1.2.2 Métricas e ferramentas para avaliação de desempenho dos modelos de ML

Os modelos de ML carecem da necessidade de avaliação do seu desempenho, seja para fins de decisão sobre a sua adequação ao problema dado seja para fins de comparação de resultados alcançados. Diversas métricas e ferramentas para avaliação de modelos de ML estão disponíveis na literatura e as utilizadas nesta pesquisa estão apresentadas a seguir (Halkidi, Batistakis e Vazirgiannis, 2001, p. 107–145).

A avaliação do desempenho de algoritmos de agrupamento se diferencia substancialmente da avaliação de algoritmos de classificação supervisionada. Ao contrário da contagem direta de erros ou da análise de precisão e recuperação em cenários supervisionados, a avaliação de algoritmos de agrupamento exige uma abordagem mais profunda e abrangente. A ausência de rótulos, impedindo a comparação direta com resultados reais, a subjetividade na interpretação de qual a melhor separação dos dados e a escolha das métricas adequadas às particularidades do problema em questão e das características dos dados são desafios que podem ser citados. A combinação de diferentes técnicas de avaliação pode fornecer uma visão mais completa e robusta do desempenho do algoritmo de agrupamento.

Dado que os modelos de agrupamento utilizados na presente pesquisa são não supervisionados, as métricas disponíveis na biblioteca SKLearn listadas a seguir foram utilizadas.

O coeficiente de silhueta ou *silhouette_score* da classe *sklearn.metrics*, considera tanto a coesão interna dos *clusters* (similaridade entre pontos dentro do mesmo *cluster*) quanto a separação entre clusters diferentes. O cálculo do coeficiente é dado pela equação (7).

$$s = \frac{b-a}{max\ (a,b)} \tag{7}$$

Onde:

s: valor que representa o *silhouette_score*, definido no intervalo [-1;1];

a: a distância média entre uma amostra e todos os outros pontos da mesma classe;

b: a distância média entre uma amostra e todos os outros pontos no próximo *cluster* mais próximo.

Os valores são limitados ao intervalo [-1;1], onde valores próximos a 1 indicam *clusters* bem definidos, com alta coesão interna e boa separação entre *clusters*, valores próximos a -1 indicam agrupamentos ruins, com baixa coesão interna ou separação insuficiente entre *clusters* ou incorretos. Valores em torno de zero apontam clusters sobrepostos (Ousseuw, P. J., 1987, p. 53-65).

O índice Calinski-Harabasz (Caliński e Harabasz, 1974, p. 1-27), também conhecido como Critério de Razão de Variância (VCR), implementado na SKLearn pelo método *calinski_harabasz_score* da classe *sklearn.metrics*, foi outro índice utilizado para avaliar o modelo. O índice é a razão entre a soma da dispersão entre clusters e da dispersão dentro do cluster para todos os clusters (onde a dispersão é definida como a soma das distâncias ao quadrado) e é calculado pelas equações (8a), (8b) e (8c).

$$\mathrm{s} = \frac{tr(B_k)}{\mathrm{tr}(W_k)} * \frac{n_E - k}{k-1} \tag{8a}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{8b}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \tag{8c}$$

Onde:

$tr(B_k)$: Representa o traço da matriz de dispersão entre os grupos (indica a dispersão geral dos pontos de dados em torno da média);

$tr(W_k)$: Representa o traço da matriz de dispersão dentro dos grupos (indica a dispersã**o** geral dos pontos de dados em torno da média);

C_q: Representa o conjunto de pontos do *cluster q*;

c_q: Representa o centro do *cluster q*;

c_E: Representa o centro do conjunto de dados *E*;

n_q: Representa o conjunto de pontos no *cluster q*.

Uma pontuação Calinski-Harabasz mais alta se refere a um modelo com clusters mais bem definidos.

O índice Davies-Bouldin (Davies e Bouldin, 1979, p. 224-227), disponibilizado na SKLearn pelo método *davies_bouldin_score* da classe *sklearn.metrics*, foi utilizado para medir a similaridade média entre os *clusters*, comparando a distância entre *clusters* com o tamanho do próprio *cluster*.

O índice é definido como a similaridade média entre cada cluster *Ci* para *i = 1, 2, ..., k* e seu *cluster* mais similar *Cj*. As equações (9a) e (9b) definem o cálculo do índice, denominado nas próximas seções como similaridade.

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} R_{ij} \qquad (9a)$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \qquad (9b)$$

Onde:

R_{ij}: Representa a medida que equilibra s_i e d_{ij};

s_i: Representa a distância média em cada ponto do *cluster i* e o centroide do *cluster* (diâmetro do *cluster*);

d_{ij}: Representa a distância entre os centroides dos *clusters i* e *j*;

k: Representa o número de *clusters*.

Valores próximos a zero, menor valor possível, indicam uma melhor separação entre os *clusters*.

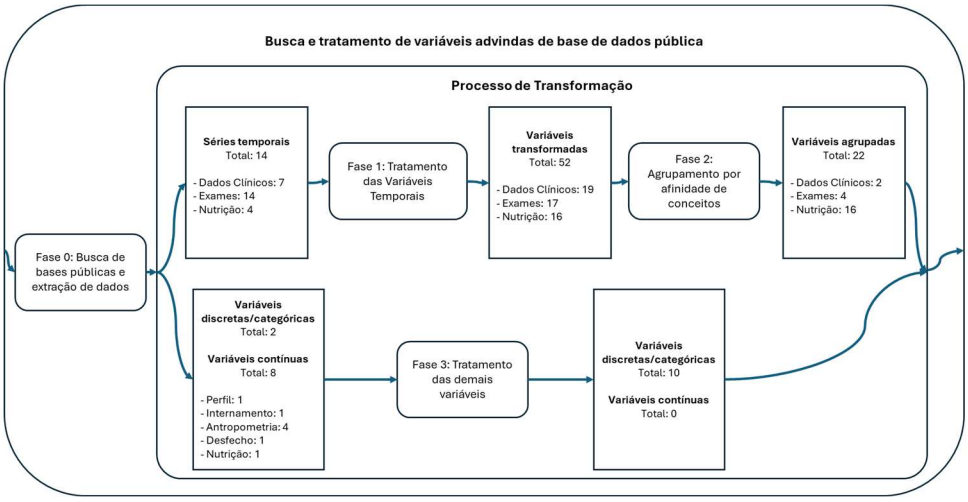
O índice Davies-Boulding é geralmente mais alto para *clusters* convexos do que outros conceitos de *clusters* e o uso da distância centroide limita a métrica de distância ao espaço euclidiano (Davies e Bouldin, 1979, p. 224-227).

========== METODOLOGIA ==========

O conjunto de variáveis selecionado na Etapa 1 compreende: variáveis numéricas contínuas com alta variação de valores (*peso, altura, idade, IMC*); e, séries temporais, sendo ideal estudar o comportamento das variáveis ao longo de todo o período de internação, em lugar

de analisar os seus valores diários, de modo isolado. Desafios como a falta de uniformidade no tempo de internamento dos pacientes e a ausência de valores em variáveis cuja coleta não é necessariamente diária (ex.: exames) ou onde não houve inserção de dados, surgiram e foram solucionados aplicando um conjunto de regras de transformação no conjunto de variáveis. A Figura 13 apresenta o passo a passo dessa transformação, que procurou zelar pela melhoria da performance dos modelos, tanto em relação à precisão quanto à confiabilidade, pela redução no tempo de treinamento e pela facilitação da comunicação e colaboração da equipe, especialmente dos especialistas. As fases foram numeradas apenas para fins de identificação e não representam, necessariamente, a ordem de execução ou representam prioridade.

Figura 1. Fluxo de aplicação das regras de transformação aplicadas na Etapa 2



Fonte: Própria autoria.

A Fase 1 consistiu em traduzir as séries temporais em um conjunto de variáveis matemáticas e estatísticas que representassem o comportamento apresentado pelo paciente dentro do período de internamento, independente do tempo de internamento. Medidas como moda, média, valor máximo, valor mínimo e tendência, considerando o período de internação do paciente, em que as variáveis foram coletadas, foram utilizadas nesse processo. Para o cálculo de tendência foi utilizado o teste de tendência Mann-Kendall, um teste estatístico não paramétrico comumente utilizado para avaliar a presença de tendências monotônicas (aumento ou diminuição) em uma série temporal (Dos Santos, 2020, p. e87). Ainda no cálculo da tendência, foram desconsiderados os dados cujo conjunto de anotações era menor que três.

As transformações foram realizadas utilizando consultas SQL sobre as tabelas armazenadas no *BigQuery*, bem como utilizando a linguagem Python, na plataforma gratuita

baseada em nuvem *Google Colab* (**GOOGLE COLAB.** Disponível em: https://colab.research.google.com/. Último acesso em: 03 de junho de 2024). Para o cálculo de tendência foi utilizada a biblioteca *pymannkendall* versão 1.4.3 (Hussain e Mahmud, 2019).

A Tabela 2 apresenta a lista resumida das variáveis resultantes da Fase 1. Observe que a série temporal *Balanço hídrico diário* foi transformada em quatro novas variáveis com o objetivo de capturar os aspectos mais relevantes da série de dados inicial – balanço hídrico em 72h (*BHD72h*), variação do balanço hídrico nas 72h iniciais de internamento em UTI (*VariacaoBH72h*), tendência do balanço hídrico nas 72h iniciais de internamento em UTI (*TendenciaBH72h*) e tendência do balanço hídrico diário durante todo o período de internamento (*BHTendencia*). O mesmo ocorreu com as demais séries temporais. A identificação dos aspectos mais relevantes, bem como as regras de transformação, foi fruto de uma série de discussões com os especialistas. Para uma análise mais aprofundada das variáveis adicionadas, pode-se consultar o Apêndice A.

Tabela 1 - Lista de variáveis temporais transformadas geradas na Fase 1

(continua)

Natureza	**Série temporal original (Granularidade: um ou mais registros por dia de internação)**	**Fase 1: Variáveis transformadas (Granularidade: um registro por período inteiro de internação)**
Dados clínicos	Balanço hídrico diário	BHD72h
		VariacaoBH72h
		TendenciaBH72h
		BHTendencia
	Evacuações	Diarreia
		Constipação
	Hemodiálise	Hemodiálise
	Ventilação mecânica invasiva	VMInicio
		VMReintubacao
		VMTempoPermanencia
		VMTempodesmame
	Hemoglucoteste	ProporcaoHGTHipo
		ProporcaoHGTHiper
	Temperatura corporal	ProporcaoTempElevada
	Uso de droga vasoativa	ProporcaoDiasSemNora
		ProporcaoDiasNoraAte025
		ProporcaoDiasNora025_050
		ProporcaoDiasNora050Mais
		UsoVasopressina
Exames	Ureia	ProporcaoUreiaAlta

	Creatinina	ProporcaoCreatininaAlta
	Sódio	ProporcaoSodioHipo
		ProporcaoSodioHiper
	Potássio	ProporcaoPotassioHipo
		ProporcaoPotassioHiper
	Magnésio	ProporcaoMagnesioHipo
		ProporcaoMagnesioHiper
	Fósforo	ProporcaoFosforoHipo
	Albumina	ProporcaoAlbuminaBaixa

Tabela 2 – Lista de variáveis temporais transformadas geradas na Fase 1

(conclusão)

Natureza	**Série temporal original (Granularidade: um ou mais registros por dia de internação)**	**Fase 1: Variáveis transformadas (Granularidade: um registro por período inteiro de internação)**
Exames	Linfócitos totais	ProporcaoLinfocitosTotaisBaixo
	Hemoglobina	ProporcaoHemoglobinaBaixa
	Aspartato (TGO ou AST)	AspartatoMaximo
	Bilirrubinas	ProporcaoBilirrubinasAlta
	Alanina (TGP ou ALT)	AlaninaMaxima
	Triglicérides	ProporcaoTrigliceridesAlto
	Fosfatase alcalina	FosfataseMaxima
Nutrição	Oferta calórica diária	MediaOfertaCaloricaDiaAte7DiasInternamento
		MediaOfertaCaloricaDiaAte7DiasTN
		MediaOfertaCaloricaDiaTotal
		MediaOfertaCaloricaDiaTotalTN
	Oferta calórica diária por Kg	MediaOfertaCaloricaDiaKgAte7Dias
		MediaOfertaCaloricaDiaKgAte7DiasTN
		MediaOfertaCaloricaDiaKgTotal
		MediaOfertaCaloricaDiaTotalTN
	Oferta proteica diária	MediaOfertaProteicaDiaAte7DiasInternamento
		MediaOfertaProteicaDiaAte7DiasTN
		MediaOfertaProteicaDiaTotal
		MediaOfertaProteicaDiaTotalTN
	Oferta proteica diária por Kg	MediaOfertaProteicaDiaKgAte7Dias
		MediaOfertaProteicaDiaKgAte7DiasTN
		MediaOfertaProteicaDiaKgTotal
		MediaOfertaProteicaDiaKgTotalTN

Na Fase 2, no intuito de caracterizar melhor o estado geral dos pacientes e facilitar a validação com os especialistas na área de saúde, as variáveis da Fase 1 foram organizadas em

diferentes perspectivas por similaridade de conceito e, sobre esses, foram aplicados algoritmos de ML (aprendizagem não supervisionada – *clustering*) para propor grupos contendo similaridades, a partir dos padrões encontrados nos dados de cada perspectiva. As variáveis relacionadas a dados clínicos foram organizadas em duas perspectivas:

a) Condição clínica: *BHD72h*, *VariacaoBH72h*, *TendenciaBH72h*, *BHTendencia*, *Diarreia*, *Constipação*, *ProporcaoHGTHipo*, *ProporcaoHGTHiper*, *ProporcaoTempElevada*; e,
b) Gravidade: *Hemodialise*, *VMInicio*, *VMReintubacao*, *VMTempoPermanencia*, *VMTempoDesmame*, *ProporcaoDiasSemNora*, *ProporcaoDiasNoraAte025*, *ProporcaoDiasNora025_050*, *ProporcaoDiasNora050Mais*, *UsoVasopressina*.

As variáveis relacionadas a exames foram organizadas em quatro perspectivas:

a) Função Renal: *ProporcaoUreiaAlta*, *ProporcaoCreatininaAlta*;
b) Eletrólitos: *ProporcaoSodioHipo*, *ProporcaoSodioHiper*, *ProporcaoPotassioHipo*, *ProporcaoPotassioHiper*, *ProporcaoMagnesioHipo*, *ProporcaoMagnesioHiper*, *ProporcaoFosforoHipo*;
c) Perfil Nutricional: *ProporcaoAlbuminaBaixa*, *ProporcaoLinfocitosTotaisBaixo*, *ProporcaoHemoglobinaBaixa*;
d) Perfil Hepático: *AspartatoMaximo*, *AlaninaMaxima*, *ProporcaoTriglicerideAlto*, *FosfataseMaxima*, *ProporcaoBilirrubinasAlta*.

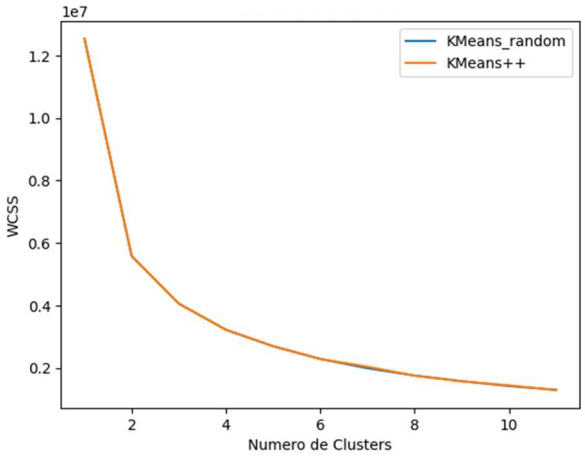
Em cada perspectiva foram aplicados os modelos de clusterização não supervisionados *k-Means* e o *Mean-Shift* disponíveis na biblioteca SKLearn (SCIKIT-LEARN, 2019). Inicialmente o modelo *Mean-Shift*, implementado como o método *MeanShift* na classe *sklearn.cluster*, foi aplicado com o parâmetro *bandwidth*, que controla a granularidade dos clusters, definido pela função *estimate_bandwidth*, da mesma biblioteca, e demais parâmetros com valores default. A escolha do *MeanShift* se deu principalmente pela capacidade de encontrar *clusters* de formas irregulares e tamanhos variados e por não requerer a predefinição do número de clusters.

Diante da dificuldade inerente à avaliação de *clusters* em cenários de aprendizagem não supervisionada, optou-se por empregar um modelo de agrupamento distinto, visando confrontar

os resultados e validar a robustez da abordagem inicial. O *k-Means* foi escolhido por ser versátil, eficiente e por sua interpretabilidade, porém fez-se necessário definir o *k* previamente.

A Curva de Elbow foi utilizada para buscar o valor de *k* utilizado como parâmetro de entrada para o *k-Means*. Para gerar a curva foi aplicado o modelo *k-Means* implementado pelo método *KMeans* da classe *sklearn.cluster* disponível na biblioteca SKLearn para diferentes valores de *k*, incluindo o valor sugerido pelo *MeanShift*. Para cada perspectiva foi plotado um gráfico contendo as curvas geradas utilizando duas formas de inicialização (parâmetro *init* com valores *'k-means++'* e *'random'*, sendo que o primeiro seleciona os centroides iniciais, de forma iterativa, usando uma amostragem baseada em uma distribuição de probabilidade empírica da contribuição dos pontos para a inércia geral e o segundo seleciona os centroides aleatoriamente). O cálculo do WCSS é disponibilizado pela biblioteca através do atributo *inertia_*. A título ilustrativo, a Figura 14 apresenta a Curva de Elbow gerada para a perspectiva Condição Clínica, apontando uma primeira quebra para dois clusters, neste caso o número de grupos sugerido pelo *Mean-Shift*, e uma suavização da curva a partir do número de grupos igual a quatro, neste caso o valor utilizado como entrada para o *k-Means*.

Figura 2. Curva de Elbow gerada para a perspectiva Condição Clínica

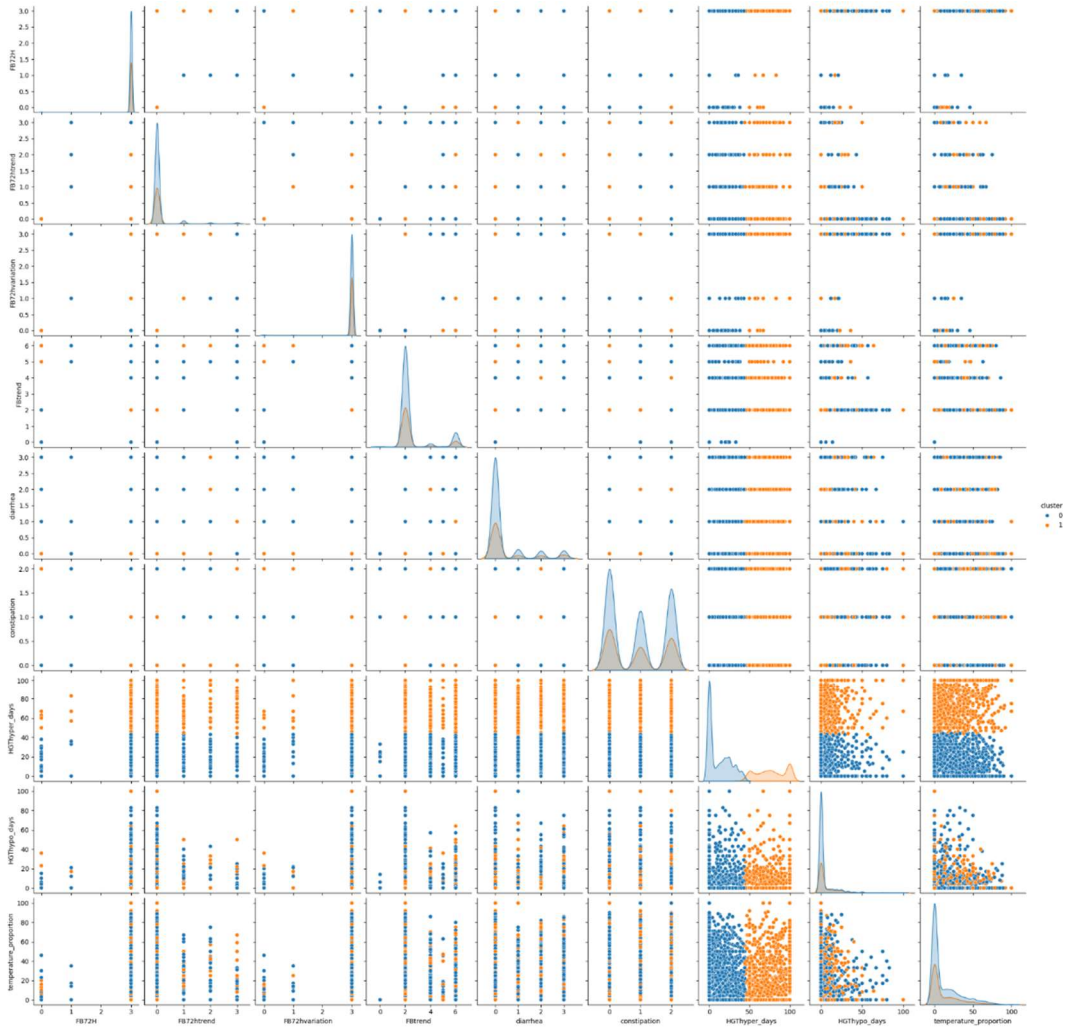


Fonte: Autoria própria.

Para cada perspectiva, os resultados produzidos por ambos os modelos foram analisados graficamente, confrontando as variáveis aos pares com os clusters gerados, utilizando o gráfico *PairGrid* da biblioteca *seaborn* (SEABORN, 2012), e em relação ao balanceamento dos *clusters* gerados. A título de exemplo, a Figura 15 apresenta o gráfico *PairGrid* gerado para a

perspectiva Condição Clínica após a aplicação do modelo *Mean-Shift* com dois grupos (plotados em diferentes cores).

Figura 3. *PairGrid* com grupos sugeridos pelo *Mean-Shift* para a perspectiva Condição Clínica.



Fonte: Autoria própria.

A implementação dos agrupamentos foi feita em *Python*, no *Google Colab*, utilizando as bibliotecas SKLearn (SCIKIT-LEARN, 2019), pandas (PANDAS, 2024) e *numpy* (Harris et

al., p. 357–362, 2020; PANDAS, 2018). Na maioria dos casos, a Curva de Elbow sugeriu um valor de *k* próximo ou igual ao sugerido pelo *Mean-Shift*, porém geralmente menor.

Os modelos não supervisionados foram aplicados em cada uma das perspectivas e os resultados obtidos foram analisados considerando as métricas: inércia (extraído apenas para o *k-Means* com diferentes modos de inicialização – valores *"k-means++"* e *"random"* utilizados no parâmetro *init* – utilizado para fins de desempate), silhueta (*silhouette_score*), VCR (*Calinski and Harabasz score*) e similaridade (*Davies-Bouldin score*). A quantidade de grupos, o balanceamento dos grupos gerados e a análise da distribuição dos *clusters*, considerando pares de variáveis, também foram considerados para a decisão. Para o subconjunto de variáveis da perspectiva Condição Clínica, o domínio de valores para as variáveis *BHD72h* e *VariacaoBH72h*, originalmente definido como (-1, 0, 1, 99) foi ajustado para (1, 2, 3, 0) buscando uma melhor performance dos modelos. O desempenho dos modelos utilizados nesta etapa do trabalho, com respectivos parâmetros utilizados, é apresentado e discutido no capítulo Resultados preliminares.

Para auxiliar na explicabilidade dos grupos gerados, fez-se uso de um modelo classificatório (árvores de decisão - aprendizado supervisionado), onde as árvores foram geradas a partir das variáveis consideradas e dos rótulos dados aos diferentes grupos pelos algoritmos de agrupamento aplicados. Nesse caso, foi possível apresentar um conjunto de regras para a formação de cada grupo (a partir das árvores de decisão geradas), possibilitando explicar o comportamento de cada diferente grupo, estratégia utilizada no trabalho de Moreira e Namen (2018).

Para implementação das árvores de decisão foi utilizada a linguagem *Python*, utilizando a biblioteca SKLearn, separando os dados de treinamento e teste numa proporção de 70-30, aplicando-se o método *train_test_split* com o parâmetro *stratify=y* para garantir que as proporções de classes fossem preservadas nos conjuntos de treino e teste para os casos de distribuições de classes desbalanceadas. Para fins de facilitação da discussão com os especialistas, as árvores foram podadas em até 6 níveis, quando necessário, porém priorizando-se garantir um bom desempenho do modelo, considerando-se as métricas: precisão, revocação e *f1-score*. As árvores foram geradas pelo método *DecisionTreeClassifier* utilizando o critério de *Gini* para determinar as quebras. O parâmetro *random_state=42* foi utilizado aqui. A coluna *cluster* foi definida como variável alvo.

==========

1. RESULTADOS PRELIMINARES

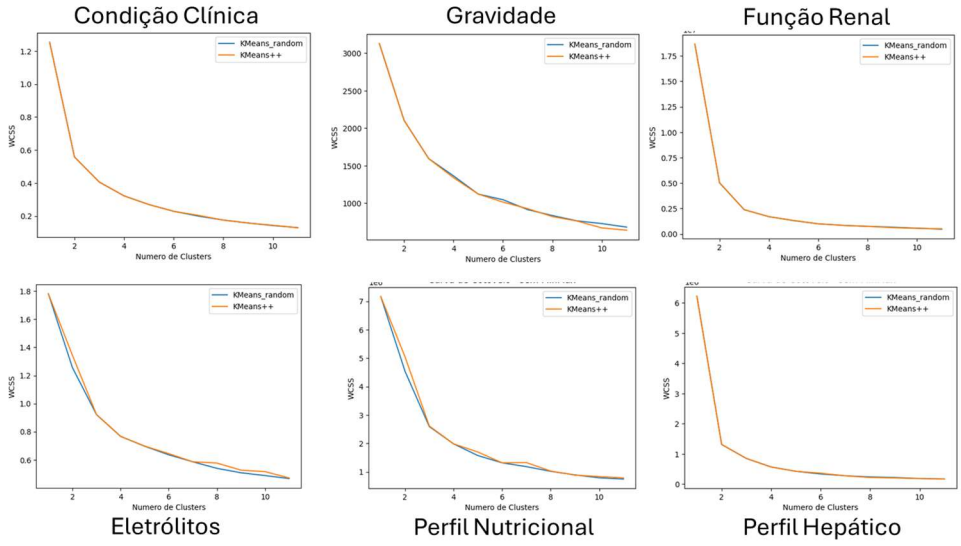
O presente capítulo apresenta os resultados preliminares obtidos até o momento. A seção inicial discute os resultados provenientes da aplicação de modelos não supervisionados, enquanto a seção subsequente aborda os resultados dos modelos classificatórios.

3.1 Discussão de resultados dos modelos não supervisionados utilizados para geração dos grupos para as diferentes perspectivas

Esta seção apresenta os resultados obtidos até o presente momento na Etapa 2. Conforme descrito na metodologia discutida no Capítulo 2, após a transformação das variáveis temporais em variáveis estatísticas – Fase 1 da Etapa 2 – e posterior organização das variáveis resultantes por afinidade de conceito – Fase 2 da Etapa 2 – os modelos de agrupamento *MeanShift* e *k-Means* foram aplicados sobre os dados de cada uma das perspectivas. O modelo *MeanShift* foi aplicado com o parâmetro *bandwidth* definido pelo método *estimate_bandwidth*, executado com o parâmetro *quantile=0,3* e *random_state=42*.

A definição do *k* inicial para aplicação do *k-Means* foi definido a partir da análise da Curva de Elbow considerando valores de *k* no intervalo [2:12]. A Figura 18 apresenta as curvas geradas para cada perspectiva.

Figura 4. Curvas de Elbow geradas para cada perspectiva



Fonte: Autoria própria.

Os resultados obtidos com a aplicação dos modelos em cada perspectiva com os respectivos parâmetros utilizados estão apresentados na Tabela 4, onde a coluna “Quantidade de grupos” apresenta o valor sugerido pelo modelo de agrupamento nas linhas que apresentam os resultados obtidos com o *MeanShift* e o k, sugerido pela Curva de Elbow, nas linhas que apresentam os resultados obtidos com o *k-Means*.

Tabela 2 – Resultados dos modelos não supervisionados para agrupamento

(continua)

Perspectiva	**Modelo**	**Tempo de execução (em segundos)**	**Inércia (menor melhor)**	**Silhueta (maior melhor)**	**VCR (maior melhor)**	**Similaridade (menor melhor)**	**Quantidade de grupos**	**Balanceamento dos grupos (em quantidade de elementos por grupo)**	**Considerações sobre a análise visual do gráfico de pares de variáveis**
Condição clínica	***MeanShift***	**35,002**	**-**	**0,527**	**9336.018**	**0,759**	**2**	**(5119, 2374)**	**Grupos bem definidos considerando os pares envolvendo a variável *HGTHiper***
	k-Means (init="k-means++", n_init=8)	1,649	3219486	0,417	7224,302	0,870	4	(1785, 3029, 1702, 977)	Grupos bem definidos considerando os pares envolvendo a variável *HGTHiper*
	k-Means (init="random", n_init=8)	0,142	3219549	0,417	7224,109	0,869	4	(1701, 1009, 2997, 1786)	Grupos bem definidos considerando os pares envolvendo a variável *HGTHiper*
Gravidade	*MeanShift**	37,984	-	0,405	248,846	1,044	6	(7080, 302, 72, 15, 5, 19)	Grupos mal definidos e com muitas sobreposições

Tabela 4 – Resultados dos modelos não supervisionados para agrupamento

(continua)

Perspectiva	Modelo	Tempo de execução (em segundos)	Inércia (menor melhor)	Silhueta (maior melhor)	VCR (maior melhor)	Similaridade (menor melhor)	Quantidade de grupos	Balanceamento dos grupos (em quantidade de elementos por grupo)	Considerações sobre a análise visual do gráfico de pares de variáveis
	k-Means (init="k-means++", n_init=8)	**1,019**	**2447688**	**0,680**	**9027,656**	**0,943**	**4**	**(5175, 336, 1312, 670)**	**Grupos bem formados considerando os pares envolvendo as variáveis *norafreedays* e *vaso_days***
	k-Means (init="random", n_init=8)	0,159	2447688	0,680	9027,656	0,943	4	(336, 5175, 1312, 670, 330)	Grupos bem formados considerando os pares envolvendo as variáveis *norafreedays* e *vaso_days*
Função Renal	*MeanShift*	17,127	-	0,695	24238,347	0,727	6	(4508, 836, 750, 499, 498, 402)	Grupos mal definidos e com muitas sobreposições
	k-Means (init="k-means++")	**0,289**	**1699084**	**0,711**	**24909,819**	**0,725**	**4**	**(4461, 1135, 1083, 814)**	**Grupos bem formados**
	k-Means (init="random")	0,038	1699080	0,711	24909,880	0,726	4	(815, 4461, 1133, 1084)	Grupos bem formados

Tabela 4 – Resultados dos modelos não supervisionados para agrupamento

(continua)

Perspectiva	**Modelo**	**Tempo de execução (em segundos)**	**Inércia (menor melhor)**	**Silhueta (maior melhor)**	**VCR (maior melhor)**	**Similaridade (menor melhor)**	**Quantidade de grupos**	**Balanceamento dos grupos (em quantidade de elementos por grupo)**	**Considerações sobre a análise visual do gráfico de pares de variáveis**
Eletrólitos	*MeanShift*	38,364	-	0,223	179,301	1,359	10	(7014, 27, 55, 26, 119, 95, 5, 94, 20, 38)	Grupos mal definidos e com muitas sobreposições
	k-Means (init=" k-means++", n_init="auto")	**0,235**	**6975053**	**0,308**	**2909,599**	**1,207**	**5**	**(3337, 1355, 849, 1274, 678)**	**Grupos relativamente bem formados**
	k-Means (init="random", n_init="auto")	1,151	6965856	0,271	2915,748	1,250	5	(3037, 1293, 617, 973, 1573)	Grupos relativamente bem formados
Perfil Nutricional	*MeanShift*	23,959	-	0,557	2214,028	1,133	5	(6552, 580, 326, 30, 5)	Grupos relativamente bem formados
	k-Means (init=" k-means++", n_init="auto")	**0,184**	**1985597**	**0,455**	**6511,267**	**0,792**	**4**	**(3959, 1304, 449, 1781)**	**Grupos relativamente bem formados**
	k-Means (init="random", n_init="auto")	0,145	1984648	0,453	6515,572	0,796	4	(3900, 449, 1266, 1878)	Grupos relativamente bem formados

Tabela 4 – Resultados dos modelos não supervisionados para agrupamento

(conclusão)

Perspectiva	**Modelo**	**Tempo de execução (em segundos)**	**Inércia (menor melhor)**	**Silhueta (maior melhor)**	**VCR (maior melhor)**	**Similaridade (menor melhor)**	**Quantidade de grupos**	**Balanceamento dos grupos (em quantidade de elementos por grupo)**	**Considerações sobre a análise visual do gráfico de pares de variáveis**
Perfil Hepático	*MeanShift*	16,643	-	0,758	10673,058	0,537	14	(5709, 673,0507, 410, 100, 37, 25, 12, 11, 4, 2, 1, 1, 1)	Grupos mal definidos, com muitas sobreposições e grupos pequenos)
	k-Means (init=" k-means++", n_init="auto")	**0,129**	**849906**	**0,778**	**23672,588**	**0,529**	**3**	**(6032, 736, 725)**	**Grupos bem formados**
	k-Means (init="random", n_init="auto")	0,152	849906	0,778	23672,588	0,529	3	(6032, 736, 725)	Grupos bem formados

* originalmente 57 grupos foram propostos, sendo que muitos deles possuíam poucas amostras (<5). O parâmetro *quantile* foi ajustado para 0,5 e os dados de entrada pré-processados, utilizando a função *MinMaxScaler* da biblioteca *sklearn.preprocessing*, para fins de normalização dos dados visando à de redução do número de grupos.

Os modelos eleitos para gerar os grupos que seguiriam para as próximas fases da pesquisa, destacados na Tabela 4, foram os que apresentaram o melhor resultado na maioria dos critérios de análise (silhueta, VCR e similaridade). A inércia foi utilizada para fins de desempate. A quantidade de grupos sugeridos, o balanceamento dos grupos e a análise visual a partir dos gráficos gerados serviram de suporte para a decisão final, sendo a preferência por um menor número de grupos. Para a perspectiva Condição Clínica, os grupos gerados pelo *MeanShift* foram selecionados para seguir para a etapa de classificação. Para as demais perspectivas, os grupos gerados pelo *k-Means* com *init="k-means++"* foram os eleitos.

Conforme mencionado no Capítulo 2 - Metodologia, árvores de decisão foram geradas para fins de explicação das regras utilizadas na formação dos grupos. A divisão da base para treino e teste respeitou a proporção 70%-30% utilizando a função *train_test_split* da classe *model_selection* da biblioteca SKLearn com o parâmetro *random_state=42* e demais parâmetros com valor padrão. A coluna *cluster* foi definida como variável alvo. Árvores geradas com muitos níveis de profundidade foram podadas para no máximo 5 níveis utilizando o parâmetro *max_depth=5* na chamada do classificador. Para todas as perspectivas, mesmo após a poda da árvore, as métricas precisão, revocação, *f1-score* e acurácia tiveram valores superiores a 0,95 nas bases de treino e teste; exceção para a perspectiva Gravidade, que obteve métricas piores na classe com menos instâncias, porém a acurácia ficou em 0,98 na base de treino e 0,99 na base de teste. A Tabela 5 apresenta os resultados em detalhe para cada perspectiva, bem como as variáveis utilizadas pelo modelo para geração das árvores após a poda, quando houve, listadas em ordem de importância. O atributo *feature_importances_* foi utilizado para recuperar a lista com a respectiva pontuação.

Tabela 3 – Informações sobre as árvores geradas para fins de explicabilidade dos grupos gerados para cada perspectiva

(continua)

Perspectiva	**Poda**	**Níveis da árvore final**	**Acurácia**	**Variável alvo**	**Precisão**	**Revocação**	***f1-score***	**Variáveis por ordem de importância**	
Condição Clínica	Não	5	1	0	1	1	1	*ProporcaoHGTHiper*	0,993645
				1	1	1	1	*ProporcaoTempElevada*	0,004453
				Média geral	1	1	1	*ProporcaoHGTHipo*	0,001902
				Média ponderada	1	1	1		
Gravidade	Sim	5	0,99	0	0,98	0,82	0,89	*ProporcaoDiasSemNora*	0,839481
				1	1	1	1	*UsoVasopressina*	0,094946
				2	0,99	0,98	0,99	*ProporcaoDiasNoraAte025*	0,062506
				3	0,9	0,99	0,94	*ProporcaoDiasNora025_050*	0,002519
				Média geral	0,97	0,95	0,96	*ProporcaoDiasNora050Mais*	0,000548
				Média ponderada	0,99	0,99	0,99		
Função Renal	Sim	4	0,99	0	1	1	1	*ProporcaoCreatininaAlta*	0,731955
				1	1	0,96	0,98	*ProporcaoUreiaAlta*	0,268045
				2	0,95	1	0,97		
				3	1	1	1		
				4	1	1	1		
				Média geral	0,99	0,99	0,99		
				Média ponderada	0,99	0,99	0,99		

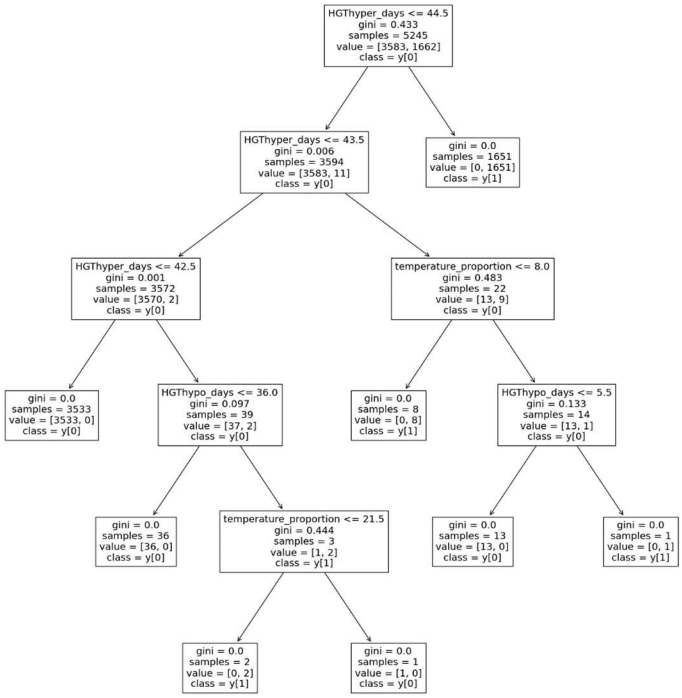
Tabela 5 – Informações sobre as árvores geradas para fins de explicabilidade dos grupos gerados para cada perspectiva

(conclusão)

Perspectiva	**Poda**	**Níveis da árvore final**	**Acurácia**	**Variável alvo**	**Precisão**	**Revocação**	***f1-score***	**Variáveis por ordem de importância**	
Eletrólitos	Sim	4	0,99	0	1	1	1	*ProporcaoSodioHiper*	0,367509
				1	0,96	1	0,98	*ProporcaoMagnesioHiper*	0,333967
				2	1	1	1	*ProporcaoSodioHipo*	0,297192
				3	0,99	0,93	0,96	*ProporcaoFosforoHipo*	0,001274
				4	1	0,99	1	*ProporcaoPotassioHiper*	0,000059
				Média geral	0,99	0,98	0,99		
				Média ponderada	0,99	0,99	0,99		
Perfil Nutricional	Sim	5	1	0	0,99	1	1	*ProporcaoAlbuminaBaixa*	0,463093
				1	1	0,99	1	*ProporcaoHemoglobinaBaixa*	0,374458
				2	1	1	1	*ProporcaoLinfocitosTotaisBaixo*	0,162449
				3	0,99	1	1		
				4	1	1	1		
				5	1	1	1		
				Média geral	1	1	1		
				Média ponderada	1	1	1		
Perfil Hepático	Não	4	1	0	1	1	1	*AlaninaMaxima*	0,478299
				1	1	1	1	*ProporcaoBilirrubinasAlta*	0,25922
				2	1	1	1	*AspartatoMaximo*	0,243279
				Média geral	1	1	1	*FosfataseMaxima*	0,018965
				Média ponderada	1	1	1	*ProporcaoTriglicerideAlto*	0,000237

A partir da análise de cada árvore de decisão gerada, foram extraídas as regras utilizadas para cada uma das perspectivas. A título de exemplo, a Figura 19 apresenta a árvore de decisão gerada para a perspectiva Condição Clínica. O Apêndice A apresenta a tradução das variáveis para seu respectivo nome interno utilizado no *dataset* de entrada para o modelo de agrupamento.

Figura 5. Árvore de decisão gerada para explicabilidade dos grupos da perspectiva Condição Clínica



Fonte: Autoria própria.

As regras obtidas a partir das árvores de decisão, possibilitaram auxiliar na interpretação dos grupos gerados para cada perspectiva, conforme apresentado na Tabela 6. A validação dos grupos se encontra em andamento pelo grupo de especialistas em saúde.

Tabela 4 – Perspectivas, grupos, regras e interpretação

(continua)

Perspectiva	Grupos	Regras
Condição Clínica	G0	- Pacientes com anotações de HGTHiper em até 43% dos dias de internamento; OU, - Pacientes com anotações de HGTHiper entre 43 e 45% dos dias de internamento E com anotações HGTHipo em pelo menos 8% dos dias de internamento E anotações de HighTemp em até 6% dos dias de internamento; OU, - Pacientes com anotações de HGTHiper entre 43 e 45% dos dias de internamento E com anotações HGTHipo em pelo menos 22% dos dias de internamento E anotações de HighTemp em pelo menos 36% dos dias de internamento;
	G1	- Pacientes com anotações de HGTHiper em mais de 45% dos dias de internamento; OU, - Pacientes com anotações de HGTHiper entre 43 e 45% dos dias com anotações HGTHipo em pelo menos 5% dos dias e anotações de HighTemp em pelo menos 8% dos dias; OU, - Pacientes com anotações de HGTHiper entre 43 e 45% dos dias de internamento E com anotações de HighTemp em até 8% dos dias de internamento; OU, - Pacientes com anotações de HGTHiper entre 43 e 45% dos dias de internamento E com anotações de HGTHipo em até 22% dos dias de internamento E anotações de HighTemp em mais de 36% dos dias de internamento;
Gravidade	G0	1) Pacientes com uso de nora(025) em até 39% dos dias de internamento E com uso de vaso em até 44% dos dias de internamento; OU, 2) Pacientes sem uso de nora em até 51% dos dias de internamento E com uso de nora(25-50) em pelo menos 10% dos dias de internamento E com uso de vaso em 45%(média) dos dias de internamento; OU, 3) Pacientes sem uso de nora em até 51% E com uso de nora(025) em até 65% dos dias de internamento E com uso de vaso em pelo menos 49% dos dias de internamento; OU, 4) Pacientes sem uso de nora entre 51% e 85% dos dias de internamento E com uso de vaso em 57%(média) dos dias de internamento E com uso de nora(+050) em pelo menos 9% dos dias de internamento; OU, 5) Pacientes sem uso de nora entre 51% e 85% dos dias de internamento E com uso de vaso em pelo menos 61% dos dias de internamento;
	G1	1) Pacientes sem uso de nora em até 51% E com uso de nora(025) em até 10% dos dias de internamento E com uso de vaso em pelo menos 14% dos dias de internamento; OU, 2) Pacientes sem uso de nora em até 51% E com uso de vaso em até 14% dos dias de internamento;