERADO	NAO	SIM	NAO	NAO	NAO
RECUPERADO	NAO	SIM	SIM	NAO	NAO
RECUPERADO	NAO	SIM	NAO	SIM	SIM
RECUPERADO	NAO	NAO	NAO	NAO	NAO
OBITO	SIM	SIM	NAO	SIM	NAO
RECUPERADO	SIM	NAO	NAO	SIM	NAO
RECUPERADO	SIM	SIM	NAO	SIM	SIM
RECUPERADO	SIM	SIM	NAO	NAO	SIM

Figura 4. *Dataframe* contendo os sintomas e a evolução do paciente

Após a transformação dos dados para o formato de atributos binários, onde cada coluna representa a presença (“SIM”) ou ausência (“NAO”) de um determinado sintoma, os dados foram submetidos ao algoritmo Apriori para a geração de regras de associação. Este algoritmo requer a definição de parâmetros como suporte e confiança. O suporte é a frequência com que um conjunto de itens aparece juntos nos dados, enquanto a confiança é a probabilidade de que um item ocorra dado outro já ocorreu. Ambos são parâmetros essenciais para o algoritmo Apriori na geração de regras de associação. Inicialmente, o algoritmo foi executado com uma confiança de 0.001, visando apenas a verificação do suporte mínimo necessário, além de diferentes níveis de suporte.

Os dados foram filtrados para exibir somente as regras que envolvem a evolução da doença, permitindo que o visualizador identifique se o paciente sobreviveu ou não. O suporte mínimo foi determinado por meio de uma série de testes, com base na quantidade de regras relevantes geradas. Ao estabelecer um suporte mínimo de 0.5, foram obtidas 7 regras; esse número aumentou para 27 quando o suporte mínimo foi ajustado para 0.3 e chegou a 104 com um suporte mínimo de 0.1. No entanto, ao analisar os resultados, foi observado que apenas regras relacionadas ao desfecho “RECUPERADO” foram geradas.

Existem algumas abordagens para gerar regras com melhores métricas (aquelas com maior suporte e confiança), mas que não sejam óbvias. Nesta pesquisa, duas delas são utilizadas:

Tabela 1. Regras geradas com um suporte mínimo de 0.01 e confiança mínima de 0.001, buscando regras com óbito como rhs

lhs	rhs	support	confidence
DISPNEIA=SIM	EVOLUCAO=OBITO	0.01188607	0.11504782
GARGANTA=NAO	EVOLUCAO=OBITO	0.01225512	0.02097029
OUTROS=NAO	EVOLUCAO=OBITO	0.01036071	0.01643620

1. Excluir as instâncias em que os pacientes foram classificados como "RECUPERADO" como desfecho.
2. Realizar o balanceamento do conjunto de dados, garantindo que haja uma proporção igual de casos de óbito e de recuperação.

Entretanto, o primeiro método apresenta a limitação de perder a métrica de confiança, uma vez que, ao restringir as análises apenas a uma classe resultante, a confiança será sempre máxima (1.0). Na Tabela 2, são apresentadas as regras geradas após o refiltro do conjunto de dados, com um único sintoma no lado esquerdo e um suporte mínimo de 0.5, com o intuito de reduzir o número de regras geradas e focar na progressão do paciente.

Tabela 2. Principais regras de associação geradas omitindo os casos em que a evolução é "OBITO"

lhs	rhs	support
GARGANTA=NÃO	EVOLUCAO=OBITO	0.858
DISPNEIA=SIM	EVOLUCAO=OBITO	0.831
OUTRO=NÃO	EVOLUCAO=OBITO	0.725
FEBRE=SIM	EVOLUCAO=OBITO	0.590
TOSSE=NÃO	EVOLUCAO=OBITO	0.520

Pode-se observar nas regras geradas que, por alguma razão, 85% das pessoas que faleceram não tinham dor de garganta, 83% tinham dispneia (falta de ar), etc[4]. A Tabela 3 exibe a quantidade de regras geradas com cada nível de suporte testado. No entanto, desta vez, o conjunto de dados "balanceado"[5] e a confiança padrão da função Apriori, 0.8, foram utilizados.

Tabela 3. Número de regras geradas usando o conjunto de dados balanceado, diferentes suportes e confiança $\geq$ 0.8

Suporte	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Nº de regras	0	0	0	0	0	0	1	3	7	27

Ao considerarmos as regras utilizando um suporte mínimo de 0.2 para os casos de óbito por COVID-19 e as ordenarmos por confiança, o conjunto principal de regras identificado foi: "dispneia (falta de ar), sem dor de garganta e sem outros sintomas não listados", com uma confiança de 94%. Todas as sete regras geradas com essas configurações podem ser visualizadas na Tabela 4.

[4]Mais regras geradas estão disponíveis em *github.com/gvheisler/KDD-COVID-19*

[5]O balanceamento foi feito de maneira que sejam usados todos os casos nos quais a evolução é óbito, e o mesmo número de casos que a evolução é recuperado (os quais são definidos aleatoriamente)

Tabela 4. Regras geradas usando o conjunto de dados balanceado, com suporte de 0.2 e confiança de 0.8

lhs	rhs	suporte	confiança
GARGANTA=NÃO, DISPNEIA=SIM, OUTRO=NÃO	ÓBITO	0.265	0.944
GARGANTA=NÃO, DISPNEIA=SIM	ÓBITO	0.355	0.935
DISPNEIA=SIM, OUTRO=NÃO	ÓBITO	0.309	0.916
TOSSE=NÃO, DISPNEIA=SIM	ÓBITO	0.215	0.907
DISPNEIA=SIM	ÓBITO	0.415	0.903
TOSSE=SIM, DISPNEIA=SIM	ÓBITO	0.200	0.898
FEBRE=SIM, DISPNEIA=SIM	ÓBITO	0.252	0.887

5.2. Classificação

Há uma variedade de métodos empregados para a classificação de dados [Phyu 2009]. Um método que se destaca por sua facilidade de uso, replicabilidade e interpretabilidade visual são as árvores de decisão [Apté and Weiss 1997]. Neste estudo, as árvores de decisão foram adotadas para classificar os sintomas que podem indicar a gravidade das infecções. Essa escolha foi motivada principalmente pela facilidade de visualização que as árvores proporcionam.

Durante a classificação dos dados, foi enfrentada uma situação semelhante à tarefa anterior, que envolveu a identificação de regras de associação representativas. Houve um desequilíbrio significativo entre o número de casos resultantes na recuperação da pessoa e o número de casos resultantes no óbito da pessoa. Por esse motivo, para utilizar o algoritmo "Recursive Partitioning and Regression Trees" (RPART) [Therneau et al. 1997], o conjunto de dados foi balanceado novamente, da mesma maneira que o anterior, o que resultou em um conjunto com exatamente 50% de cada classe (OBITO, RECUPERADO). Foram selecionadas apenas as colunas de sintomas e de evolução do paciente, as quais são necessárias para a construção da árvore de decisão.

Após essas etapas, uma árvore de decisão foi gerada utilizando uma função do pacote *rpart* [Therneau et al. 2015] e visualizada com a função *rpart.plot* [Milborrow and Milborrow 2019]. A coluna EVOLUCAO é utilizada como classe-alvo. No entanto, a árvore final não revela muitos padrões interessantes nos dados, como pode ser observado na Figura 5.

Esse fenômeno ocorre devido à importância crítica da falta de ar como um sintoma determinante para a recuperação do paciente da doença, o que também é revelado nos resultados obtidos utilizando o algoritmo Apriori. Como resultado, a árvore de decisão depende fortemente desse sintoma específico. Para melhorar a visualização, a coluna "DISPNEIA"foi omitida e ajustes foram feitos no parâmetro de complexidade da árvore. A árvore resultante é representada na Figura 6.

A árvore atualizada oferece uma compreensão mais abrangente da relação entre os sintomas e a mortalidade, pois incorpora sintomas além da falta de ar. Esta árvore expandida revela *insights* intrigantes sobre a COVID-19, como casos em que pacientes que apenas tiveram dor de garganta foram capazes de se recuperar, em maioria significativa.

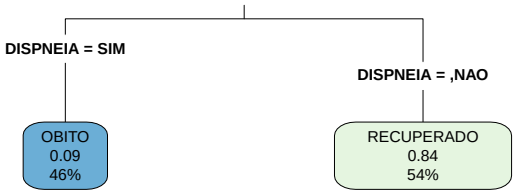


Figura 5. Árvore de decisão usando dados balanceados

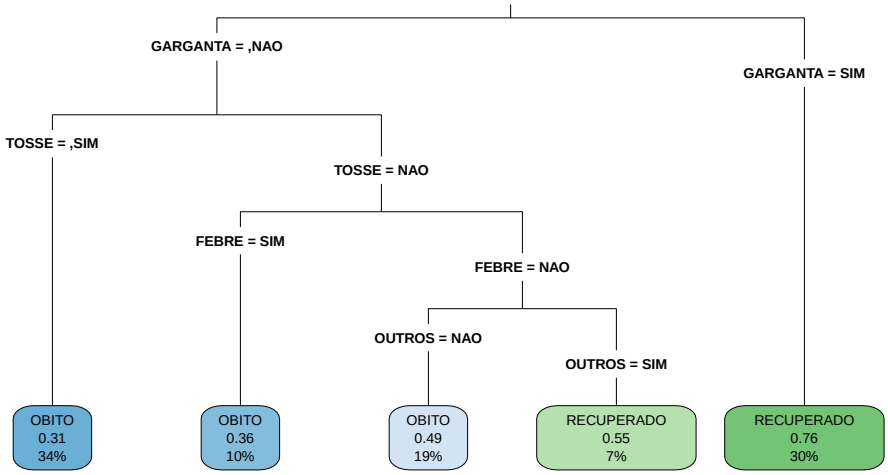


Figura 6. Árvore de decisão usando dados balanceados sem o atributo "DISPNEIA"

6. Resultados obtidos

Este estudo identificou padrões nos dados da COVID-19 no estado do Rio Grande do Sul, especificamente focando nos sintomas predominantes em pacientes que faleceram. Estes passos iniciais foram dados para obter uma compreensão mais abrangente da pandemia dentro do estado, dada a disponibilidade limitada de informações estatísticas detalhadas. Os resultados apresentados revelaram uma correlação significativa entre certos sintomas e desfechos fatais, como por exemplo a presença da dispneia (falta de ar) no paciente. As regras de associação mostradas na tabela 4 e a árvore de decisão apresentada nas Figuras 5 e 6 fornecem uma representação visual clara dos principais sintomas associados a óbitos. Além disso, foi apresentado um exemplo do processo de Descoberta de Conhecimento em Bases de Dados (KDD) aplicado em tais casos.

Gostaríamos de ressaltar que este trabalho não tem a intenção de servir como um

guia médico ou instrumento de diagnóstico. Além disso, não buscamos fornecer respostas definitivas sobre por que alguns sintomas são mais importantes do que outros. No entanto, este trabalho revela padrões ocultos que podem ser relevantes para compreender os dados da COVID-19 no estado do Rio Grande do Sul.

7. Trabalhos futuros

Em trabalhos futuros, temos a intenção de aprofundar a análise deste conjunto de dados investigando os padrões evolutivos dos sintomas ao longo do tempo. Esta exploração envolverá o desenvolvimento de abordagens inovadoras para prever picos de COVID-19 e outras doenças infecciosas, além de avaliar a eficácia da vacinação contra a COVID-19 através da análise de metadados e do cruzamento de dados de diferentes entidades. Acreditamos que tais esforços contribuirão significativamente para uma compreensão mais completa da dinâmica das doenças e possibilitarão o desenvolvimento de estratégias mais eficazes para intervenções em saúde pública.

Referências

Agrawal, R., Mehta, M., Shafer, J. C., Srikant, R., Arning, A., and Bollinger, T. (1996). The quest data mining system. In *KDD*, volume 96, pages 244–249.

Apté, C. and Weiss, S. (1997). Data mining with decision trees and decision rules. *Future generation computer systems*, 13(2-3):197–210.

Cucinotta, D. and Vanelli, M. (2020). Who declares covid-19 a pandemic. *Acta Bio Medica: Atenei Parmensis*, 91(1):157.

Dagnino, R., Weber, E., and Panitz, L. (2020). Monitoramento do coronavírus (covid-19) nos municípios do Rio Grande do Sul.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

Hahsler, M., Grün, B., and Hornik, K. (2005). arules-a computational environment for mining association rules and frequent item sets. *Journal of statistical software*, 14:1–25.

Hallal, P. C., Horta, B. L., Barros, A. J., Dellagostin, O. A., Hartwig, F. P., Pellanda, L. C., Struchiner, C. J., Burattini, M. N., Silveira, M. F. d., Menezes, A., et al. (2020). Evolução da prevalência de infecção por covid-19 no Rio Grande do Sul, Brasil: inquéritos sorológicos seriados. *Ciência & Saúde Coletiva*, 25:2395–2401.

Milborrow, S. and Milborrow, M. S. (2019). Package 'rpart. plot'. *Plot'rpart'Models: An Enhanced Version of'plot. rpart*.

Phyu, T. N. (2009). Survey of classification techniques in data mining. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 727–731. Citeseer.

Silveira, M. F., Barros, A. J., Horta, B. L., Pellanda, L. C., Victora, G. D., Dellagostin, O. A., Struchiner, C. J., Burattini, M. N., Valim, A. R., Berlezi, E. M., et al. (2020). Population-based surveys of antibodies against sars-cov-2 in southern brazil. *Nature Medicine*, 26(8):1196–1199.

Therneau, T., Atkinson, B., Ripley, B., and Ripley, M. B. (2015). Package 'rpart'. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)*.

Therneau, T. M., Atkinson, E. J., et al. (1997). An introduction to recursive partitioning using the rpart routines. Technical report, Technical report Mayo Foundation.