Avaliação dos Fatores Determinantes da Duração da Hospitalização em Casos de Dengue no Brasil: Uma Análise Baseada em Descoberta de Subgrupos

Alexandre Henrique Martins, Ana Luísa Araújo Bastos, Fernanda Luiza Tobias, Jorge Augusto de Lima e Silva, Rodrigo Sales Nascimento, Vitor Emanuel Ferreira Vital

Departamento de Ciência da Computação – Universidade Federal de Minas Gerais - Belo Horizonte, MG – Brasil

{alexandrehm, alab, fernandalt, rodrigosalesn, vitor-ef}@ufmg.br,

jorge.lima2407@gmail.com

Abstract. *Dengue is a viral disease endemic in over 120 countries, causing approximately 400 million infections annually. In 2024, Brazil led the world ranking with more than 4.8 million confirmed cases. Analyzing the hospitalization duration of patients affected by the disease is crucial for understanding treatment costs and the burden on hospitals. This study uses subgroup discovery, an advanced data analysis technique, to identify factors influencing the hospitalization duration of dengue patients in Brazil. Utilizing data from the Unified Health System (SUS), 31,488 patients were analyzed, of which 30,815 (97.86%) were hospitalized for at least one day. Algorithms such as Beam Search, Depth First Search (DFS), and SSD++ were applied and compared in terms of result quality. The analysis revealed that factors such as geographic location, specific comorbidities, intervals between symptom onset and hospitalization, as well as demographic characteristics, significantly influence hospitalization duration. The findings provide valuable insights for optimizing hospital resources and formulating more effective public health policies.*

Resumo. *A dengue é uma doença viral endêmica em mais de 120 países, causando cerca de 400 milhões de infecções anuais. Em 2024, o Brasil liderou o ranking mundial com mais de 4,8 milhões de casos confirmados. Analisar o tempo de hospitalização dos pacientes acometidos pela doença é crucial para entender os custos de tratamento e a carga sobre os hospitais. Este estudo utiliza a descoberta de subgrupos, uma técnica avançada de análise de dados, para identificar fatores que influenciam o tempo de hospitalização de pacientes com dengue no Brasil. Utilizando dados do Sistema Único de Saúde (SUS), foram analisados 31.488 pacientes, dos quais 30.815 (97,86%) ficaram internados pelo menos um dia. Algoritmos como Beam Search, Depth First Search (DFS) e SSD++ foram aplicados e comparados na qualidade dos resultados. A análise revelou que fatores como localização geográfica, comorbidades específicas, intervalos entre sintomas e internação, além de características demográficas, influenciam significativamente a duração da hospitalização. As descobertas fornecem insights valiosos para a otinização dos recursos hospitalares e a formulação de políticas de saúde pública mais eficazes.*

1. Introdução

A dengue é uma das doenças virais transmitidas por mosquitos mais significativa globalmente (Ooi and Gubler, 2009). Segundo a Organização Mundial de Saúde (OMS), a dengue é endêmica em mais de 120 países, causando cerca de 400 milhões de

infecções anuais, principalmente nas regiões tropicais e subtropicais, como o Brasil [OMS 2024a].

Essa doença continua a ser uma das principais preocupações de saúde pública no Brasil, com surtos recorrentes e um impacto considerável sobre os sistemas de saúde (Dias et al., 2024). No ano de 2024, o país já registrou mais de 4,8 milhões de casos confirmados, ocupando a primeira posição no ranking mundial da OMS [OMS 2024a]. Nota-se, portanto, a necessidade de encontrar formas eficientes de combater e lidar com a dengue no Brasil. Entre os aspectos importantes a serem analisados para minimizar esse problema, está o tempo de hospitalização dos pacientes infectados, que pode variar amplamente e impacta diretamente não só os custos de tratamento como também a carga sobre os hospitais. Por isso, identificar e entender os fatores que influenciam o período de internação pode ajudar na alocação eficiente de recursos, no planejamento de intervenções de saúde pública e na melhoria dos cuidados aos pacientes.

Neste trabalho, propomos utilizar técnicas avançadas de análise de dados, especificamente a descoberta de subgrupos, para explorar e identificar os fatores que influenciam o tempo de hospitalização de pacientes com dengue no Brasil. A descoberta de subgrupos é uma técnica poderosa que permite identificar padrões entre diferentes propriedades e variáveis em relação a uma variável alvo e assim descobrir subgrupos da população que sejam estatisticamente mais interessantes (Herrera, 2011). No nosso contexto, a variável alvo é o tempo de internação e as outras variáveis são, por exemplo, comorbidades, idade e sexo.

Os dados utilizados neste estudo são provenientes do Sistema Único de Saúde (SUS) do Brasil, extraídos do portal Base dos Dados, que possuem um amplo conjunto de informações sobre casos de dengue, incluindo dados demográficos, clínicos e de tratamento dos pacientes.

O restante do artigo está organizado da seguinte forma: a Seção 2 revisa pesquisas anteriores sobre dengue e aplicação de métodos de descoberta de subgrupos; a Seção 3 descreve os algoritmos aplicados neste trabalho; a Seção 4 detalha a coleta e o pré-processamento dos dados, assim como a metodologia de análise utilizada; a Seção 5 apresenta os principais achados, incluindo subgrupos identificados e desempenho dos algoritmos; por fim, a Seção 6 resume os principais resultados e discute suas implicações, sugerindo direções para futuras pesquisas.

2. Trabalhos Relacionados

No contexto da dengue, o estudo conduzido por Jain et al. (2017) realizou uma análise estatística univariada e multivariada para identificar associações significativas entre variáveis demográficas, clínicas e laboratoriais de pacientes de um centro de atendimento na Índia e os desfechos de interesse, incluindo mortalidade e formas graves da doença. Liew et al. (2016) também realizaram análises multivariadas para investigar associações entre variáveis sociodemográficas e clínicas com a mortalidade de pacientes doentes na Malásia. Esses estudos destacam a importância de características clínicas específicas na gestão de pacientes com dengue. No entanto, nenhum desses estudos apresentou uma análise focada no tempo de internação dos pacientes e toda a metodologia foi fundamentada em métodos estatísticos. Em contraste, o nosso estudo

busca identificar as características clínicas que possam influenciar o período de internação utilizando métodos automatizados de classificação dos pacientes em subgrupos.

Considerando o domínio médico, diversos métodos de *Subgroup Discovery* (*SD*) têm sido aplicados para encontrar subgrupos com algum valor preditivo no controle de doenças e gestão hospitalar. Vagliano et al. (2023) realizaram uma comparação dos algoritmos *Improved Subgroup Set Discovery* (*SSD++*), *Patient Rule Induction Method* (*PRIM*) e *APRIORI-SD* na identificação de subgrupos de pacientes COVID-19 internados em UTIs que apresentem padrões relevantes na dedução da probabilidade de desfechos em óbitos. Nannings et al. (2008) aplicou o método *PRIM* e a regressão logística para encontrar subgrupos de pacientes idosos em UTIs com alto risco de mortalidade.

Os resultados apresentados por essas investigações demonstram o potencial de aplicabilidade das metodologias *SD* na identificação de padrões em dados médicos, oferecendo uma base para as análises conduzidas neste estudo.

3. Algoritmos

Os algoritmos de descoberta de subgrupos, em geral, possuem três fases: geração de candidatos, poda e pós-processamento. Na primeira, cada algoritmo utiliza uma estratégia para buscar e extrair os subgrupos. Já a poda consiste em empregar um esquema que seleciona apenas candidatos significativos. Por último, no pós-processamento, os grupos são ranqueados a partir de certas medidas de qualidade (Helal, 2016).

Para identificar fatores que interferem no tempo de hospitalização de pacientes com dengue, este estudo emprega diversos algoritmos de descoberta de subgrupos. Cada algoritmo possui características diferentes. A seguir, é apresentada uma breve descrição de cada algoritmo utilizado.

3.1. Beam Search

O *Beam Search* é uma técnica de busca heurística que explora o espaço de busca de maneira restrita, mantendo um número fixo de soluções parciais de alta qualidade em cada nível de profundidade, conhecido como *beam width*. A cada iteração, os melhores candidatos (de acordo com uma função de avaliação) são expandidos para gerar novos candidatos, e os menos promissores são descartados. Essa abordagem é eficiente para problemas de otimização em espaços de busca grandes, pois limita o crescimento exponencial de possibilidades, focando nas mais promissoras (Helal, 2016).

3.2. *Depth-First Search* (*DFS*)

O *DFS* é um algoritmo clássico de busca que explora o espaço de busca de maneira "profunda", ou seja, avança por um caminho até que não seja possível continuar antes de retroceder (*backtrack*) e explorar novos caminhos (Cormen et al. 2009). É particularmente útil para problemas que requerem a exploração completa e

eficiente em termos de uso de memória, pois mantém apenas o caminho atual e os nós já visitados, mas pode não ser o mais eficiente em termos de tempo para espaços de busca vastos e complexos.

3.3. SSD++

O *SSD++* é um algoritmo heurístico desenvolvido para encontrar listas de subgrupos de alta qualidade em dados complexos, abordando o problema de encontrar listas ótimas de subgrupos, que é NP-Difícil. O *SSD++* combina a busca por feixe (*Beam Search*) com a estratégia "dividir para conquistar" para iterativamente adicionar o melhor subgrupo encontrado à lista de subgrupos. Uma vantagem é que a utilização da abordagem gulosa gera resultados interpretáveis para os usuários (Proença et al., 2022).

4. Materiais e Métodos

4.1. Dados

Esse estudo utilizou dados coletados dos pacientes admitidos entre janeiro de 2023 e janeiro de 2024, com dengue confirmada, no Brasil, extraídos do Sistema de Informação de Agravos de Notificação (SINAN) fornecidos pelo Ministério da Saúde(MS) provenientes do Sistema Único de Saúde (SUS) do Brasil.

Os microdados da dengue são obtidos através da ficha de investigação do paciente, garantindo certa padronização devido às definições e verificações rigorosas. Essa ficha contempla informações gerais do paciente, dados de residência, características da notificação individual, dados clínicos e dados laboratoriais. A variável alvo durante nossos estudos foi o intervalo de internação, sendo obtida a partir da data do início da internação à data de encerramento.

A identificação de cada ficha é feita através do número de notificação, data de notificação e município de notificação. Os dados públicos, disponibilizados pelo DATASUS, não incluem o número de notificação, por isso, não há uma chave única para essa tabela. Dessa forma, não é possível verificar o número exato de pacientes admitidos, consideramos, assim, que cada uma das linhas é um paciente.

Nesse estudo, foram considerados um total de 31.488 pacientes com dengue, dos quais 30.815 (97,86%) ficaram internados por pelo menos um dia. Foram analisados 32 atributos, selecionados e derivados das 151 colunas do banco de dados original.

Entre os pacientes analisados, 895 morreram (2,84%), um percentual que está de acordo com a taxa de mortalidade da dengue reportada pelo gov.br, considerando que a análise incluiu apenas aqueles que foram hospitalizados devido à dengue. Em média, o tempo de internação foi semelhante entre pacientes do sexo feminino e masculino, sendo 17,35 dias para mulheres e 17,62 dias para homens.

4.2. Pacientes Incluídos

Os pacientes considerados foram aqueles que constaram positivos para dengue e foram internados devido aos seus desdobramentos. Aquelas pessoas que tiveram classificação sobre a dengue descartado, ou seja, cujos dados laboratoriais constataram

que não era dengue, não foram consideradas.

4.3. Análises

Pré Processamento - inclui o tratamento de dados ausentes e a seleção de variáveis. Para o estudo, utilizamos 32 atributos, dos quais incluem diferentes tipos de informações sobre o paciente e o local no qual foi atendido.

Valores como ‘data_obito’ foram convertidos em uma coluna ‘houve_obito’ em que, se há a data, o resultado é considerado True, caso contrário, False. Valores de data foram convertidos para intervalo, sendo três possíveis:

intervalo_internação: a variável alvo, consideramos como o intervalo entre data de internação e data de encerramento.

intervalo_sintoma_internacao: intervalo entre a data dos primeiros sintomas e a data em que o paciente foi internado.

intervalo_busca_atendimento: intervalo entre a data dos primeiros sintomas e a data em que o paciente buscou ajuda médica.

Para além disso, o banco de dados possuía algumas datas incorretas, pois não condizem com o intervalo ou com o desencadeamento lógico de um processo de atendimento de paciente, e.g. paciente sendo internado devido à dengue antes da data dos primeiros sintomas da doença. Dentro desse contexto, desconsideramos tais linhas.

No quesito intervalo, existiam também alguns muito espaçados, isto é, mais de mil dias, o que não condizia com o tempo definido para estudos (2023 - 2024). Portanto, consideramos os intervalos como sendo ≥0 e 90≤ em dias.

As outras informações foram mantidas, retirando todas as linhas com informações faltantes.

Subgrupos de Pacientes - foram obtidos com base nos três algoritmos citados anteriormente, *Beam Search*, *DFS* e *SSD++*. Por algoritmo, analisamos um subgrupo independentemente uns dos outros, além disso, um paciente pode pertencer a um ou mais subgrupos, por definição de adesão às condições. Assim, se um paciente se adequa às restrições de um subgrupo ele estará inserido.

4.4. Análise Estática

Hardware: Os experimentos foram realizados em um computador Intel Core i7-8565U, com 15.8GB de RAM disponíveis.

Software: Todos os experimentos foram realizados usando *Python* v3.12, com versões dos pacotes de softwares disponíveis publicamente.

5. Resultados

A análise do tempo de internação em relação aos atributos investigados que impactam nesse alvo foi realizada a partir da execução de experimentos de descoberta de subgrupos. Nessa abordagem, foram utilizados pseudocódigos disponíveis nesse campo científico que abordam variáveis alvo numéricas e que performam em ambientes

computacionais de desenvolvimento aberto como o *Python*.

Com isso, implementou-se no notebook *Jupyter* "tratamento_banco_aplicacao_sd_2023_2024_v2.ipynb", disponível em https://github.com/alexandrehm84/projeto_dengue_internacao, os seguintes modelos de descoberta de subgrupos e configurações:

- *Beam Search*:
  - Quantidade de Subgrupos: 300;
  - Profundidade: 3;
  - Função de Qualidade Padrão do pacote pysubgroup.
- *Depth First Search* (*DFS*):
  - Quantidade de Subgrupos: 300;
  - Profundidade: 3;
  - Função de Qualidade Padrão do pacote pysubgroup.
- *SSD*++:
  - Função executada: SSDC;
  - Profundidade: 3 e 4;
  - Número de padrões selecionados em cada iteração: 25 e 50;
  - Número máximo de regras a minerar: 20;
  - Pontos de corte de discretização: 3 e 5.

O *Beam Search* e o *DFS* foram executados apenas uma vez com as configurações apresentadas. Esses algoritmos foram escolhidos para implementação devido à facilidade de implementação, por carregarem menos a memória do equipamento e por convergirem rapidamente para a formação de subgrupos.

Apesar disso, o *Beam Search* por gerar subótimos locais, já que ele limita a busca a um subconjunto de nós em cada nível, pode não encontrar a solução ótima se o feixe não incluir a melhor trajetória. Além disso, por manter múltiplos subgrupos em cada nível, ele pode definir caminhos de busca similares, o que aumenta a chance de formação de grupos redundantes na busca.

Já o *DFS*, por explorar possíveis subgrupos candidatos de forma exaustiva, pode gerar múltiplos grupos que podem tornar-se também redundantes. Com isso, tem-se também que esse código não garante que seja encontrado o caminho mais curto para a formação de subgrupos válidos.

Por fim, em contrapartida a esses modelos, temos o algoritmo *SSD*++, que é uma ferramenta aprimorada para descoberta de subgrupos proposta por Proença *et. al.* (2020). Esse algoritmo foi projetado para ser mais eficiente e escalável do que as técnicas tradicionais de *SD*, pois utiliza técnicas de otimização que permitem a descoberta de subgrupos em grandes conjuntos de dados. Isso permite a redução de redundâncias. Tem-se também a implementação de parâmetros de qualidade baseados em relevância estatística e utilidade de subgrupos.

Para o *SSD*++, foram propostos 06 (seis) modelos neste experimento cuja variável alvo é o tempo de internação de pacientes com dengue, conforme a descrição da tabela 1:

Tabela 1. Modelos implementados de descoberta de subgrupos com o *SSD++*

Modelo	Profundidade máxima	Número de padrões	Número máximo de regras	Número de pontos de corte (discretização)
1	3	25	20	3
2	4	25	20	3
3	3	25	20	5
4	4	25	20	5
5	3	50	20	3
6	4	50	20	3

Assim, serão apresentadas as avaliações de desempenho do processamento dos experimentos, a qualidade dos subgrupos por meio do |*z-score*|.

5.1. Desempenho do processamento

Com os parâmetros definidos em cada experimento, observou-se que os mais eficientes são o *Beam Search* e o *DFS*, como esperado. No processamento do banco de dados de internações por Dengue, esses algoritmos construíram 60 subgrupos válidos em 26 e 1 segundo, respectivamente.

Em contrapartida, as aplicações do *SSD++* foram significativamente mais demoradas para realizar a busca de subgrupos válidos do que os algoritmos anteriores. Nessa prática, verificou-se que o modelo 1 do *SSD++* convergiu para os resultados válidos em 14 minutos, enquanto o modelo 6, como mais peso número de padrões de busca e na profundidade, apresentou as soluções válidas em 44 minutos.

No quesito quantidade de iterações realizadas, no *SSD++*, o modelo 6 proporcionou uma maior volume de subgrupos candidatos (64) em um tempo relativamente curto. Verificou-se também que o modelo 1 gerou o menor número de iterações (42) e o modelo 4 o segundo maior número de subgrupos (62 candidatos em 27 minutos). Tais observações podem ser constatadas na tabela 6.

Tabela 2. Perfomance dos modelos de descoberta de subgrupos

Modelo	Iterações	Tempo
Modelo SSD++ 1	42	14 minutos
Modelo SSD++ 2	59	20 minutos
Modelo SSD++ 3	49	15 minutos
Modelo SSD++ 4	62	27 minutos
Modelo SSD++ 5	47	22 minutos

Modelo	Iterações	Tempo
Modelo SSD++ 6	64	44 minutos
Beam Search	60	26 segundo
DFS	60	1 segundo

5.2. A qualidade dos subgrupos

A medida de qualidade utilizada foi o |*z-score*| e a escolha considerou a relevância da métrica na literatura e sua priorização de subgrupos maiores [Meeng et al., 2021]. Identificar padrões que contemplam um maior número de pacientes parece mais relevante no âmbito de criação de políticas públicas ou mesmo administrativo.

A qualidade de cada subgrupo pode ser medida a partir da fórmula abaixo, onde n é o número de pacientes que satisfazem as condições descritas, μ_S e μ_D são os tempos médios de internação no subgrupo e na população, respectivamente, e σ_D o desvio padrão na população: $|(\sqrt{n} \cdot (\mu_S - \mu_D))/\sigma_D|$

O valor absoluto permite a identificação de subgrupos com desvios superiores e inferiores na distribuição alvo.

5.3. Top 5 grupos por modelo

A fim de caracterizar os subgrupos encontrados por cada modelo e comparar os resultados de acordo com o método utilizado, os cinco subgrupos com maiores valores de |z-score| foram selecionados. Em termos populacionais, o tempo médio de internação foi de 17,5 dias, com um desvio padrão de 17,4 dias.

Para o SSD++ apresentaremos apenas os resultados obtidos pelos modelos 1, 4 e 6, dado que esses são os que consideramos mais interessantes e os outros possuem subgrupos bastante similares.

A tabela 3 exibe um breve resumo sobre os cinco subgrupos referentes ao modelo *SSD++* 1. Embora a quantidade de pacientes que se enquadram em cada subgrupo não seja desprezível, o menor suporte equivale a 96, trata-se de uma fração muito pequena da população. Todas as descrições são compostas por três atributos. O fator localização, retratado pela sigla da Unidade Federativa (UF) de notificação, aparece em todos os conjuntos. O primeiro, terceiro e quarto subgrupos apresentam médias elevadas em relação à população, ao passo que o segundo e quinto subgrupos têm médias inferiores. Em algumas das descrições aparecem regras que condicionam um dos atributos, raça/cor ou escolaridade, a um valor irrelevante do ponto de vista prático, na base rotulado como “Ignorado”. Outros fatores considerados na caracterização dos subgrupos são comorbidades (artrite), sintomas (cefaleia), óbito, intervalo entre primeiros sintomas e internação, intervalo busca por atendimento, classificação final do diagnóstico sobre dengue e gestação.

Tabela 3. Top-5 subgrupos usando o modelo SSD++ 1

Descrição	Suporte	Média (internação)	Desvio Padrão (internação)	|z-score|
sigla_uf_notificacao = SE AND raca_cor_paciente = Ignorado AND apresenta_artrite = Não	134	48,9	19,6	20,9
sigla_uf_notificacao = GO AND escolaridade_paciente = Ignorado AND apresenta_cefaleia = Sim	704	8,5	8,7	13,6
sigla_uf_notificacao = MG AND houve_obito = sim AND 2 <= intervalo_sintoma_internacao <= 5	96	39,4	25,0	12,4
intervalo_busca_atendimento >= 7 AND sigla_uf_notificacao = MG AND raca_cor_paciente = Branca	603	26,0	20,1	12,1
classificacao_final = Dengue com Sinais de Alarme AND sigla_uf_notificacao = SC AND gestante_paciente = Não	696	9,9	9,8	11,5

Os top 5 subgrupos encontrados com o modelo *SSD*++ 4 estão apresentados na Tabela 4. Os atributos destacados na descrição são similares aos das demais aplicações, destacando o fator de localização, mas os subgrupos passam a possuir mais características. Todos os subgrupos têm média de intervalo de internação superior ao valor populacional. Alguns subgrupos trazem pouca informação ou informações confusas, como o 3, por exemplo, que apesar de possuir um tempo médio de internação longo, tem como características, além do estado de Sergipe, a ausência dos sintomas: vômito, petéquias e dor retro orbital.

Tabela 4. Top-5 subgrupos usando Modelo SSD++ 4

Descrição	Suporte	Média (internação)	Desvio Padrão (internação)	|z-score|
sigla_uf_notificacao = SE AND apresenta_artrite = Não AND raca_cor_paciente = Ignorado AND 1 <= intervalo_sintoma_internacao <= 4	80	51,8	17,3	17,6
sigla_uf_notificacao = MG AND houve_obito = sim AND intervalo_sintoma_internacao <= 5 AND possui_doenca_autoimune = Não	122	38,7	25,0	13,5
sigla_uf_notificacao = SE AND apresenta_vomito = Não AND apresenta_petequias = Não AND	80	43,3	21,5	13,3

apresenta_dor_retroorbital = Não				
intervalo_busca_atendimento >= 8 AND sigla_uf_notificacao = PR AND intervalo_sintoma_internacao <= 5 AND escolaridade_paciente = Ignorado	108	39,0	25,3	12,9
intervalo_busca_atendimento >= 8 AND sigla_uf_notificacao = MG AND idade_paciente >= 11 AND apresenta_exantema = Não	627	25,3	19,8	11,3

A tabela 5 apresenta os cinco subgrupos com maiores |*z-score*| resultantes da aplicação do modelo SSD++ 6. Nota-se que os subgrupos 1 e 2 são quase iguais, sendo a característica intervalo_sintoma_internacao <= 3 contida no primeiro, a única diferença.

Tabela 5. Top-5 subgrupos usando o modelo SSD++ 6

Descrição	Suporte	Média (internação)	Desvio Padrão (internação)	|z-score|
sigla_uf_notificacao = SE AND raca_cor_paciente = Ignorado AND apresenta_artrite = Não AND intervalo_sintoma_internacao <= 3	59	52,3	17,7	15,4
apresenta_artrite = Não AND sigla_uf_notificacao = SE AND raca_cor_paciente = Ignorado	75	46,2	20,4	14,3
sigla_uf_notificacao = MG AND intervalo_busca_atendimento >= 7 AND raca_cor_paciente = Branca AND apresenta_leucopenia = Não	493	27,2	20,5	12,4
sigla_uf_notificacao = MG AND houve_obito = sim AND idade_paciente >= 59 AND 2 <= intervalo_sintoma_internacao <= 5	63	44,0	23,8	12,1
raca_cor_paciente = Parda AND sigla_uf_notificacao = MG AND classificacao_final = Dengue AND idade_paciente >= 16	1194	22,9	20,6	10,7

Finalmente, as tabelas 6 e 7 exibem os top 5 subgrupos encontrados com *Beam Search* e *DFS*. O suporte nesses dois cenários é muito superior ao dos modelos *SSD++*. Isso ocorre em razão das regras extremamente simples. No caso do *DFS*, por exemplo, nenhuma das descrições envolve mais que dois atributos. Em todos os subgrupos, a UF de notificação equivale a MG. Nas duas aplicações, um dos subgrupos encontrados é caracterizado exclusivamente por esse atributo. As métricas incluídas nas tabelas são bastante similares entre si. Em suma, o tempo médio desses subgrupos é mais próximo

da média populacional que nos subgrupos discutidos anteriormente.

Tabela 9. Top-5 subgrupos usando *Beam Search*

Descrição	Suporte	Média (internação)	Desvio Padrão (internação)	|z-score|
ano = 2023 AND sigla_uf_notificacao = MG	5328	21,1	20,1	15,4
sigla_uf_notificacao = MG	5412	21,1	20,1	15,3
ano = 2023 AND possui_hepatopatias = Não AND sigla_uf_notificacao = MG	5253	21,1	20,1	15,3
ano = 2023 AND possui_doenca_acido_peptica = Não AND sigla_uf_notificacao = MG	5303	21,1	20,0	15,2
ano = 2023 AND possui_doenca_autoimune = Não AND sigla_uf_notificacao = MG	5237	21,1	20,0	15,2

Tabela 10. Top-5 subgrupos usando *DFS*

Descrição	Suporte	Média (internação)	Desvio Padrão (internação)	|z-score|
classificacao_final = Dengue AND sigla_uf_notificacao = MG	4191	21,6	20,4	20,9
ano = 2023 AND sigla_uf_notificacao = MG	5328	21,1	20,1	15,4
sigla_uf_notificacao = MG	5412	21,1	20,1	15,3
possui_hepatopatias = Não AND sigla_uf_notificacao = MG	5337	21,1	20,1	15,1
possui_doenca_acido_peptica = Não AND sigla_uf_notificacao = MG	5385	21,1	20,0	15,1

As variações no intervalo de internação nos cinco subgrupos com maior |*z-score*| por modelo são evidenciadas na figura 1. Conforme comentado previamente, alguns subgrupos estão centrados em tempos de internação menores e outros em valores maiores. Além disso, algumas distribuições possuem picos mais acentuados, enquanto em outros subgrupos, há maior variabilidade no valor do alvo. As curvas geradas a partir do *Beam Search* e *DFS* são praticamente equivalentes.

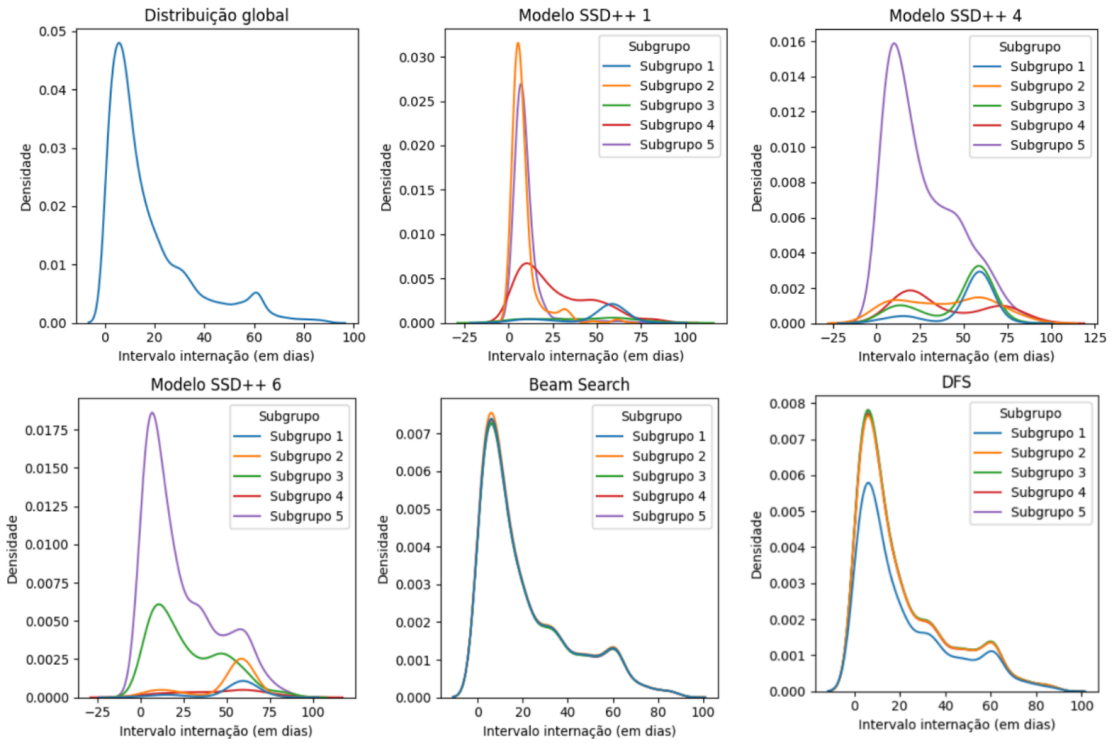

Figura 1. Distribuição tempo de internação por modelo e subgrupo

6. Conclusão e Trabalhos Futuros

Este estudo investigou os fatores determinantes da duração da hospitalização em casos de dengue no Brasil por meio da técnica de descoberta de subgrupos. A análise foi realizada com dados do Sistema Único de Saúde (SUS), aplicando algoritmos como *Beam Search*, *Depth First Search* (*DFS*) e *SSD++* para identificar padrões que influenciam o tempo de internação.

Os principais achados indicam que a técnica de descoberta de subgrupos é eficaz para revelar padrões significativos entre variáveis clínicas e demográficas dos pacientes. Os resultados mostram que fatores como comorbidades, intervalos entre os primeiros sintomas e a internação, e características sociodemográficas, como a Unidade Federativa de notificação, desempenham papéis importantes na determinação da duração da hospitalização.

Os algoritmos utilizados apresentaram desempenho variado. O *Beam Search* e o *DFS* se mostraram mais rápidos, mas com maior propensão a gerar subgrupos redundantes. Por outro lado, o *SSD++* demonstrou ser mais eficiente na criação de subgrupos relevantes, apesar de requerer maior tempo de processamento. Os subgrupos identificados pelo *SSD++* forneceram *insights* mais precisos e úteis para a gestão hospitalar e políticas de saúde pública.

Estes achados destacam a importância de considerar múltiplas variáveis na análise do tempo de hospitalização por dengue e apontam para a necessidade de

estratégias de tratamento e intervenção mais personalizadas. Futuros trabalhos podem expandir esta análise para incluir outros fatores potencialmente influentes e explorar a aplicação de outras técnicas de mineração de dados para aprimorar a qualidade dos resultados.

Além disso, recomenda-se a implementação de tais modelos em sistemas de saúde para auxiliar na previsão de tempos de internação e na alocação eficiente de recursos, contribuindo para a melhoria do atendimento aos pacientes e a gestão hospitalar no combate à dengue.

Referências

Bosc, G., Boulicaut, JF., Raïssi, C. et al. Anytime discovery of a diverse set of patterns with Monte Carlo tree search. Data Min Knowl Disc 32, 604–650 (2018). https://doi.org/10.1007/s10618-017-0547-5.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Introduction to Algorithms (3rd Edition). MIT Press.

Dias, R. I. R., Oliveira, T. d. S., Farias, B. R. D., Diniz, M. d. L. P., Oliveira, A. G. d. S. C., Carvalho, K. A. d. O., Araújo, N. H. d. F., Costa, V. M., Costa, A. D., Santos, F. M. C. S., Cavalcanti, B. B., and Neto, J. M. d. S. (2024). Impacto nas medidas de prevenção e promoção da saúde na epidemiologia da dengue no brasil: Uma revisão sistemática. *Brazilian Journal of Implantology and Health Sciences*, 6(3):1069–1078.

Helal, S. Subgroup Discovery Algorithms: A Survey and Empirical Evaluation. J. Comput. Sci. Technol. 31, 561–576 (2016). https://doi.org/10.1007/s11390-016-1647-1.

Herrera, F., Carmona, C.J., González, P. et al. An overview on subgroup discovery: foundations and applications. Knowl Inf Syst 29, 495–525 (2011). https://doi.org/10.1007/s10115-010-0356-2.

Jain, S., Mittal, A., Sharma, S. K., Upadhyay, A. D., Panday, R. M. et al. (2017). Predictors of Dengue-Related Mortality and Disease Severity in a Tertiary Care Center in North India. OFID - Open Forum Infectious Diseases. https://doi.org/10.1093/ofid/ofx056.

Liew S. M., Khoo E. M., Ho B. K., Lee Y. K., Omar M., Ayadurai V., et al. (2016) Dengue in Malaysia: Factors Associated with Dengue Mortality from a National

Registry. PLoS ONE 11(6): e0157631. https://doi.org/10.1371/journal.pone.0157631.

Meeng, M., & Knobbe, A. (2021). For real: a thorough look at numeric attributes in subgroup discovery. Data Mining and Knowledge Discovery, 35(1), 158-212. https://doi.org/10.1007/s10618-020-00703-x.

Nannings, B., Abu-Hanna, A., Jonge, E. (2008). Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. International Journal of Medical Informatics 77, 272–279. https://doi.org/10.1016/j.ijmedinf.2007.06.007.

OMS (2024a). Dengue and severe dengue. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue. online, acesso em 24 de julho de 2024.

OMS (2024b). Global dengue surveillance. https://worldhealthorg.shinyapps.io/dengue_global/. online, acesso em 24 de julho de 2024.

Ooi, E.-E. and Gubler, D. J. (2009). Global spread of epidemic dengue: The influence of environmental change. Future Virology, 4(6):571–580.

Proença, H.M., Grünwald, P., Bäck, T. et al. Robust subgroup discovery. Data Min Knowl Disc 36, 1885–1970 (2022). https://doi.org/10.1007/s10618-022-00856-x.

Vagliano, I., Kingma, M. Y., Dongelmans, D. A., Lange, D. W., Keizer, N. F, Schut, M. C. (2023). Automated identification of patient subgroups: A case-study on mortality of COVID-19 patients admitted to the ICU. Computers in Biology and Medicine 163. https://doi.org/10.1016/j.compbiomed.2023.107146.