Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciência
Instituto de Matemática e Estatística

Sabrinna Rodrigues de Oliveira de Souza

Métodos de Inteligência Artificial aplicados em dados de biomassa para a caracterização dos diferentes tipos de pirólise

Rio de Janeiro
2023

Sabrinna Rodrigues de Oliveira de Souza

Métodos de Inteligência Artificial aplicados em dados de biomassa para a caracterização dos diferentes tipos de pirólise

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Vinicius Layter Xavier
Coorientador: Prof. Dr. Aderval Severino Luna

Rio de Janeiro
2023

CATALOGAÇÃO NA FONTE

UERJ/REDE SIRIUS/BIBLIOTECA CTC/A

S729 Souza, Sabrinna Rodrigues de Oliveira de.
Métodos de inteligência artificial aplicados em dados de biomassa para a caracterização dos diferentes tipos de pirólise. – 2023.
74 f.: il.

Orientador: Vinicius Layter Xavier
Coorientador: Aderval Severino Luna
Dissertação (Mestrado em Ciências Computacionais) - Universidade do Estado do Rio de Janeiro, Instituto de Matemática e Estatística.

1. Inteligência artificial - Teses. 2. Agrupamento de dados relacionados - Teses. 3. Biomassa - Teses. 4. Aprendizado de máquina - Teses. I. Xavier, Vinicius Layter. II. Luna, Aderval Severino. III. Título.

CDU 004.8

Patricia Bello Meijinhos CRB7/5217 - Bibliotecária responsável pela elaboração da ficha catalográfica

Sabrinna Rodrigues de Oliveira de Souza — Assinatura

05/05/2023 — Data

Sabrinna Rodrigues de Oliveira de Souza

Métodos de Inteligência Artificial aplicados em dados de biomassa para a caracterização dos diferentes tipos de pirólise

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Aprovada em 24 de março de 2023.
Banca Examinadora:

Prof. Dr. Vinicius Layter Xavier (Orientador)
Instituto de Matemática e Estatística – UERJ

Prof. Dr. Aderval Severino Luna (Coorientador)
Instituto de Química – UERJ

Prof. Dr. Alexandre Rodrigues Tôrres
Faculdade de Tecnologia – UERJ

Prof.ª Dra. Maria Clícia Stelling de Castro
Instituto de Matemática e Estatística– UERJ

Prof.ª Dra. Daniela Polessa Paula
Escola Nacional de Ciências Estatísticas – ENCE

Prof. Dr. Igor Campos de Almeida Lima
Instituto de Matemática e Estatística– UERJ

Rio de Janeiro
2023

DEDICATÓRIA

A minha família.

AGRADECIMENTOS

Primeiro e antes de tudo, agradeço a Deus, por me guiar, iluminar e me conceder o dom da fortaleza. À Nossa Senhora, São José e Santa Teresinha que me ampararam nos momentos de dificuldades e foram meu auxílio e proteção, desatando todos os nós e abrindo os caminhos necessários.

Ao meu amável esposo André, por todo apoio, incentivo e palavras sábias nos momentos mais difíceis. Nos momentos de maiores dificuldades sua companhia e apoio foram primordiais.

À minha mãe Matilde, que acompanhou cada passo comigo, torcendo, vibrando e me encorajando dia após dia. Nas horas de incerteza suas palavras e orações foram combustíveis para seguir minha caminhada.

Ao meu pai Ivan e minha irmã Samantha, que sempre me apoiaram em tudo.

À minha vó Izaura (in memoriam), pelo grande exemplo de garra e determinação que deixou para nós.

Aos meus amigos de vida e de mestrado, que tanto me apoiaram e torceram por mim.

Ao meu querido orientador, professor Vinicius Layter Xavier, por todo carinho, atenção, apoio e dedicação que me deu ao longo desta caminhada, tenho um enorme carinho e admiração pelo professor que é. Também agradeço ao meu querido coorientador, professor Aderval Luna, por todos os ensinamentos, trocas, incentivos e apoio que meu deu ao longo desse processo. Sem o apoio e disponibilidade de vocês, eu não chegaria até aqui.

Agradeço também o apoio da FAT (Faculdade de Tecnologia da UERJ), pelo fornecimento dos dados de biomassa para que este trabalho pudesse ser realizado.

A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original.

Albert Einstein

RESUMO

SOUZA, Sabrinna Rodrigues de Oliveira de.*Métodos de Inteligência Artificial aplicados em dados de biomassa para a caracterização dos diferentes tipos de pirólise.* 2023. 77 f. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística,Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

Este estudo aborda o problema de classificação de dados de biomassa. Um dos objetivos é identificar as variáveis de maior relevância para classificar o tipo de pirólise de biomassa. Além disso, avaliar se as classes dos tipos de pirólise são suficientes para caracterizar este processo químico. O algoritmo de Floresta Aleatória foi aplicado para identificar quais variáveis são relevantes no processo de classificação do tipo pirólise, obtendo uma exatidão em torno de 97%. Foi identificado que as variáveis mais importantes são: Tempo de residência médio no reator para o gás e de arraste, Porcentagem de carbono em base seca livre de cinza na matéria-prima, Tamanho da partícula média no reator e Porcentagem de hidrogênio em base seca livre de cinza na matéria-prima. Os seguintes métodos de agrupamentos foram usados com as variáveis de maior relevância encontradas: k-means, pam, clara, diana, fanny, hierárquico, som, sota e model. Para avaliar os métodos de agrupamentos, foram usadas as medidas de validação interna com as métricas de índice de Dunn e silhueta. As medidas de validação indicaram o agrupamento hierárquico e k-means com os melhores resultados para número de grupos maior do que o número de classes de pirólise já existentes. Assim, o conjunto de dados pode ser dividido em um número maior de tipos de pirólise, levando em consideração à conclusão que apenas as classes disponíveis são muito limitadas para caracterizar os tipos de pirólise, uma vez que os algoritmos de classificação não supervisionados indicam o número de agrupamentos como maior ou igual a cinco.

Palavras-chave: Análises de agrupamento. Inteligência Artificial. Biomassa. Aprendizado de máquina.

ABSTRACT

SOUZA, Sabrinna Rodrigues de Oliveira de.*Artificial Intelligence methods applied to biomass data to characterize the different types of pyrolysis.* 2023. 77 f. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística , Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

This study addresses a biomass data classification problem. One of the objectives is to identify the most relevant variables for classifying the pyrolysis type of biomass. Also, to evaluate if the pyrolysis type classes are sufficient to characterize this chemical process. The Random Forest algorithm was applied to identify which variables are relevant in the pyrolysis-type classification process, obtaining an accuracy of around 97%. It was identified that the most important variables are: Average residence time in the reactor for the gas and carrier, Percentage of ash-free dry basis carbon in the feedstock, Average particle size in the reactor, and Percentage of ash-free dry basis hydrogen in the feedstock. The clustering methods were used with the most relevant variables: k-means, pam, clear, diana, fanny, hierarchical, sound, sota, and model. To evaluate the clustering methods, internal validation measures with Dunn's index metrics and silhouette were used. The validation measures indicated the hierarchical clustering and k-means with better results for several groups greater than the number of existing pyrolysis classes. Thus, the dataset can be divided into more pyrolysis-type groups, considering only the available classes are too limited to characterize the pyrolysis type since the unsupervised classification algorithms indicate the number of clusters as greater than or equal to five.

Keywords: Cluster Analysis. Artificial Intelligence. Biomass. Machine Learning.

LISTA DE ILUSTRAÇÕES

Figura 1 - Esquema do processo de Pirólise. 16
Figura 2 - Aprendizado Supervisionado. 20
Figura 3 - Aprendizado não Supervisionado. 22
Figura 4 - Aprendizado semi-Supervisionado. 22
Figura 5 - Algoritmo de Floresta Aleatória 24
Figura 6 - Esquema representativo método *k-fold*. 26
Figura 7 - Análise de agrupamento. 28
Figura 8 - Método hierárquico de agrupamento. 30
Figura 9 - Representação gráfica da curva *u-shaped*. 43
Figura 10 - Histograma das variáveis (Batelada). 47
Figura 11 - Correlação das variáveis (Batelada). 48
Figura 12 - Histograma das variáveis (Contínuo). 49
Figura 13 - Correlação das variáveis (Contínuo). 50
Figura 14 - Representação gráfica da taxa de erro *Out of bag* em relação ao número de árvores (Batelada). 52
Figura 15 - Representação gráfica da importância de cada variável na classificação considerando a Acurácia e o ìndice de Gini (Batelada). 53
Figura 16 - Representação gráfica do Índice de Dunn em relação ao número de grupos (Batelada). 55
Figura 17 - Representação gráfica da Silhueta em relação ao número de grupos (Batelada). 56
Figura 18 - Representação gráfica de radar do Índice de Dunn (Batelada). 57
Figura 19 - Representação gráfica de radar da silhueta (Batelada). 58
Figura 20 - Representação gráfica da silhueta (Batelada). 59
Figura 21 - Representação gráfica da silhueta (Batelada). 60
Figura 22 - Representação gráfica do dendrograma (Batelada). 61
Figura 23 - Representação gráfica da redução de dimensionalidade do plano R^4 para R^2 (Batelada) 62
Figura 24 - Representação gráfica da taxa de erro *Out of bag* em relação ao número de árvores (Contínuo). 63
Figura 25 - Representação gráfica da importância de cada variável na classificação considerando a Acurácia e o ìndice de Gini(Contínuo). 64
Figura 26 - Representação gráfica do Índice de Dunn em relação ao número de grupos (Contínuo). 66
Figura 27 - Representação gráfica da silhueta em relação ao número de grupos (Contínuo). 66

Figura 28 - Representação gráfica de radar do Índice de Dunn (Contínuo). 67
Figura 29 - Representação gráfica de radar da silhueta (Contínuo). 67
Figura 30 - Representação gráfica da silhueta (Contínuo). 68
Figura 31 - Representação gráfica da silhueta (Contínuo). 69
Figura 32 - Representação gráfica do dendrograma (Contínuo). 70
Figura 33 - Representação gráfica da redução de dimensionalidade do plano R^4 para R^2 (Contínuo). 71

LISTA DE TABELAS

Tabela 1 - Variáveis que agrupam os tipos de pirólise. 44
Tabela 2 - Variáveis numéricas e suas respectivas medidas (Batelada). 46
Tabela 3 - Variáveis numéricas e suas respectivas medidas (Contínuo). 48
Tabela 4 - Variáveis de maior relevância (Batelada). 54
Tabela 5 - Resultados da validação interna (Batelada). 54
Tabela 6 - Resultados consolidados (Batelada). 56
Tabela 7 - Resultados dos agrupamentos para o método hierárquico (Batelada). . 60
Tabela 8 - Resultados dos agrupamentos para o método *k-means* (Batelada). . . . 61
Tabela 9 - Variáveis de maior relevância (Contínuo). 64
Tabela 10 - Resultados da validação interna (Contínuo). 65
Tabela 11 - Resultados consolidados (Contínuo). 68
Tabela 12 - Resultados dos agrupamentos para o método hierárquico (Contínuo). . 70
Tabela 13 - Resultados dos agrupamentos para o método *k-means* (Contínuo). . . . 70

LISTA DE ABREVIATURAS E SIGLAS

AGNES	Agglomerative Nesting
Clara	Clustering Large Applications
Diana	Divisive Analysis
EMLS	Escalonamento Multidimensional Local Suavizado
Fanny	Fuzzy Analysis
FCM	Fuzzy C-means
LC	Continuidade Local
MDS	Escalonamento multidimensional
MDS Local	Escalonamento Multidimensional Local
MONA	Monothetic Analysis
OOB	Out-of-bag
Pam	Partitioning Around Medóides
RF	Random Foresta (Floresta Aleatória)

LISTA DE SÍMBOLOS

max	máximo
min	mínimo

SUMÁRIO

	INTRODUÇÃO	15
1	**JUSTIFICATIVAS E OBJETIVOS**	18
1.1	**Objetivos gerais**	18
1.2	**Objetivos específicos**	18
2	**REVISÃO BIBLIOGRÁFICA**	19
2.1	**Métodos de Inteligência Artificial**	19
2.1.1	Aprendizado de Máquina	19
2.1.2	Árvores de decisão	23
2.1.3	Floresta Aleatória (Random Forest)	23
2.1.3.1	Índice Gini	23
2.1.4	Técnicas para validação de modelos	25
2.1.4.1	Validação cruzada *k-fold*	25
2.2	**Agrupamento de dados**	25
2.2.1	Definição de Agrupamento	25
2.2.2	Medidas de Similaridade	31
2.2.2.1	Distância Euclidiana:	31
2.2.2.2	Distância *Manhattan*:	32
2.2.2.3	Distância de *Chebychev*:	32
2.2.2.4	Distância de *Mahalanobis*:	32
2.2.3	Algoritmos de agrupamento	32
2.2.3.1	*k-means*	33
2.2.3.2	*Pam - Partitioning Around Medóides*	33
2.2.3.3	*Clara - Clustering Large Applications*	34
2.2.3.4	*Fanny - Fuzzy Analysis*	35
2.2.3.5	Hierárquico	36
2.2.3.6	*Diana - Divisive Analysis*	37
2.2.3.7	*Som - Self-Organizing Map*	38
2.2.3.8	*Sota - State-Of-The-Art*	38
2.2.3.9	*Model*	39
2.2.4	Medidas de validação	39
2.3	**Método Escalonamento Multidimensional Local**	40
3	**BANCO DE DADOS**	44
4	**ANÁLISE EXPLORATÓRIA DOS DADOS**	46
4.1	**Análise do subconjunto do tipo de reator Batelada**	46
4.2	**Análise do subconjunto do tipo de reator Contínuo**	47
5	**RESULTADOS E ANÁLISES**	51

5.1 **Tipo de reator Batelada** 51
5.2 **Tipo de reator Contínuo** 62
CONCLUSÃO 72
REFERÊNCIAS 74

INTRODUÇÃO

Todos os dias a sociedade gera grandes quantidades de dados, que podem ser submetidos a análises e gerenciamento. Uma forma vital de lidar com esses dados é classificá-los ou agrupá-los em um conjunto de categorias ou agrupamentos. Para aprender um novo objeto ou compreender um novo fenômeno, as pessoas buscam recursos para descrevê-lo. Adicionalmente, eles o comparam com outros objetos ou fenômenos conhecidos com base na semelhança ou dissimilaridade, de acordo com certos padrões ou regras (LAM; WUNSCH, 2014).

Na classificação não supervisionada, também chamada de análises de agrupamentos, nenhum dado rotulado está disponível. Conforme observado por Backer e Jain, "na análise de agrupamento, um grupo de objetos é dividido em vários subgrupos mais ou menos homogêneos com base em uma medida de similaridade frequentemente escolhida subjetivamente (ou seja, escolhida subjetivamente com base em sua capacidade de criar agrupamentos "interseccionados"), de modo que a similaridade entre objetos dentro de um subgrupo é maior do que a similaridade entre objetos pertencentes a diferentes subgrupos" (BACKER; JAIN, 1981).

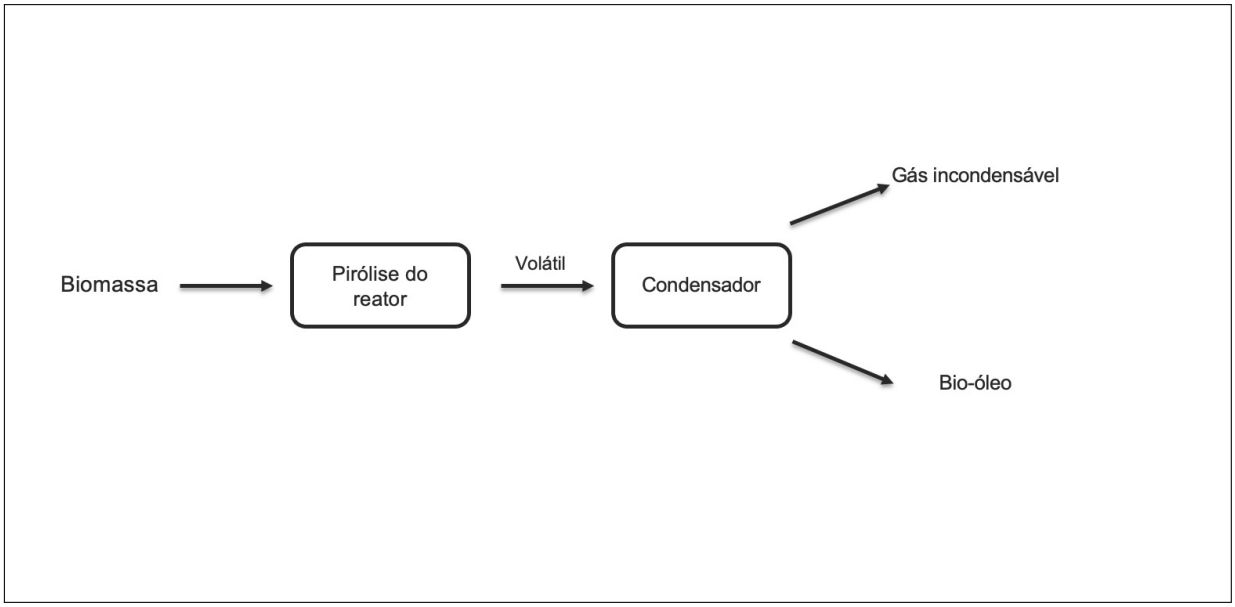
Nesta pesquisa, abordamos dados de biomassa, que é um termo utilizado para caracterizar todo material orgânico que se origina de plantas. O material orgânico é produzido através da fotossíntese, que usa a luz solar para converter dióxido de carbono e água em matéria orgânica. A biomassa é uma fonte alternativa de energia, e o bio-óleo obtido da pirólise da biomassa, é utilizado como combustível, produtos químicos e biomateriais (GUEDES; LUNA; TORRES, 2018). A pirólise é a degradação térmica do material orgânico que ocorre em concentrações de O_2 muito baixas, ou mesmo em um ambiente com uma concentração de O_2 capaz de barrar a gaseificação intensiva do material orgânico. Em geral, a pirólise ocorre a uma temperatura, variando de 400°C, até o início do sistema de gaseificação (VIEIRA et al., 2014).

O processo de pirólise é dividido em quatro tipos, dependendo das condições de operação utilizadas: *fast, slow, flash* e *catalytic*. A pirólise *slow* ocorre a uma temperatura de processo mais baixa, menor taxa de aquecimento e maior tempo de residência, o que favorece a produção de carvão. A pirólise *flash* é o processo no qual o tempo de reação é de apenas alguns segundos, ou até menos, e a taxa de aquecimento é muito alta. A pirólise *fast* favorece a formação de bio-óleo e ocorre a uma temperatura moderada, curto tempo de residência do vapor e alta taxa de aquecimento, mas não tão alta quanto na pirólise *flash* e a pirólise *catalytic* serve para melhorar a qualidade do óleo produzido (GUEDES; LUNA; TORRES, 2018).

Os tipos de pirólise são agrupados nas respectivas classes (*fast, slow, flash* e *catalytic*) de acordo com as doze variáveis elementares, complementares e de processo, re-

presentadas na Tabela 1 no capítulo 3. A Figura 1 ilustra o processo de pirólise com a produção de líquidos, gases e sólidos.

Figura 1 - Esquema do processo de Pirólise.



Fonte: Adaptado de (GUEDES; LUNA; TORRES, 2018).

O banco de dados utilizado contém diferentes processos de pirólise e condições de operação e foi construído para entender melhor como esses fatores influenciam a composição e o rendimento dos produtos da pirólise. Além disso, foi desenvolvido a partir de dados experimentais obtidos de mais de 200 pesquisas de pirólise de biomassa disponíveis na literatura desde 1984, incluindo os artigos mais citados, contendo dados experimentais, bem como pesquisas recentes na área. Este banco de dados foi desenvolvido pelos autores (GUEDES; LUNA; TORRES, 2018).

Algumas análises aplicadas em dados de biomassa são encontradas na literatura. Algumas delas são aplicações voltadas para redes neurais artificiais, como por exemplo o estudo do autor (MOSCATO, 2019) que gerou análises exergéticas - capacidade de transformar energia em trabalho, ou seja, em energia organizada - de uma caldeira de biomassa baseada em redes neurais artificiais. Outro estudo é o do autor (MERDUN, 2018) que apresentou a aplicação de dois métodos de redes neurais artificiais (feed-forward network e cascade-forward network) na modelagem de rendimentos de produtos de pirólise (Bio-carvão, Bio-óleo e Mistura de gás) usando nove tipos de biomassa e dois parâmetros de processo de pirólise como variáveis de entrada para os modelos.

Um outro estudo é o dos autores (ÖZBAY; KÖKTEN, 2020) que desenvolveram um modelo confiável de rede neural artificial para modelar o produto líquido da pirólise, considerando os tipos de pirólise *flash* e *slow*. Outro exemplo sobre redes neurais é o estudo dos autores (CAO; XIN; YUAN, 2016) que desenvolveram modelos de Inteligência

Artificial, baseados em redes neurais artificiais e máquina de vetor de suporte, para prever a distribuição de produtos e alto valor de aquecimento de bio-óleo de pirólise *fast* de biomassa em leitos fluidizados borbulhantes.

Outras aplicações encontradas na literatura abordam os conceitos de mínimos quadrados com uma abordagem inteligente de máquina de vetor de suporte. É o caso do estudo realizado por (CAO; XIN; YUAN, 2016) onde fizeram uma previsão do rendimento de Bio-carvão da pirólise de estrume de gado. Porém, não há muitos estudos na área de biomassa com métodos de Inteligência Artificial voltados para métodos de agrupamentos. Também não há estudos que avaliam se as classes dos tipos de pirólise são suficientes para caracterizar os dados de biomassa. Além disso, não existem muitos estudos associando as variáveis mais importantes na classificação do tipo de pirólise. Esta dissertação tem como diferencial realizar o que não é encontrado na literatura sobre a seleção de variáveis mais importantes dos dados de biomassa e classificação dos tipos de pirólise para dados de biomassa, como forma de contribuir e preencher essa lacuna.

Nesta dissertação, é realizada uma aplicação do algoritmo de Floresta Aleatória (BREIMAN, 2001) no intuito de identificar as variáveis com maior relevância para a classificação do tipo de pirólise. Com as variáveis identificadas, foram aplicados nove métodos de agrupamento: *k-means*, Pam, Clara, Diana, Fanny, Hierárquico, Som, Sota, Model. Para avaliar os métodos de agrupamentos, foi utilizada a medida de validação interna.

- Organização do trabalho:

Esta dissertação está dividida em seis capítulos. O primeiro capítulo descreve a introdução, a motivação do estudo e o que será abordado ao longo da dissertação. O segundo capítulo destaca as principais justificativas e os principais objetivos da pesquisa. O terceiro capítulo fornece uma revisão bibliográfica das principais técnicas e métodos utilizados, ou seja, os detalhamentos da fundamentação teórica dos métodos de Inteligência Artificial aplicados nesta dissertação. As técnicas específicas do aprendizado de máquina são descritas de forma breve, como o algoritmo de Floresta Aleatória do aprendizado supervisionado, os métodos de agrupamentos do aprendizado não supervisionado e as medidas de validação utilizadas. O quarto capítulo descreve o banco de dados utilizado. O quinto capítulo descreve a análise exploratória dos dados. O sexto capítulo apresenta os resultados obtidos pela aplicação dos algoritmos, assim como as discussões, avaliando principalmente a performance dos métodos. Por fim, o sétimo capítulo aborda as conclusões finais.

1 JUSTIFICATIVAS E OBJETIVOS

1.1 Objetivos gerais

A meta é encontrar as variáveis de maior relevância para a classificação dos tipos de pirólises de biomassa e avaliar se os tipos de pirólises já existentes são suficientes para caracterizar o processo químico. Em seguida, avaliar esses resultados, propondo um número de agrupamentos ideal e o melhor método de agrupamento.

1.2 Objetivos específicos

- Análise exploratória do conjunto de dados;
- Aplicação do algoritmo Floresta Aleatória, um método supervisionado, no conjunto de dados para encontrar as variáveis de mais importância;
- Aplicação de nove métodos de algoritmos de agrupamento, métodos não supervisionados, sendo eles: k-means, pam, clara, diana, fanny, hierárquico, som, sota, model.
- Avaliação e análise dos resultados dos melhores métodos de agrupamentos por meio da validação interna com os indicadores índice Dunn e silhueta.
- Aplicação de redução de dimensionalidade com o método de Escalonamento Multidimensional Local com suavização hiperbólica como meio de avaliar se os grupos estariam bem divididos projetando-os em espaços de menor dimensão com máxima conservação das conexões de vizinhança entre os dados.

2 REVISÃO BIBLIOGRÁFICA

2.1 Métodos de Inteligência Artificial

Diante dos dados disponíveis, esta dissertação engloba conceitos da área de Inteligência Artificial. Essa área ganhou força no meio computacional pelo fato de possuir um campo muito importante: a Aprendizagem de Máquina. Muita das vezes, depois de visualizar os dados, não é possível interpretar as informações extraídas dos dados e, neste caso, são utilizadas as técnicas de Aprendizagem de Máquina. Com a abundância dos conjuntos de dados disponíveis, a demanda por Aprendizado de Máquina tem encorajado as indústrias a usarem para a extração de dados relevantes (MAHESH, 2020).

2.1.1 Aprendizado de Máquina

O Aprendizado de Máquina é um subcampo da Inteligência Artificial que estuda métodos computacionais como forma de obter novas competências e meios de organizar as informações. Em um processo de aprendizado, os algoritmos se baseiam em dados já existentes para obter novos conhecimentos. O aprendizado de máquina pode ser realizado de duas formas: supervisionado ou não supervisionado (SOUTO et al., 2003).

Não há apenas um tipo único de algoritmo de aprendizado de máquina que seja melhor para resolver um problema. O tipo de algoritmo empregado depende do tipo de problema que se deseja resolver, o número de variáveis, o tipo de modelo que melhor se adequa aos dados e assim por diante. Segue abaixo uma breve descrição de alguns dos algoritmos que são usados em aprendizado de máquina:

- **Método supervisionado**

  O aprendizado supervisionado utiliza exemplos previamente rotulados, ou seja, é fornecido ao sistema um conjunto de dados da seguinte forma:

  $$X = \{X_1\, X_2 \ldots X_n\} \tag{1}$$

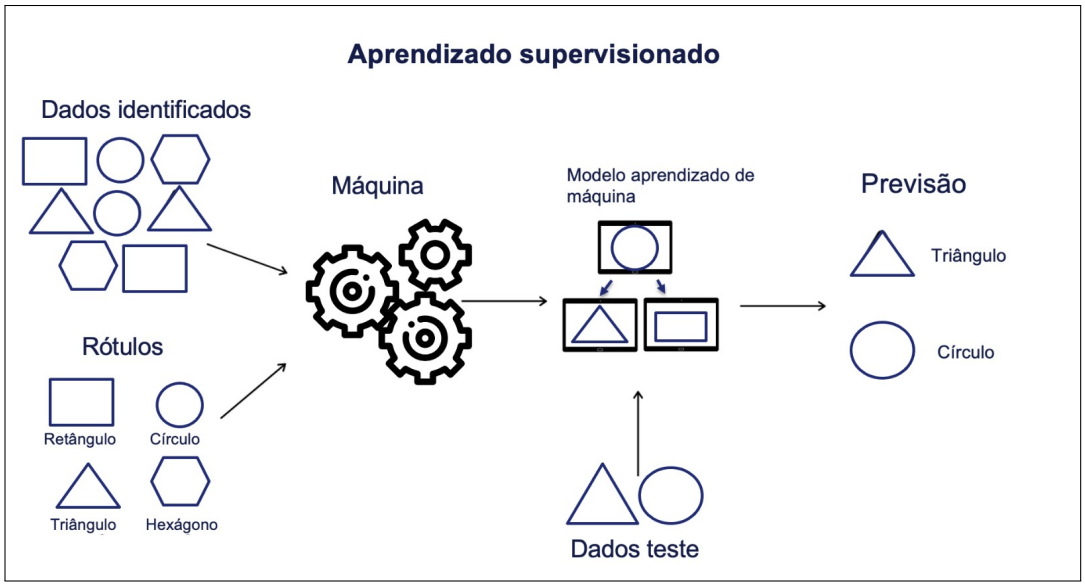
  Sendo que cada dado $\mathbf{X}_i \in \mathbf{X}$ possui um rótulo associado (MILARÉ, 2003). Este rótulo define a classe a qual o dado pertence, ou seja, cada dado $\mathbf{X}_i \in \mathbf{X}$ é uma tupla: $\mathbf{X}_i = (\mathbf{x}_i, \mathbf{y}_i)$ onde $\mathbf{x}_i$ é um vetor de valores que representam as características de $\mathbf{X}_i$ e $\mathbf{y}_i$ é o valor da classe desse dado (MILARÉ, 2003).

  Em resumo, conforme afirmam os autores (LAM; WUNSCH, 2014), os sistemas

supervisionados consideram que existe um conhecimento anterior da estrutura dos dados em que a especificação da pertinência a um dos grupos é conhecida a priori para cada uma das observações. Assim, na classificação supervisionada, o desafio é poder classificar uma nova observação dentro de um dos grupos, usando a estrutura intrínseca aos dados de entrada.

Nesse tipo de aprendizado, estima-se a taxa de acerto e taxa de erro obtidas por um classificador e os dados são separados em dois subconjuntos: treinamento e teste. O subconjunto de treinamento é utilizado no aprendizado do classificador e o de teste mede a capacidade de generalização na predição (LORENA; CARVALHO, 2003), conforme apresentado na Figura 2.

Figura 2 - Aprendizado Supervisionado.



Legenda: Fluxo de Trabalho de Aprendizagem Supervisionada.
Fonte: Adaptado de (RAJ, 2023).

Nesse cenário, o poder de generalização de um classificador pode levar a duas situações: o sobreajuste que ocorre quando o modelo se ajusta muito bem aos dados, ou seja, acontece um superajustamento dos dados, porém se mostra ineficaz para prever novos resultados, apresentando uma baixa exatidão quando confrontado com novos dados e o subajuste em que ocorre uma baixa exatidão no conjunto de treinamento (LORENA; CARVALHO, 2003).

Um outro ponto importante é que o conjunto de dados deve possuir uma quantidade de observações suficientes sob a ótica da Estatística, para ter uma divisão coerente em treinamento e teste. Se isso não acontecer, é necessário utilizar um método de Validação Cruzada conhecido como *k-fold* (SILVA; SPATTI; FLAUZINO, 2010).

Em resumo, no aprendizado supervisionado, o conjunto de dados original é, geralmente, dividido em três conjuntos denominados conjunto de treinamento, de teste e de validação, descritos a seguir:

- **conjunto treinamento**: esse conjunto é a principal entrada dos algoritmos de aprendizado supervisionado. A partir dele são induzidas as hipóteses, e, portanto, ele deve representar a distribuição da população para que se possa realizar com sucesso a indução de classificadores (METZ, 2006).
- **conjunto teste**: esse conjunto é utilizado para avaliar o modelo induzido. Para que essa avaliação seja válida (estatisticamente), os exemplos contidos nesse conjunto devem ser exemplos não utilizados pelo algoritmo durante a construção da hipótese, i.e., a intersecção desse conjunto com o conjunto de treinamento deve ser o conjunto vazio (METZ, 2006).
- **conjunto validação**: em alguns casos, pode ser necessário utilizar exemplos para realizar ajustes no modelo induzido. Esses exemplos não são utilizados diretamente na indução do modelo, mas são utilizados na escolha da complexidade mais adequada para o modelo. Dessa maneira, esses casos são indiretamente "vistos" pelo algoritmo durante o processo de indução, o que implica que os exemplos de validação devam ser distintos dos exemplos de teste e treinamento (METZ, 2006).

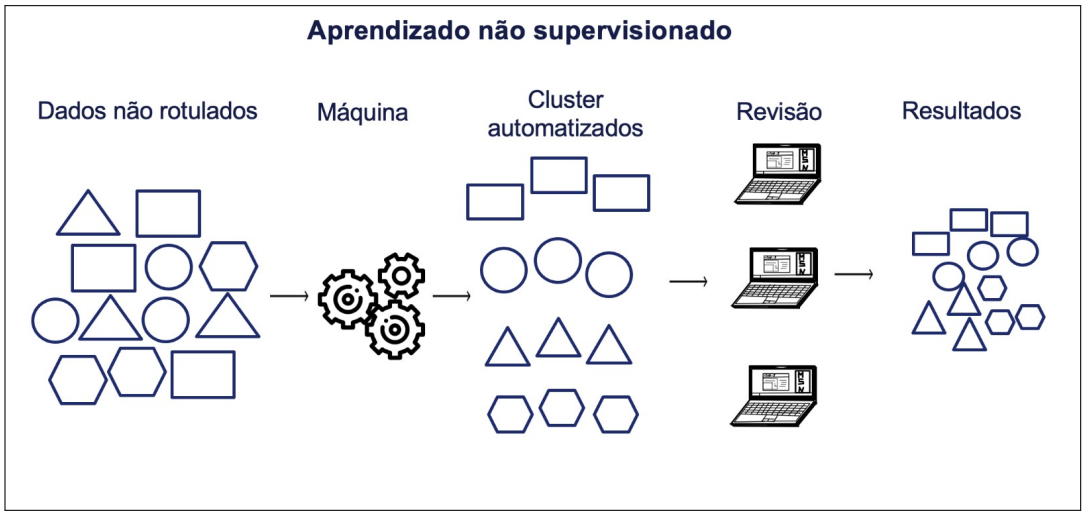
- **Método não supervisionado**

  Neste caso, os dados não possuem uma classe previamente rotulada. Os dados fornecidos são analisados e, em seguida, é feita a tentativa de determinar se alguns deles podem ser agrupados de alguma forma. Ou seja, é fornecido ao sistema um conjunto de dados $\mathbf{X}$, no qual cada dado consiste somente de vetores $\mathbf{x}$, não incluindo a informação sobre a classe $\mathbf{y}$ à qual ele pertence. Na Figura 3 é possível observar o funcionamento deste algoritmo de forma sintetizada.

- **Método semi-supervisionado**

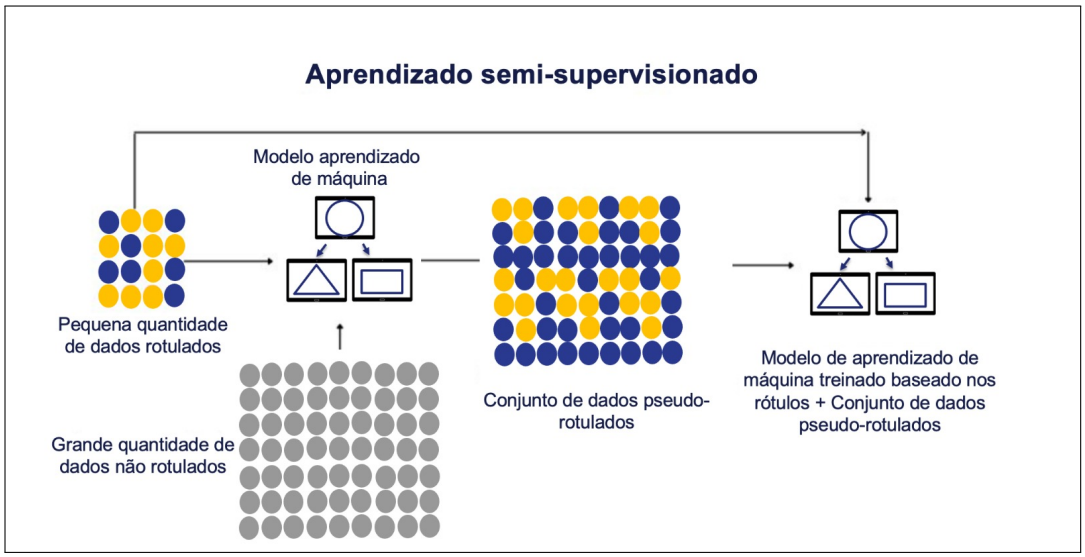
  Devido às limitações que o aprendizado supervisionado apresenta no sentido de necessitar de um conjunto de dados com uma quantidade grande de exemplos rotulados para chegar a um bom resultado, o modo de aprendizado semi-supervisionado utiliza poucos exemplos rotulados e muitos exemplos não rotulados. A ideia principal desses algoritmos é rotular um maior número de exemplos para os quais a classe não é conhecida com a finalidade de melhorar a performance de algoritmos de aprendizado supervisionado (METZ, 2006). Na Figura 4 é possível observar o funcionamento deste algoritmo de forma sintetizada.

Figura 3 - Aprendizado não Supervisionado.



Legenda: Fluxo de Trabalho de Aprendizagem não supervisionada.
Fonte: Adaptado de (RAJ, 2023).

Figura 4 - Aprendizado semi-Supervisionado.



Legenda: Fluxo de Trabalho de Aprendizagem semi-supervisionada.
Fonte: Adaptado de (RAJ, 2023).

2.1.2 Árvores de decisão

As árvores de decisão são métodos utilizados como uma forma de dividir os dados. Para tomar uma decisão utilizando os modelos baseados em árvores, as amostras vão se dividindo conforme os critérios estabelecidos para uma determinada variável, caso a resposta seja "sim"as amostras vão para um lado específico, por exemplo, para a esquerda, e caso seja "não", seguem para o lado oposto. O processo de selecionar uma variável e dividir as amostras continua para cada novo nó subsequente até que alguma regra seja atingida (SANTANA, 2020).

2.1.3 Floresta Aleatória (Random Forest)

O algoritmo de Floresta Aleatória (BREIMAN, 2001), aprendizado supervisionado, é aplicado nesta dissertação como classificador e como ferramenta de seleção do melhor conjunto de atributos. Este algoritmo é uma ferramenta popular de aprendizado de máquina baseada em árvore e indicada para análise de dados de alta dimensão.

O algoritmo Floresta Aleatória gera uma coleção de centenas a milhares de árvores. Na construção de cada árvore é utilizada uma amostra dos dados originais. As amostras são geradas com o método Bootstraping. (CHEN; ISHWARAN, 2012)

A Floresta Aleatória tem sido tradicionalmente aplicada em configurações de classificação e regressão. A construção da Floresta Aleatória para classificação e regressão são descritas nas etapas do algoritmo descritos na figura 5 (HASTIE et al., 2009).

Os dados que não pertencem a amostra bootstrap, são chamados de *out-of-bag* (OOB) e são utilizados para estimar o desempenho do modelo. A partir destes dados, é possível medir a taxa de erro do modelo de predição (BREIMAN, 2001).

2.1.3.1 Índice Gini

É possível encontrar as variáveis de maior relevância por meio do índice *Gini* que avalia a impureza de um nó (JAMES et al., 2013). O índice *Gini* em um determinado nó é dado pela seguinte equação 2 (BARBOSA; CARNEIRO; TAVARES, 2012):

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{2}$$

onde, p_i é a frequência relativa de cada classe em cada nó e c é o número de classes.

Figura 5 - Algoritmo de Floresta Aleatória

Algoritmo 1 - Algoritmo de Floresta Aleatória

1. Para b = 1 até B:

   - Gerar uma amostra *bootstrap* Z^* de tamanho N dos dados de treinamento.
   - Criar uma árvore T_b para os dados de *bootstrap*, repetindo recursivamente as seguintes etapas para cada nó da árvore, até que o tamanho mínimo do nó seja atingido:
     - Selecionar m variáveis aleatoriamente das p variáveis (preditoras).
     - Escolher o melhor ponto de divisão entre os m.
     - Dividir o nó em dois nós filhos.

2. Saída do conjunto de árvores $\{T_b\}_1^B$

Para fazer uma previsão em um novo ponto x :

Regressão: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Classificação: Seja $\hat{C}_b(x)$ seja a previsão de classe da b-ésima árvore da floresta aleatória.
Então $\hat{C}_{rf}^b(x)$ = voto majoritário $\hat{C}^b(x)_1^B$
FIM ALGORITMO
FIM DOCUMENTAÇÃO

Legenda: Descrição do algoritmo de Floresta Aleatória.
Fonte: (HASTIE et al., 2009)

Quando este índice é igual a zero, o nó é considerado puro. Mas quando ele se aproxima de um, o nó é considerado impuro, pois aumenta o número de classes uniformemente distribuídas neste nó (BARBOSA; CARNEIRO; TAVARES, 2012).

2.1.4 Técnicas para validação de modelos

Atualmente a validação cruzada *k-fold* (BURMAN, 1989) é o método mais simples para estimar o erro de previsão. Este método estima diretamente o erro esperado $Err = E[L(Y, f(X))]$, o erro médio de generalização quando o método *f(X)* é aplicado a uma amostra de teste independente da distribuição conjunta de X e Y (HASTIE et al., 2009).

Avalia-se a capacidade de predição para avaliar o desempenho de um modelo. Tendo em vista o algoritmo de Floresta Aleatória abordado na seção anterior, é necessário a utilização de técnicas que validem o modelo. Sendo assim, a seguir será abordada a teoria de uma dessas técnicas.

2.1.4.1 Validação cruzada *k-fold*

No método de validação cruzada as n observações da amostra original $\mathbf{X}$ são divididas em k conjuntos disjuntos de observações, sendo eles $k_1, k_2, ..., k_k$, cada um de tamanho m_k aproximadamente igual, tal que $n = \sum_{k=1}^{K} m_k$. A partir disso, a amostra de validação é composta pela partição $\mathbf{X}_k$, enquanto a amostra de treino engloba as outras *k-1* partições que não incluem a k-ésima partição, ou seja, o conjunto de treino é dado por $\mathbf{X}_{(-k)} = \mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_{k-1}, \mathbf{X}_{k+1}, ..., \mathbf{X}_K$ (BERTOLETTI et al.,).

Isso se repete k vezes, conforme representado na Figura 6:

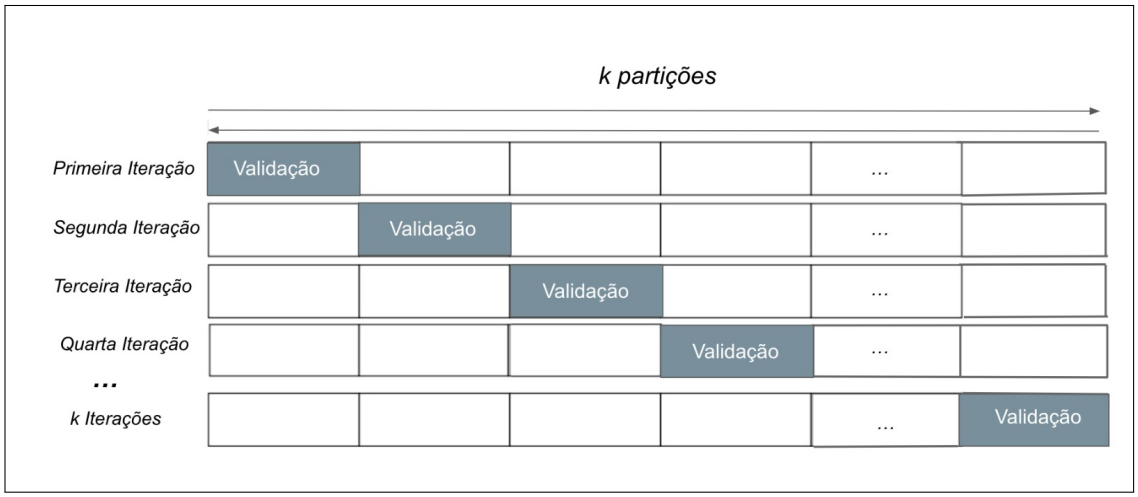
Em outras palavras, a validação cruzada *k-fold* é um método de validação cruzada que divide aleatoriamente o conjunto total de amostras em K partições. Utiliza-se *K-1* para produzir o subconjunto de treinamento, sendo o restante o subconjunto teste (SILVA; SPATTI; FLAUZINO, 2010).

2.2 **Agrupamento de dados**

2.2.1 Definição de Agrupamento

Muitas definições de agrupamento são encontradas na literatura, as mais utilizadas são (BRIAN, 1993):

Figura 6 - Esquema representativo método *k-fold*.



Legenda: Esquema representativo das K iterações inerentes ao procedimento *k-fold*.
Fonte: Adaptado de (BERTOLETTI et al.,).

Definição 1: um agrupamento é um conjunto de objetos semelhantes, e objetos que pertencem a grupos diferentes, não são iguais.

Definição 2: um agrupamento é um conjunto de pontos no espaço de teste de tal maneira que a distância entre quaisquer dois pontos em um mesmo grupo é menor que a distância entre qualquer ponto desse grupo e um outro ponto qualquer não pertencente a ele.

Definição 3: agrupamentos podem ser descritos como regiões conectadas de um espaço multidimensional contendo uma alta densidade relativa de pontos, separadas de outras regiões por uma região contendo uma baixa densidade relativa de pontos.

As análises de agrupamentos já foram aplicadas em diversos campos, desde engenharia (aprendizado de máquina, inteligência artificial, reconhecimento de padrões, etc.), ciências da computação (mineração da web, análise de banco de dados, segmentação de imagens, etc.), ciências da vida e médicas (genética, biologia, química, microbiologia, paleontologia, psiquiatria, clínica, etc.), até ciências da terra (geografia, geologia, sensoriamento remoto), ciências sociais (sociologia, psicologia, arqueologia, educação) e economia (marketing, negócios) (LANDAU; STER, 2010).

Assim, o agrupamento também é conhecido como taxonomia numérica, aprendizado sem professor (ou aprendizado não supervisionado), análise tipológica e particionamento. Embora a diversidade desses nomes reflita a importante posição do agrupamento

na pesquisa científica, as diferentes terminologias e objetivos também resultam em confusão. Algoritmos de agrupamento desenvolvidos para resolver um problema particular em um campo especializado geralmente fazem suposições em relação a aplicação de interesse. Esses vieses inevitavelmente afetam o desempenho do algoritmo em outros problemas que não satisfazem as mesmas premissas (LAM; WUNSCH, 2014)

Os algoritmos de agrupamento particionam os dados em grupos (subconjuntos ou categorias), mas esses grupos não têm uma definição universalmente aceita (LANDAU; STER, 2010). Os autores (LANDAU; STER, 2010) também afirmam que um grupo pode ser "um conjunto de entidades que são semelhantes e as entidades dentro de cada grupo não são semelhantes com as entidades dos outros grupos". (LAM; WUNSCH, 2014).

Em outras palavras, a análise de agrupamentos ou de conglomerados (cluster analysis) é um método estatístico que permite agrupar observações em grupos homogêneos em função do grau de similaridade entre as observações (AZEVEDO; ANZANELLO, 2015). São técnicas não supervisionadas e tem sido cada vez mais utilizadas. Essas análises consistem em agrupar um conjunto de observações de modo que as observações que pertençam a um mesmo grupo sejam parecidas entre si e diferentes das dos demais grupos. Assim, temos dois princípios básicos da análise de agrupamento que são homogeneidade e separação. Logo, quanto mais homogêneos são os elementos dentro de um grupo, mais separados ou diferentes são os grupos (XAVIER, 2012). Semelhanças e diferenças devem ter o potencial de serem examinadas de maneira clara e significativa.

Como forma de colocar em um mesmo grupo objetos que sejam similares de acordo com algum critério pré-determinado, se utiliza uma função de dissimilaridade como critério, função esta que recebe dois objetos e retorna a distância entre eles. Os grupos devem apresentar alta homogeneidade interna e alta separação (heterogeneidade externa). Isso quer dizer que os elementos de um determinado conjunto devem ser mutuamente similares e, preferencialmente, muito diferentes dos elementos de outros conjuntos.

Os objetos também são denominados exemplos, tuplas e/ou registros. Cada objeto representa uma entrada de dados que pode ser constituída por um vetor de atributos que são campos numéricos, como por exemplo, idade(inteiro), temperatura (real), entre outros ou categóricos, como por exemplo, bases de DNA (um dentre os valores A, C, G ou T), entre outros (GORDON, 1999).

Hruschka et al. (2005) considera um conjunto de n objetos $\mathbf{X}_n = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ a ser agrupado, onde cada $\mathbf{x}_i \ \epsilon \ R_p$ é um vetor de atributos que consiste em medidas reais "p". Os objetos devem ser classificados em grupos não sobrepostos $\mathbf{C} = \mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k$ onde k é o número de grupos, de modo que:

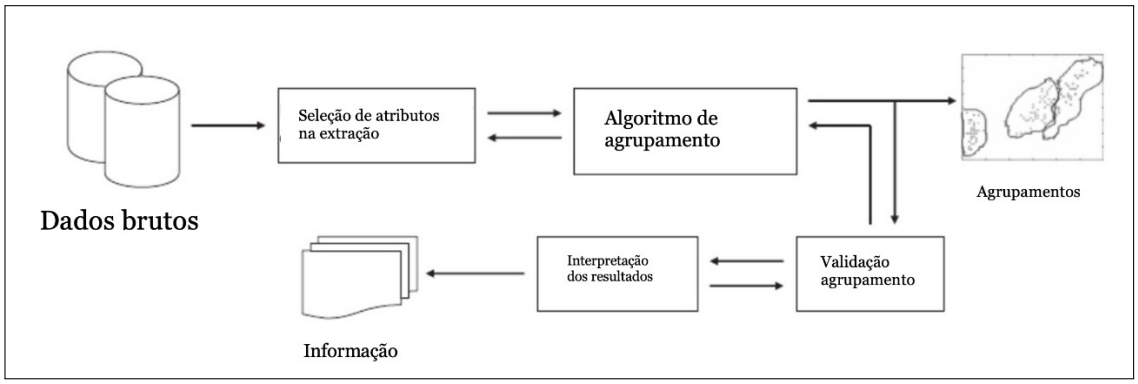
$$C_1 \cup C_2 \cup ... \cup C_k = X \quad , \tag{3}$$

$$C_i \neq \oslash \quad , \tag{4}$$

$$C_i \cap C_j = \oslash, \forall i \neq j \tag{5}$$

Ou seja, a união dos agrupamentos será definida como **X**, todo e qualquer agrupamento deve ser diferente de um conjunto vazio e a intersecção com qualquer agrupamento deverá ser um conjunto vazio. De forma sintetizada, a Figura 7 descreve as quatro etapas básicas do procedimento de análises de agrupamentos.

Figura 7 - Análise de agrupamento.



Legenda: Procedimento das técnicas de agrupamentos.
Fonte: Adaptado de (LAM; WUNSCH, 2014)

Existem duas estratégias principais para resolver problemas de agrupamento: métodos hierárquicos e métodos de partição. Métodos hierárquicos produzem uma hierarquia de partições de um conjunto de observações. Os métodos de partição, em geral, definem um determinado número de agrupamentos e, essencialmente, buscam otimizar uma função objetiva, tendo uma avaliação da homogeneidade dentro do agrupamento (XAVIER, 2012).

Também existem outras estratégias, como por exemplo, o agrupamento baseado em modelo, em que cada agrupamento é modelado por uma distribuição estatística. No agrupamento de densidade estimada, cada grupo é definido como uma região densa conectada. O agrupamento de subespaço visualiza os grupos como um subespaço. O agrupamento pode ser classificado como *hard* ou *fuzzy*. No agrupamento *hard*, cada objeto pertence a um e apenas um agrupamento, enquanto no agrupamento *fuzzy*, cada objeto tem algum grau de associação em cada agrupamento (LAM; WUNSCH, 2014). Descrevendo mais detalhadamente os métodos hierárquicos e métodos de partição, são apresentadas a seguir as seguintes definições:

- **Método hierárquico:**

  Os resultados de um algoritmo de agrupamento hierárquico geralmente são representados por uma árvore binária ou dendrograma, conforme representado na Figura 8. O nó raiz do dendrograma representa todo o conjunto de dados e cada nó da folha representa um objeto de dados. Os nós intermediários descrevem a medida em que os objetos são próximos uns aos outros e a altura do dendrograma geralmente expressa a distância entre cada par de objetos ou agrupamentos ou um objeto e um agrupamento. Um melhor resultado de agrupamento pode ser obtido cortando o dendrograma em diferentes níveis. Esta representação fornece descrições muito informativas e uma potencial visualização das estruturas de agrupamento de dados, especialmente quando existem relacionamentos hierárquicos reais nos dados (LAM; WUNSCH, 2014).

  Os autores (LAM; WUNSCH, 2014) ainda afirmam que os algoritmos hierárquicos são classificados principalmente como aglomerativos ou métodos divisivos. O agrupamento aglomerativo começa com n agrupamentos, cada um dos quais inclui exatamente um objeto. Uma série de operações de mesclagem combina todos os objetos no mesmo grupo. O agrupamento de método divisivo procede de maneira oposta. Todo o conjunto de dados inicialmente pertence a um único agrupamento e, em seguida, um procedimento o divide sucessivamente até que todos os agrupamentos sejam individuais. Para um agrupamento com n objetos, há $2^{N-1} - 1$ possíveis divisões de dois subconjuntos, o que é muito caro computacionalmente. Portanto, agrupamento divisivo não é comumente usado na prática. A variedade de algoritmos de agrupamento aglomerativo surgiu originalmente por causa das diferentes definições para a distância entre dois agrupamentos. A fórmula geral para a distância foi proposta por Lance e Williams como:

$$D(C_l, (C_i, C_j)) = \alpha_i D(C_l, C_i) + \alpha_j D(C_l, C_i) + \beta D(C_i, C_j) + \gamma |D(C_l, C_i) - D(C_l, C_j)| \tag{6}$$

  onde *D()* é a função distância, e α_i, α_j, β e γ são coeficientes cujos valores dependem do esquema utilizado.

  A fórmula descreve a distância entre um agrupamento l e um novo agrupamento formado pela fusão dos dois agrupamentos i e j. Observe que quando $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ e $\gamma =$ -1/2, a fórmula se torna:

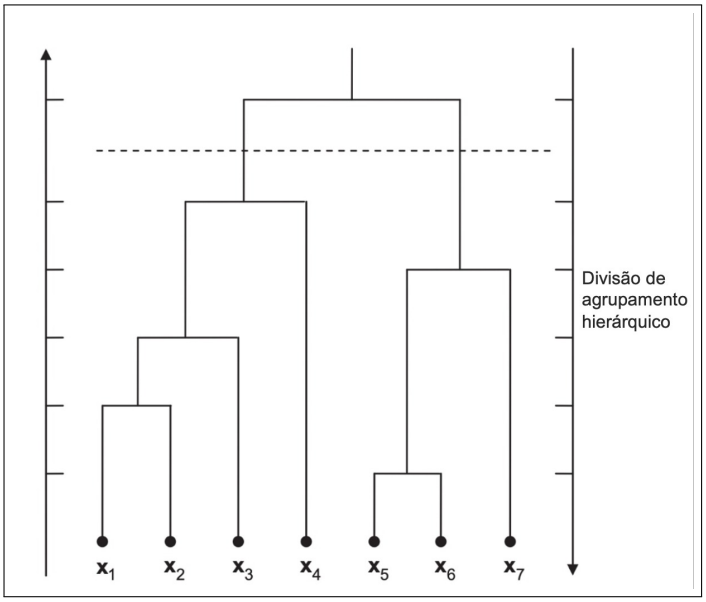
$$D(C_l, (C_i, C_j)) = min(D(C_l, C_i), D(C_l, C_j)) \tag{7}$$

que corresponde ao método de ligação simples. Quando $\alpha_i = \alpha_j = \gamma = 1/2$ e $\beta = 0$, a fórmula é:

$$D(C_l, (C_i, C_j)) = max(D(C_l, C_i), D(C_l, C_j)) \tag{8}$$

que corresponde ao método de ligação completo.

Figura 8 - Método hierárquico de agrupamento.



Legenda: Exemplo do funcionamento do método hierárquico de agrupamento.
Fonte: Adaptado de (LAM; WUNSCH, 2014)

- **Método de partição**

Ao contrário do agrupamento hierárquico, o agrupamento particionado atribui dados em K agrupamentos sem qualquer estrutura hierárquica, otimizando uma função de critério, sendo uma delas o critério de soma quadrada. Supondo que seja desejado organizar um conjunto de objetos $\mathbf{x}_j \in R^d$, j = 1,..,N, em K subconjuntos $C = C_1, ..., C_K$. O critério é então definido como:

$$J(U, M) = \sum_{i=1}^{k} \sum_{j=1}^{k} u_{ij} ||x_j - m_i||^2, \tag{9}$$

onde $m = [m_1, ..., m_k]$ é a matriz (média) do protótipo do agrupamento. $U = u_{ij}$ é uma matriz de partição, $u_{ij} \in [0, 1]$ e $\sum_{i=1}^{K} uij = 1$. u_{ij} é a associação de dados

x_j no agrupamento C_i. Em agrupamento **hard**, $u_{ij} = 0$ ou 1, como é o caso do conhecido *k-means*. No agrupamento particionado, u_{ij} pode ter qualquer valor de 0 a 1, como é o caso do "Fuzzy C-means"(FCM, Confuso C-means).

2.2.2 Medidas de Similaridade

É utilizado o conceito de distância entre o objeto a ser classificado e os objetos do conjunto de treinamento mais próximos a ele. A medida mais usada é a distância Euclidiana. As medidas de distância consideram a similaridade como a proximidade de observações umas com as outras. As medidas de distância são, na verdade, uma medida de dissimilaridade, com valores maiores denotando menor similaridade. A distância é convertida em uma medida de similaridade pelo uso de uma relação inversa (HAIR et al., 2009).

Há outras medidas de distância que também podem ser utilizadas, além da medida euclidiana, abaixo seguem conceitos de outras cinco medidas que podem ser utilizadas:

2.2.2.1 Distância Euclidiana:

É a medida mais utilizada, também conhecida como distância em linha reta. Suponha que dois pontos em duas dimensões tenham coordenadas ($\mathbf{X}_1$, $\mathbf{Y}_1$) e ($\mathbf{X}_2$, $\mathbf{Y}_2$), respectivamente. A distância em linha reta, ou distância euclidiana, entre os pontos é o comprimento da hipotenusa de um triângulo retângulo (MANLY; ALBERTO, 2008), conforme se calcula pela fórmula abaixo:

$$d = \sqrt{(X_2 - X_2)^2 + (Y_2 - Y_1)^2} \tag{10}$$

Outra distância euclidiana é a distância euclidiana elevada ao quadrado, que é a soma dos quadrados das diferenças sem calcular a raiz quadrada. A distância euclidiana elevada ao quadrado tem a vantagem de que não é necessário calcular a raiz quadrada, o que acelera o tempo de computação.

2.2.2.2 Distância *Manhattan*:

É uma disância que emprega a soma das diferenças absolutas das variáveis (isto é, os dois lados de um triângulo retângulo em vez da hipotenusa). Este procedimento é o mais simples de calcular, mas pode conduzir a agrupamentos inválidos se as variáveis forem altamente correlacionadas(QUEIROZ; PINTO, 2014). Calcula-se pela seguinte fórmula:

$$d(x, y) = \sum_{i=1}^{k} |x_i - y_i| \tag{11}$$

2.2.2.3 Distância de *Chebychev*:

A distância é a maior diferença ao longo de todas as variáveis de agrupamento, também conhecida como distância ao máximo (PALMA, 2018). Ela é particularmente suscetível a diferenças em escalas ao longo das variáveis e se calcula da seguinte forma:

$$d(x, y) = max_i(|x_i - y_i|) \tag{12}$$

2.2.2.4 Distância de *Mahalanobis*:

É uma medida generalizada de distância baseada nas correlações entre variáveis de uma maneira que pondera igualmente cada uma delas(MANLY; ALBERTO, 2008). Ela também depende de variáveis padronizadas e pode ser calculada da seguinte maneira:

$$D^2 = (x - m)^T C^{-1} (X - M) \tag{13}$$

2.2.3 Algoritmos de agrupamento

É um grande desafio decidir qual método de agrupamento utilizar na condução de um experimento. Um outro problema que enfrentamos é determinar o número de agrupamentos apropriados para os dados utilizados. Devemos analisar o número ideal

de agrupamentos não somente pelas propriedades estatísticas, mas também os resultados quimicamente relevantes, de acordo com a base de dados em questão.

Os algoritmos de agrupamento são habitualmente divididos em Hierárquicos e em Particionais. Uma excelente revisão sobre métodos de agrupamentos pode ser encontrada no artigo Data clustering: a review (JAIN; MURTY; FLYNN, 1999).

Para este trabalho, foram utilizados nove algoritmos, já citados anteriormente: k-means, sota, clara, diana, pam, fanny, hierarquical, model. Uma breve descrição de cada método de agrupamento é apresentada abaixo:

2.2.3.1 *k-means*

É um algoritmo de aprendizado não supervisionado que agrupa n dados em k-grupos. É um método que busca minimizar a distância ao quadrado de cada observação ao centro e solução do problema de agrupamento segundo o critério de minimização da soma de distâncias do grupo ao qual pertence, dado por $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$ de forma iterativa. A distância entre um ponto p_i e um conjunto de grupos, dada por $d(p_i, X)$, é definida como sendo a distância do ponto ao centro mais próximo dele (LINDEN, 2009).

O algoritmo *k-means* depende de um parâmetro k (número de agrupamentos) definido inicialmente. Isto costuma ser um problema, tendo em vista que normalmente não se sabe quantos grupos existem a priori, afetando a performance do algoritmo.

O algoritmo *k-means* básico tem complexidade computacional de O(nk) (FADEL; SEMAAN; BRITO, 2014).

2.2.3.2 *Pam - Partitioning Around Medóides*

No início dos anos 80, muitos métodos de agrupamento eram limitados para lidar com dados métricos, ou seja, coordenadas de pontos geométricos. O algoritmo *Pam* foi desenvolvido naquela época (mas apenas publicado posteriormente). O *Pam* fazia parte de um projeto para construir métodos de agrupamentos que poderiam lidar com matrizes de dissimilaridade arbitrárias (SCHUBERT; ROUSSEEUW, 2021).

Este algoritmo realiza o particionamento em torno de *medoids* - objetos representativos de um conjunto de dados cuja soma de dissimilaridades para todos os objetos no grupo é mínima - semelhante ao *K-means*, mas é considerado mais robusto porque admite o uso de outras dissimilaridades além da mínima soma de distância ao quadrado. Como *K-means*, o número de agrupamentos é fixado antecipadamente e um conjunto inicial de

centros de agrupamentos é necessário para iniciar o algoritmo (BROCK et al., 2008).

Formalmente, é definido da seguinte maneira:

Seja o conjunto **X** com n objetos:

$$X = x_1, x_2, ..., x_n \tag{14}$$

A partir deste conjunto, inicialmente, deve-se escolher k objetos, a fim de formar o conjunto de medoids:

$$M = medoid_1, medoid_2, ..., medoid_k \tag{15}$$

Posteriormente, cada um dos *(n-k)* objetos restantes são alocados ao grupo cuja distância do objeto ao medoid seja mínima. Este processo deve ser repetido até que não haja mudanças significativas no valor da função abaixo (SILVA; BRITO; OLIVEIRA,):

$$\sum_{i=1}^{K} \sum_{\forall xj \in med_j} d(x_i, x_j) \tag{16}$$

Em síntese, o algoritmo é dividido nas seguintes etapas:

- **Selecionar, dentre os *n* objetos disponíveis de X, os *k* objetos que definem um conjunto *M* de medoids;**
- **Associar os objetos restantes ao seu medoid mais próximo;**
- **Substituir os *k* medoids de forma a minimizar a soma das distâncias dos objetos ao seu medoid mais próximo;**
- **Repetir os passos 2 e 3 até que não haja nenhuma mudança significativa nas posições dos medoids.**

Este algoritmo é uma particularidade do método de agrupamento *k-medoids* (Han et al., 2012) e possui complexidade computacional $O(K(n-k)^2)$ (FADEL; SEMAAN; BRITO, 2014).

2.2.3.3 *Clara - Clustering Large Applications*

É indicado para grandes conjuntos de dados, pois utiliza um subconjunto dos dados e aplica o algoritmo *Pam* sobre ele, visando reduzir os requisitos de memória e de capacidade de processamento. Desta forma, é possível utilizar apenas uma amostra

da base de dados no qual o *Pam* é aplicado para selecionar os medóides, permitindo tempos de execução mais rápidos quando o número de observações é relativamente grande (RODRIGUES, 2009). Por conta disso, Kaufman e Rousseeuw (1986) introduziram o método *Clara*, com a ideia central de trabalhar com amostras menores do conjunto inicial. Esses subconjuntos de dados têm um tamanho fixo, quanto menor o tamanho, menor a complexidade (XAVIER, 2012).

Em outras palavras, podemos resumir o algoritmo *Clara* em um processo de duas fases: *Build* e *Swap*. Na primeira, um subconjunto pré-definido é retirado do conjunto de dados inicial e agrupado em k grupos, usando o algoritmo *Pam*. Então, o algoritmo *Clara* seleciona *medoids* sucessivos com o intuito de alcançar a menor distância entre os objetos do subconjunto. A segunda fase, conhecida como *Swap*, é uma tentativa de melhorar o conjunto de objetos representativos, assim como melhorar os agrupamentos obtidos por este conjunto de objetos.

2.2.3.4 *Fanny - Fuzzy Analysis*

Este algoritmo realiza agrupamento difuso, onde cada observação pode ter pertinência parcial em cada agrupamento (Kaufman e Rousseeuw 1990). Assim, cada observação tem um vetor que dá a adesão parcial a cada um dos agrupamentos. Um agrupamento rígido pode ser produzido atribuindo cada observação ao agrupamento onde ela tem a maior associação (BROCK et al., 2008).

O algoritmo FANNY (Kaufman e Rousseeuw 1990) foi um dos primeiros métodos de agrupamento difuso propostos (KAUFMAN; ROUSSEEUW, 1990). Este método minimiza a função objetivo descrita pela Equação 17:

$$j = \sum_{c=1}^{K} \frac{\sum_{i=1}^{N} u_{ic}^2 \sum_{j=1}^{N} u_{jc}^2 D(i,j)}{2 \sum_{j=1}^{N} u_{jc}^2} \tag{17}$$

Tendo uma inicialização aleatória das pertinências, respeitando a restrição de soma para as pertinências:

$$\sum_{c=1}^{K} u_{ic} = 1, \forall i = 1, 2, ..., N \tag{18}$$

com os passos de otimização obtidos com o método de multiplicadores de Lagrange, sendo γ_i os multiplicadores de Lagrange (KAUFMAN; ROUSSEEUW, 1990), temos:

$$l_{fanny} = \sum_{c=1}^{K} \frac{\sum_{i=1}^{N} u_{ic}^2 \sum_{j=1}^{N} u_{jc}^2 D(i,j)}{2\sum_{j=1}^{N} u_{jc}^2} - \sum_{i=1}^{N} \gamma_i (\sum_{c=1}^{K} u_{ic} - 1) \quad (19)$$

Pelo fato do modelo ser simples, as únicas entradas necessárias são a matriz de dissimilaridades e o número de agrupamentos, sem parâmetros adicionais (KAUFMAN; ROUSSEEUW, 1990).

O algoritmo *fanny* não possui um vetor específico para a ponderação das representações dos agrupamentos (u_{cj}), dessa forma, foi utilizando o próprio vetor de pertinências como a ponderação das representações (KAUFMAN; ROUSSEEUW, 1990).

2.2.3.5 Hierárquico

Os algoritmos hierárquicos podem ser aglomerativos ou divisivos. No aglomerativo ele produz um dendrograma que pode ser cortado a uma altura escolhida para produzir o número desejado de ramificações. Cada observação é inicialmente colocada em seu próprio agrupamento, e os agrupamentos são unidos sucessivamente em ordem de sua "proximidade". A proximidade de quaisquer dois agrupamentos é determinado por uma matriz de dissimilaridade, e pode ser baseado em uma variedade de métodos de aglomeração (BROCK et al., 2008). Ou seja, o aglomerativo opera criando conjuntos a partir de elementos isolados.

Em síntese, o algoritmo aglomerativo é dividido nas seguintes etapas (LINDEN, 2009).:

- **Gerar um agrupamento para cada elemento;**
- **Encontrar os pares de agrupamentos mais similares, de acordo com a medida de dissimilaridade escolhida;**
- **Fundir em um agrupamento maior e recalcular a distância deste agrupamento para todos os outros elementos;**
- **Repetir passos 2 e 3 até sobrar apenas um agrupamento.**

A estrutura hierárquica formada pela união entre os elementos é representada por meio de um dendrograma, que é a forma mais usual de representação dos resultados de algoritmos hierárquicos e mostra a ordem do agrupamento. Quanto mais alta a linha ligando dois grupos, mais tarde foi feito seu agrupamento. Logo, a altura da linha ligando dois grupos é proporcional à sua distância (LINDEN, 2009).

Diversas maneiras diferentes de medir a distância entre dois grupos foram propostas na literatura: o *Single link*, que é um dos algoritmos hierárquicos mais simples que

existe, pois ele usa a técnica do vizinho mais próximo, onde a distância entre dois grupos é determinada pela distância do par de exemplos mais próximo, sendo cada exemplo pertencente a um desses grupos. O *Complete link*, que emprega a técnica do vizinho mais distante. Diferente do algoritmo *Single link*, esse método determina a distância entre dois grupos de acordo com a maior distância entre um par de exemplos, sendo cada exemplo pertencente a um grupo distinto (METZ, 2006).

O *Average link*, onde a distância é dada pela distância entre seus centroides. O Centroid-based, a distância entre os agrupamentos é definida em termos da distância no espaço euclidiano entre o representante de cada agrupamento (METZ, 2006). O *Ward*, segundo (HAIR et al., 2009), é uma técnica onde a medida de similaridade é calculada como a soma dos quadrados entre os dois agrupamentos em relação a todo o conjunto de dados. Com isto, os agrupamentos tendem a ter tamanhos iguais.

Os algoritmos descritos acima são os algoritmos mais tradicionais da literatura de agrupamento hierárquico. Porém, existem outros algoritmos elaborados a partir das ideias implementadas nos algoritmos clássicos, como por exemplo, o AGNES - Agglomerative Nesting (Kaufman e Rousseeuw, 1990a), o MONA - Monothetic Analysis (Kaufman e Rousseeuw, 1990d), entre outros (METZ, 2006).

Já a alternativa divisiva do algoritmo hierárquico de agrupamento, inicia-se com um único agrupamento contendo todos os objetos, dividindo-os em seguida. Em cada iteração, utiliza-se um algoritmo *flat* (como o K-Means, por exemplo), para separar o conjunto corrente em grupos menores, repetindo-se o processo recursivamente até que se tenha apenas conjuntos compostos de um único elemento ou até que um critério de parada seja atendido (LINDEN, 2009). Ou seja, a divisiva começa com um grande conjunto e vai quebrando-o em partes até chegar a elementos isolados (LINDEN, 2009).

2.2.3.6 *Diana - Divisive Analysis*

É um algoritmo hierárquico divisivo que inicialmente começa com todas as observações em um único grupo e divide sucessivamente os agrupamentos até que cada grupo contenha uma única observação (BROCK et al., 2008). Possui complexidade $O(N^2 logN)$.

No processo iterativo orientado pelo *Diana*, o k-ésimo agrupamento produzido possui k grupos em que $k = 1, ..., n$, onde n é a quantidade de dados. Os autores (NIETTO; SAMPAIO,) exemplificam que considerando um grupo $\mathbf{X}$, este grupo é dividido em dois subconjuntos de tal maneira que os dois grupos resultantes possuam a maior dissimilaridade possível entre eles.

Inicialmente o algoritmo busca identificar um dado do grupo $\mathbf{X}$ cuja dissimilaridade média em relação aos dados restantes seja máxima. O dado com dissimilaridade máxima é retirado do grupo $\mathbf{X}$ e inserido em um novo grupo criado nesse momento, chamado de

tempX.

Na sequência, para cada dado $\mathbf{x} \in \mathbf{X}$, é calculada a média dos valores de dissimilaridade de $\mathbf{x}$, com todos os demais dados de $\mathbf{X}$, ou seja, média dos valores diss($\mathbf{x}$,$\mathbf{y}$), $y \in X$. De maneira análoga, é calculada a média dos valores de dissimilaridade de $\mathbf{x}$ com os dados pertencentes a *tempX*, ou seja, a média dos valores diss($\mathbf{x}$, $\mathbf{z}$), $\mathbf{z} \notin \mathbf{X}(i.e, z \in tempX)$.

Se para cada $\mathbf{x} \in \mathbf{X}$: D($\mathbf{x}$) = (média diss($\mathbf{x}$, $\mathbf{y}$), $\mathbf{y} \in \mathbf{X}, \mathbf{x} \neq \mathbf{y}$) - (média diss($\mathbf{x}$, $\mathbf{z}$), $\mathbf{z} \in$ *tempX*) para negativa, então *tempX* não receberá mais nenhum dado de $\mathbf{X}$.

Caso contrário, o dado $\mathbf{x} \in \mathbf{X}$ que produzir o valor máximo para a diferença D($\mathbf{x}$) = (média diss($\mathbf{x}$,$\mathbf{y}$), $\mathbf{y} \in \mathbf{X}$, $\mathbf{x} \neq \mathbf{y}$) - (média diss($\mathbf{x}$,$\mathbf{z}$), $\mathbf{z} \in$ *tempX*) é escolhido e retirado de $\mathbf{X}$ e, então, inserido em *tempX*. O procedimento é repetido até que cada dado seja o único em um grupo ou, então até que um critério de parada seja satisfeito (NIETTO; SAMPAIO,).

2.2.3.7 *Som - Self-Organizing Map*

Faz parte de uma classe especial de redes neurais artificiais não supervisionadas. Essas redes são formadas por grades de neurônios de uma ou de duas dimensões que modificam seus pesos sinápticos em um processo de aprendizagem competitivo, formando sobre a grade de saída um sistema de coordenadas significativas para diferentes características de entrada (MARTINS; GUIMARÃES; FONSECA, 2002). Este algoritmo é muito indicado por sua capacidade de mapear e visualizar dados de alta dimensão em duas dimensões (BROCK et al., 2008).

2.2.3.8 *Sota - State-Of-The-Art*

O algoritmo de árvore auto-organizado é uma rede não supervisionada com uma hierarquia de estrutura de árvore binária. É adequado para agrupar muitos objetos (BROCK et al., 2008).

A diferença desse algoritmo para os de agrupamento hierárquico, que seguem uma estratégia aglomerativa, é que o método SOTA segue o método divisivo, onde o processo de agrupamento começa com uma árvore binária formada por um nó principal e duas folhas. Esse método combina a estrutura de árvore do agrupamento hierárquico com a da rede neural utilizado no agrupamento SOM. Assim como o SOM, o SOTA é um algoritmo não determinístico (MÁRQUEZ; SABATÉ; CASADO, 2011).

2.2.3.9 *Model*

É um modelo estatístico que consiste em uma mistura finita de distribuições gaussianas do qual é ajustado aos dados. Cada componente da mistura representa um agrupamento, e os componentes da mistura e os membros do grupo são estimados usando o método de máxima verossimilhança (BROCK et al., 2008).

2.2.4 Medidas de validação

Quando se aplica os algoritmos de agrupamento, tanto a medida de dissimilaridade quanto os parâmetros escolhidos pelo algoritmo podem influenciar na qualidade dos agrupamentos gerados. Diante disto, é muito necessário avaliar, a posteriori, a qualidade dos agrupamentos. De modo geral, esta qualidade pode ser verificada com base em índices estatísticos. Estes índices estão associados ao que a literatura chama de critérios de validação (SOARES; OLIVEIRA; BRITO, 2014).

Diversos autores propuseram uma variedade de medidas destinadas a validar os resultados de uma análise de agrupamento e determinar qual algoritmo de agrupamento tem o melhor desempenho para um experimento específico.

Um tipo de medida de validação importante para avaliar o desempenho dos métodos de agrupamento é a classe de medida de validação interna. Esta estratégia usa apenas o conjunto de dados e a partição do grupo como entrada e usam informações intrínsecas nos dados para avaliar a qualidade do agrupamento (BROCK et al., 2008). Para esta dissertação foram utilizadas as medidas de validação interna com os indicadores índice de Dunn (DUNN, 1974) e o valor da silhueta (ROUSSEEUW, 1987). Estes métodos são ambos os exemplos de combinações não lineares da compacidade e separação (BROCK et al., 2008).

Medidas internas:

Para a validação interna, utilizou-se os indicadores de índice de Dunn e valor da silhueta.

Valor da silhueta: Existem diversos índices de validação de agrupamentos que podem ser utilizados junto com os métodos de agrupamento para encontrar um valor ótimo de grupos. A silhueta é um dos mais usados, pois é uma medida intuitiva e simples que não depende de suposições de modelos estatísticos (BATOOL; HENNIG, 2021).

O valor da silhueta mede o grau de confiança na atribuição de agrupamento de uma observação particular. Se as observações possuem valores próximos de 1, são bem agrupadas, mas o pior ocorre quando as observações possuem valores próximos de -1, pois significa que são mal agrupadas (ROUSSEEUW, 1987). Para a observação i, ela é definida

como:

$$S_i = \frac{b_i - a_i}{max(b_i, a_i)} \tag{20}$$

onde a_i é a distância média entre i e todas as outras observações no mesmo agrupamento, e b_i é a distância média entre i e as observações no "grupo vizinho mais próximo", ou seja:

$$a_i = \frac{1}{n(C_{(i)})} \sum_{j \in C_{(i)}} dist(i, j) \tag{21}$$

$$b_i = min_{Ck \in ConnCi} \sum_{j \in C_k} \frac{dist(i, j)}{n(C_k)} \tag{22}$$

onde $\mathbf{C}_{(i)}$ é o agrupamento contendo a observação i, dist(**i**, **j**) é a distância (por exemplo, Euclidiana, Manhattan, já apresentadas anteriormente) entre as observações i e j, e $n(C)$ é a cardinalidade do agrupamento C. A largura da silhueta, portanto, está no intervalo [-1, 1], e deve ser maximizada (BROCK et al., 2008).

Índice de Dunn: O índice de Dunn é a razão da menor distância entre observações que não estão no mesmo agrupamento para a maior distância intra-cluster (BROCK et al., 2008). É calculado como:

$$D(Conn) = \frac{min_{Ck,Cl \in Conn, C_k \neq C_l}(min_{i \in C_k, j \in C_l} dist(i, j))}{max_{Cm \in Conn} diam(C_m)} \tag{23}$$

2.3 Método Escalonamento Multidimensional Local

A redução da dimensionalidade é muito importante para facilitar a visualização de dados multidimensionais. O problema da redução da dimensionalidade pode ser explicado da seguinte forma: seja $X = x_1, x_2, ...x_n$ um conjunto de n pontos (vetores) de um espaço multidimensional, ou seja, $x_i \in R^m$, a redução de dimensionalidade busca encontrar um conjunto de dados equivalente de pontos de saída $Y = y_1, y_2, ...y_n$, de forma que $y_i \in R^p$, sendo $p \leq m$ e Y, a representação mais fiel possível de X no espaço de baixa dimensão. (MEDEIROS; COSTA, 2008).

Em outras palavras, um dos principais objetivos dos métodos de redução de dimensionalidade é encontrar um subconjunto de variáveis, baseado nas variáveis originais, de forma que as informações sejam preservadas. Um exemplo de redução de dimensionalidade é o Escalonamento multidimensional (MDS), é um método de redução de dimensionalidade da classe de escalonamento multidimensional métrico e possui a característica de ser não diferenciável (XAVIER et al., 2018).

O método MDS utiliza o critério de preservação de dissimilaridade. Através deste método, busca-se minimizar as distorções entre as distâncias ou as dissimilaridades medidas entre as observações no espaço original de alta dimensão e as distâncias medidas no espaço de baixa dimensão (XAVIER et al., 2018).

Um outro método pertencente a classe de métodos MDS, é o método de Escalonamento Multidimensional Local (*Local MDS*), em que são feitas transformações sobre os dados projetando-os em espaços de menor dimensão buscando preservar as relações de distância, ou seja, o método prioriza preservar as observações próximas (vizinhanças). Além disso, o valor do parâmetro k utilizado no *Local MDS* preserva as K observações vizinhas e afasta as observações não vizinhas.

Desta forma, o método *Local MDS* consiste em obter um novo conjunto de observações, $\mathbf{x}_i; i = 1, \ldots, n$, pertencentes ao espaço de menor dimensão d, ou seja, $\mathbf{x}_i \in R^d$, de modo que gere uma matriz de distâncias, que se aproxime da melhor forma à matriz de distâncias $D_(n \text{ x } n)$, tendo com base a seguinte função de escalonamento a ser minimizada (XAVIER et al., 2018):

$$f_{(LocalMDS)}(x) = \sum_{(i,j)\in N} (D_{ij} - ||x_i - x_j||)_2)^2 - t \sum_{(i,j)\notin N} ||x_i - x_j||_2 \tag{24}$$

O problema intrinsecamente não diferenciável é aproximado por um continuamente diferenciável, denominado Escalonamento Multidimensional Local Suavizado (EMLS).

$$f_{EMLS}(x, \gamma_1, \gamma_2) = \sum_{(i,j)\in N} (D_{ij} - \theta(||x_i - x_j||_2, \gamma_1))^2 - t \sum_{(i,j)\notin N} \theta(||x_i - x_j||_2, \gamma_2) \tag{25}$$

onde a suavização da distância euclidiana é dada por $\theta(u, \gamma)$:

$$\theta(u, \gamma) = \sqrt{(u^2 + \gamma^2)} \tag{26}$$

É utilizado o critério Continuidade Local *(Local continuity ou LC meta-criterion)* a fim de comparar o método *Local MDS* com outros métodos. Este critério utiliza a relação de vizinhança nos espaços de alta dimensão e de baixa dimensão, tendo como ideia central

a interseção dos K vizinhos mais próximos *K-NN* de uma observação no espaço de alta dimensão e os *K-NN* no espaço de baixa dimensão (XAVIER et al., 2018).

Conforme os autores (CHEN; BUJA, 2009), o critério Continuidade Local é baseado nas três definições a seguir:

1. $N_k^s(i) = j_1..., j_k$, conjunto dos $k^{'}$ vizinhos mais próximos da observação i com respeito às observações no espaço de alta dimensão;

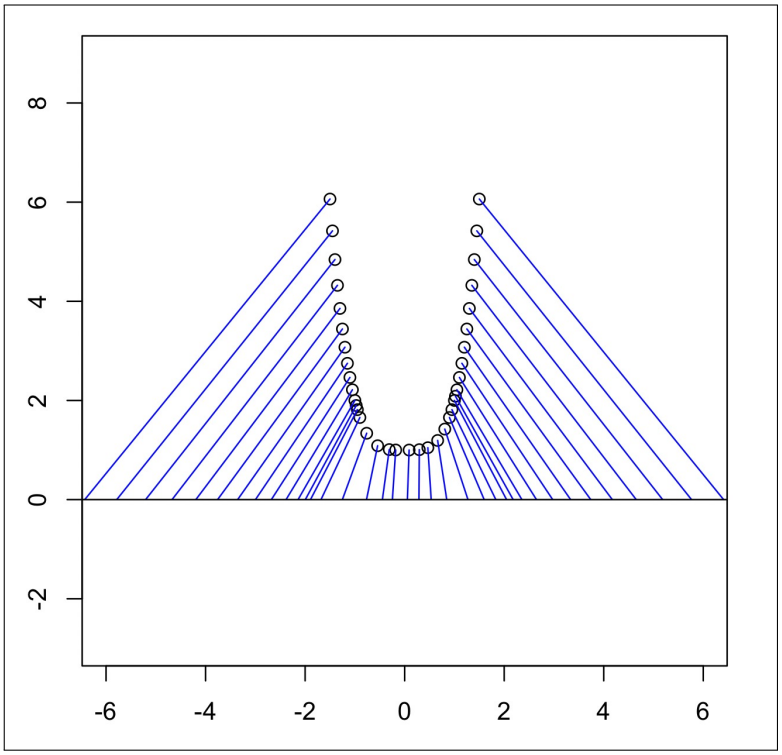
2. $N_k^d(i) = k_1, ..., k_k$, conjunto dos $k^{'}$ vizinhos mais próximos da observação i com respeito às observações no espaço de baixa dimensão;

3. $N_k(i) = |N_K^S(i) \cup N_K^d(i)|$, cardinalidade da interseção dos dois conjuntos anteriores.

Quanto maior for a interseção entre os dois conjuntos $N_k^s(i)$ e $N_k^d(i)$, maior será a preservação das vizinhanças. Ou seja, a projeção no espaço de baixa dimensão terá uma maior preservação das vizinhanças no espaço de alta dimensão.

Um exemplo prático de redução de dimensionalidade com o método *Local MDS*, com parâmetro *k=3*, é a curva *u-shaped*, projetada em um plano R^1, apresentado no gráfico da Figura 9. Avaliando a projeção com a métrica continuidade local, com parâmetro *k=3*, pode-se ver que os três vizinhos mais próximos foram preservados na projeção.

Figura 9 - Representação gráfica da curva *u-shaped*.



Legenda: Representação gráfica da curva *u-shaped* projetada para espaço dimensional R^1.

Fonte: A autora, 2022.

3 BANCO DE DADOS

O Banco de dados utilizado é composto por dados referentes ao processo de biomassa, fornecido por (GUEDES; LUNA; TORRES, 2018). A biomassa é considerada como tendo potencial para ser utilizada como fonte alternativa de energia. O bio-óleo obtido da pirólise da biomassa é utilizado como combustíveis e produtos químicos.

O rendimento e a composição do óleo de pirólise dependem da composição da biomassa e dos parâmetros operacionais do processo. Este banco de dados foi criado com foco em parâmetros que influenciam o processo, como temperatura, tempo de reação, taxa de aquecimento, vazão de gás, taxa de alimentação, tamanho de partícula e composição de biomassa e discute o efeito desses parâmetros no rendimento e qualidade de bio-óleo.

Este banco de dados foi desenvolvido a partir de dados experimentais obtidos de mais de 200 pesquisas de pirólise de biomassa disponíveis na literatura. A base de dados inclui os artigos mais citados na literatura, desde 1984, contendo dados experimentais, bem como pesquisas recentes na área (GUEDES; LUNA; TORRES, 2018).

A base de dados continha diferentes processos de pirólise e diferentes condições de operação e foi construída para entender melhor como esses fatores influenciam a composição e o rendimento dos produtos da pirólise.

Os tipos de pirólise são agrupados nas respectivas classes de acordo com as variáveis químicas descritas na Tabela 1:

Tabela 1 - Variáveis que agrupam os tipos de pirólise.

Variável	Descrição	Número
p_c_mp	Porc. carbono em base seca livre de cinza na matéria-prima	1
p_h_mp	Porc hidrogênio em base seca livre de cinza na matéria-prima	2
p_n_mp	Porc. nitrogênio em base seca livre de cinza na matéria-prima	3
p_o_mp	Porc. oxigênio em base seca livre de cinza na matéria-prima	4
p_umid_mp	Porc. umidade na matéria-prima	5
p_cfix_mp	Porc. carbono fixo em base seca da matéria-prima	6
p_cinz_mp	Porc. cinzas em base seca da matéria-prima	7
p_vola_mp	Porc. voláteis em base seca da matéria-prima	8
tp_med_reator	Tamanho da partícula média no reator	9
t_res_vap_reator	Tempo de residência médio no reator para o gás e arraste	10
temp_reator	Tempreatura de operação do reator	11
rend_gp	Rendimendo em gás - porcentagem	12
tipo_piro_reator	Tipo de pirólise (classes)	13

Legenda: Subconjuntos das variáveis que compõem o banco de dados e a sua respectiva numeração associada para citações ao longo do texto.

Fonte: (GUEDES; LUNA; TORRES, 2018)

Para este estudo é considerado os casos de regime de operação do reator igual a Batelada e o tipo do reator igual a **Leito Fixo** que contemplam os tipos de pirólise: *Fast*, *Slow* e *Catalytic*. No nível de comparação, também foram considerados os casos de regime de operação do reator igual a Contínuo e o tipo do reator igual a **Leito Fixo** que também contemplam os tipos de pirólise: *Fast*, *Flash* e *Catalytic*.

Para o regime de operação do reator igual a Batelada, a amostra utilizada possui 246 observações, sendo 165 para tipo de pirólise *Fast*, 71 para tipo de pirólise *Slow* e 10 para tipo de pirólise *Catalytic* .

Já para o regime de operação do reator Contínuo, a amostra utilizada possui 171 observações, sendo 135 para tipo de pirólise *Fast*, 35 para tipo de pirólise *Flash* e 1 para tipo de pirólise *Catalytic*. Neste caso, os resultados que envolvem toda a volumetria possuem menos influência por parte da classe *Catalytic* e tem um peso maior para as classes *Fast* e *Flash*. Por isso, após diversos experimentos, a classe *Catalytic* foi excluída da amostra para o tipo de reator Contínuo pelo fato de possuir apenas uma amostra. Para uma descrição mais detalhada do conjunto de dados sugerimos a leitura de (GUEDES; LUNA; TORRES, 2018).

4 ANÁLISE EXPLORATÓRIA DOS DADOS

Em toda a análise de dados foi utilizado o programa R, livre e *open source* (TEAM, 2016). De forma geral, um conjunto de dados que aborda análises químicas envolve um determinado número de amostras, descritas por determinado número de variáveis. Antes de ir para as ferramentas de fato e para as aplicações dos algoritmos, é necessário compreender a natureza dos dados, possíveis medidas de posição, dispersão e distribuições.

4.1 Análise do subconjunto do tipo de reator Batelada

Na Tabela 2 são apresentadas as seguintes estatísticas descritivas das variáveis: valor mínimo, primeiro quartil (Q1), mediana, média, terceiro quartil (Q3), valor máximo e variância. As variáveis 3, 5, 7, 3 e 9 tem suas observações bem próximas a zero no valor mínimo. Já as variáveis 10 e 11 possuem os maiores valores máximos. A média está abaixo da mediana no caso das variáveis 4, 5, 6 e 8, indicando uma assimetria negativa, ou seja, a cauda da curva da distribuição declina para esquerda.

Tabela 2 - Variáveis numéricas e suas respectivas medidas (Batelada).

Variável	Mínimo	Q1	Mediana	Média	Q3	Máximo	Variância
1	39,98	49,74	51,65	52,76	55,89	67,49	31,47
2	4,32	6,09	6,50	6,76	7,58	9,36	2,29
3	0,08	0,62	2,05	2,46	3,89	9,29	5,14
4	22,07	28,25	40,86	37,83	42,62	55,25	67,90
5	0,44	5,70	7,50	7,48	8,38	21,00	5,90
6	6,87	10,81	15,71	14,99	17,62	32,03	20,67
7	0,83	2,33	4,65	5,38	6,14	22,55	18,18
8	60,84	76,37	80,97	79,62	85,68	90,78	37,99
9	0,11	0,64	0,92	1,35	1,16	15,00	3,86
10	3,00	24,00	24,00	317,40	113,10	7200,00	1133618
11	300,00	500,00	550,00	551,90	600,00	1000,00	15168.01
12	3,39	19,12	24,50	25,46	29,10	61,30	91,34

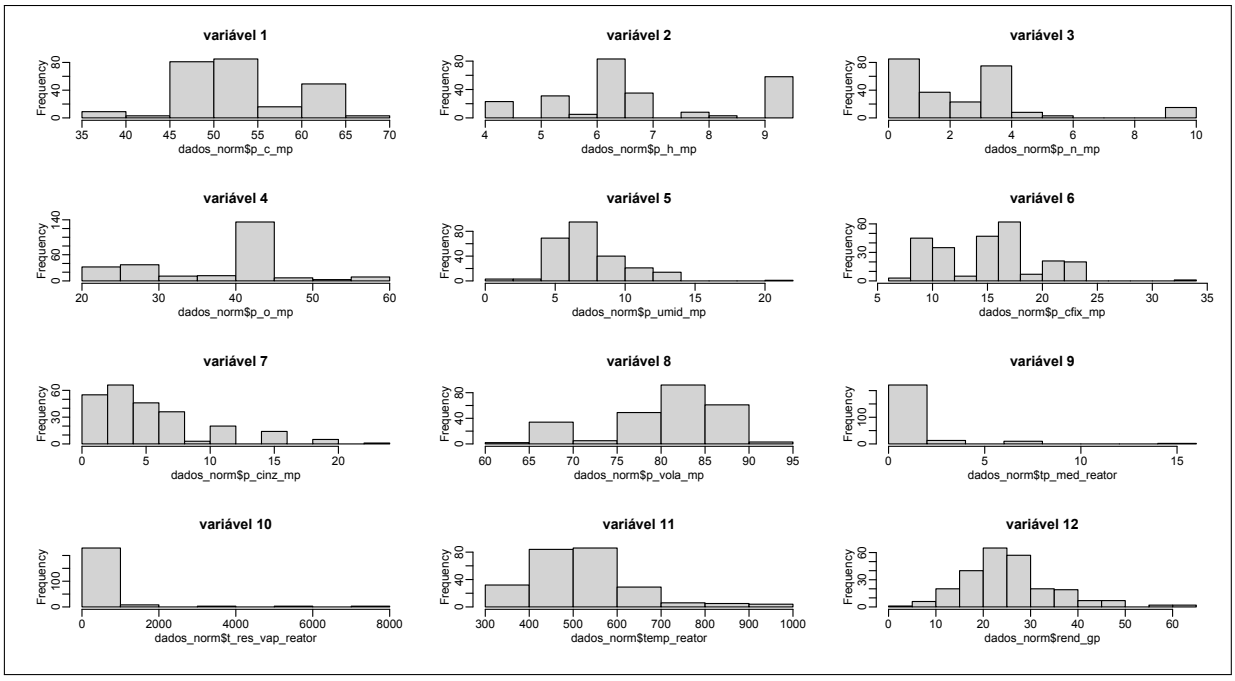
Legenda: Sumarização das variáveis numéricas do regime tipo de reator Batelada.
Fonte: A autora, 2022.

Na Figura 10 é possível analisar as distribuições da densidade dessas variáveis mais detalhadamente, as variáveis de número 5, 7, 9, 10, 11 e 12 apresentam assimetria à direita.

Com a compreensão de como estão distribuídos os valores das variáveis, pode-

se averiguar as relações entre as variáveis, assim como a interdependência entre elas. Na Figura 11 podemos visualizar pelo mapa de calor das correlações que as variáveis que estão dentro da escala de 1 a 0,7 (+ ou -), na tonalidade azul escuro (no caso das positivas) e na tonalidade vermelho escuro (no caso das negativas), possuem uma forte correlação. Já as variáveis que estão entre 0,7 a 0,5 (+ ou -) na tonalidade azul claro (no caso das positivas) e na tonalidade laranja (no caso das negativas), possuem correlação moderada. As variáveis que possuem escala de 0,5 a 0,25 (+ ou -) possuem baixa correlacão.

Figura 10 - Histograma das variáveis (Batelada).



Legenda: Histograma das variáveis numéricas do banco de dados de biomassa do regime tipo reator Batelada.

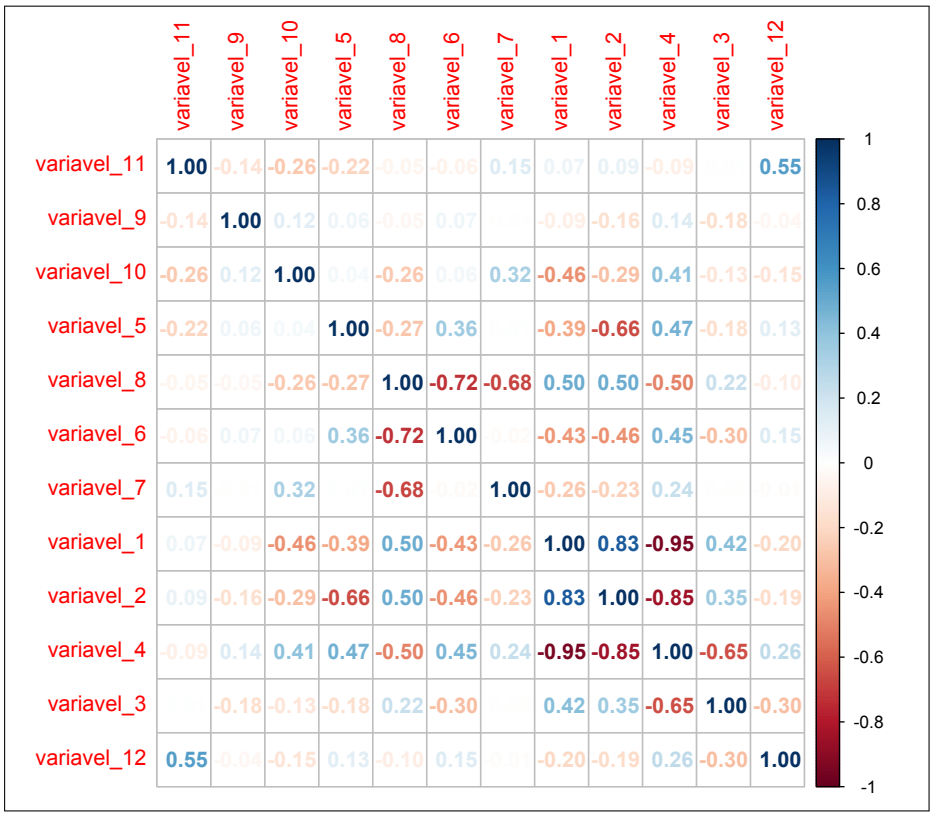
Fonte: A autora, 2022.

4.2 **Análise do subconjunto do tipo de reator Contínuo**

As mesmas observações foram geradas para o subconjunto do tipo reator Contínuo. Na Tabela 3, é perceptível que as variáveis 3, 5, 6, 7, 9 e 10 também possuem suas observações bem próximas a zero no valor mínimo. Já as variáveis 10 e 11 também possuem os maiores valores máximos. A média está abaixo da mediana no caso das variáveis 4, 5, 8 e 11, indicando uma assimetria negativa, ou seja, a cauda da curva da distribuição declina para esquerda.

Na Figura 12 é possível analisar as distribuições da densidade de cada uma das

Figura 11 - Correlação das variáveis (Batelada).



Legenda: Mapa de calor de correlação das variáveis numéricas do banco de dados de biomassa do regime tipo reator Batelada.

Fonte: A autora, 2022.

Tabela 3 - Variáveis numéricas e suas respectivas medidas (Contínuo).

Variável	Mínimo	Q1	Mediana	Média	Q3	Máximo	Variância
1	42,49	45,10	47,36	48,44	51,40	55,80	10,744
2	5,10	5,79	6,14	6,27	6,85	8,10	0,52
3	0,10	0,39	0,74	1,00	1,30	4,54	0,80
4	34,70	40,46	44,41	44,13	48,80	50,63	17,57
5	0,68	6,00	7,90	7,15	8,64	12,30	7,22
6	0,11	10,78	14,55	14,55	18,77	30,55	51,76
7	0,55	2,64	3,60	4,36	6,60	13,98	6,68
8	64,07	77,70	81,50	80,80	83,86	98,06	49,89
9	0,12	0,43	0,53	1,04	1,50	10,00	1,40
10	0,50	1,27	4,10	13,19	7,54	101,40	590,99
11	300,00	450,00	495,00	483,30	546,00	700,00	5222,07
12	5,00	14,50	19,00	20,84	24,90	60,00	75,829

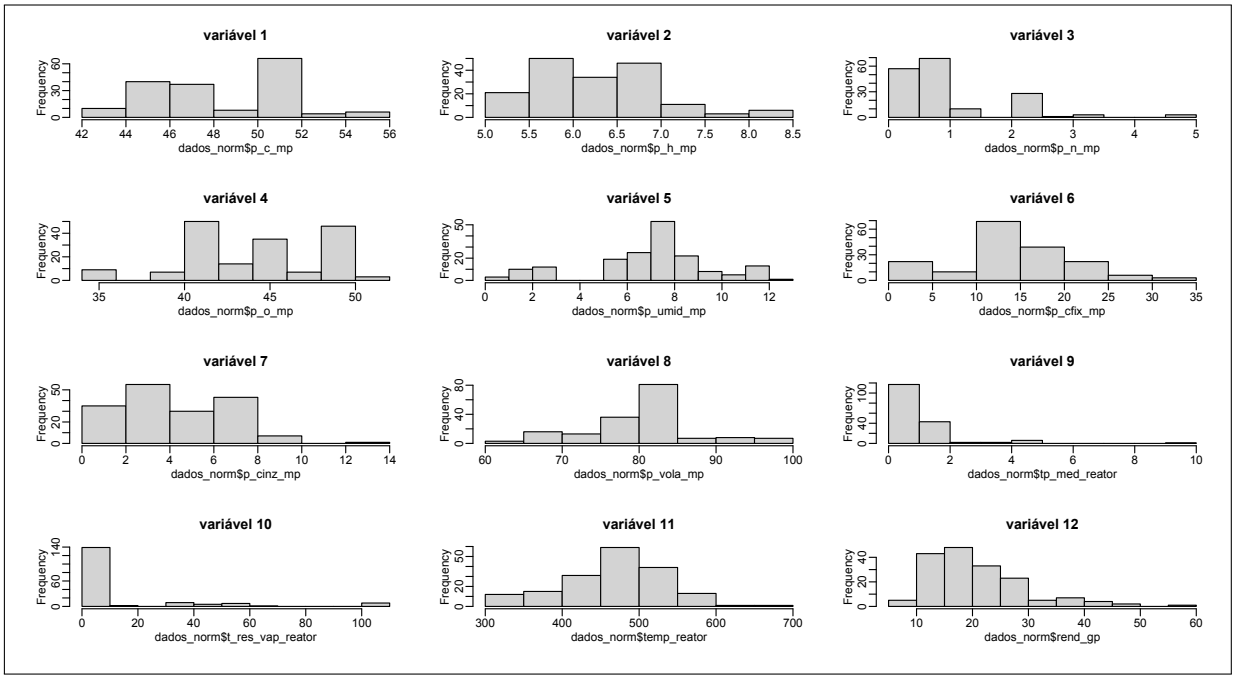
Legenda: Sumarização das variáveis numéricas do regime tipo de reator Contínuo.

Fonte: A autora, 2022.

variáveis mais detalhadamente, as variáveis de número 3, 7 e 12 apresentam assimetria à direita. Já a variável de número 5 apresenta assimetria à esquerda.

Com a compreensão de como estão distribuídos os valores das variáveis, pode-se averiguar as relações entre as variáveis, assim como a interdependência entre elas. Na Figura 13 podemos visualizar pelo mapa de calor das correlações que as variáveis que estão dentro da escala de 1 a 0,7 (+ ou -), na tonalidade azul escuro (no caso das positivas) e na tonalidade vermelho escuro (no caso das negativas), possuem uma forte correlação. Já as variáveis que estão entre 0,7 a 0,5 (+ ou -) na tonalidade azul claro (no caso das positivas) e na tonalidade laranja (no caso das negativas), possuem correlação moderada. As variáveis que possuem escala de 0,5 a 0,25 (+ ou -) possuem baixa correlacão.

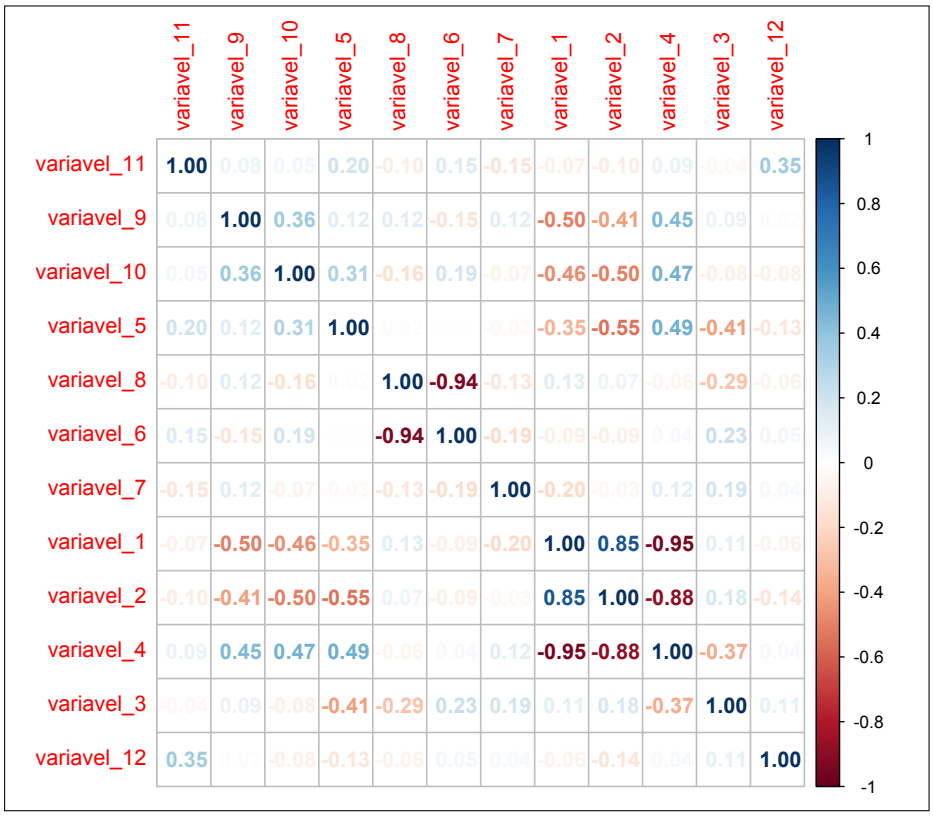
Figura 12 - Histograma das variáveis (Contínuo).



Legenda: Histograma das variáveis numéricas do banco de dados de biomassa do regime tipo reator Contínuo.

Fonte: A autora, 2022.

Figura 13 - Correlação das variáveis (Contínuo).



Legenda: Mapa de calor de correlação das variáveis numéricas do banco de dados de biomassa do regime tipo reator Contínuo.

Fonte: A autora, 2022.

5 RESULTADOS E ANÁLISES

Visto que as variáveis apresentam diferenças de variâncias e assimetrias e dado que o objetivo consiste em identificar grupos de tipos de pirólise com variáveis de processos semelhantes, foi realizada a padronização dos dados para a amostra do reator Batelada e para a amostra do reator Contínuo. Ou seja, foi feita uma transformação de todas as variáveis para a mesma ordem de grandeza.

5.1 Tipo de reator Batelada

Por meio do algoritmo Floresta Aleatória, foi feito um ajuste do número de variáveis e número de árvores e, em seguida, foi realizada a validação cruzada 10-*fold*. Para o número ideal de árvores, obteve-se um pico registrado em 1.000 árvores com quatro variáveis. Além disso, analisando os valores de exatidão e *Kappa*, foi obtido em torno de 97% para exatidão e em torno de 95% para o *Kappa*, sendo o número de variáveis igual a quatro e o número de árvores igual a 1.000.

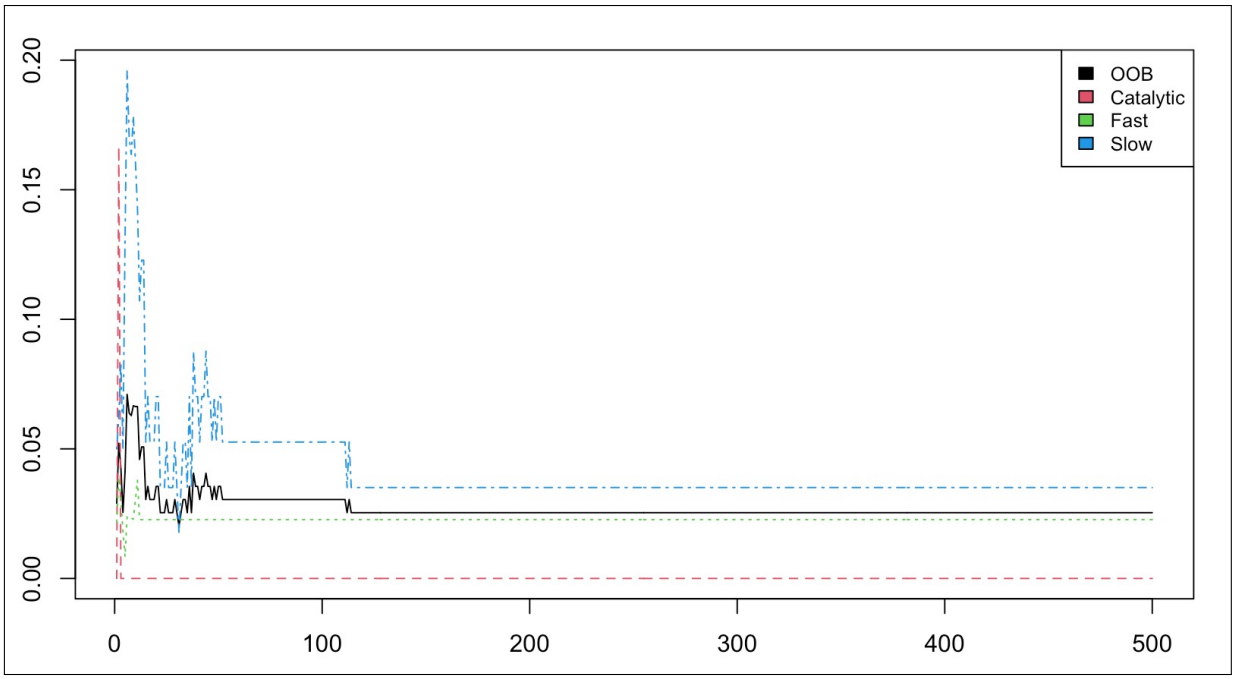
Como há uma classe muito desbalanceada, a exatidão não é uma boa métrica a ser usada, por isso, também foram gerados os valores da exatidão balanceada, obtendo 100% para *Catalytic*, 98% para *Fast* e 98% para *Slow*. Foram obtidos também os valores de *F1-score*, que é uma média ponderada entre a precisão e o *recall*, sendo 100% para *catalytic*, 98% para *fast* e 97% para *slow*.

Na Figura 14 é possível visualizar a taxa de erro *out-of bag* (OOB) em que o erro geral está representado pela linha preta. O erro se inicia com um valor aproximado de 0.05, em seguida cresce e logo depois vai decrescendo e se mantém com uma sazonalidade bem imperceptível um pouco antes de 100 árvores no eixo x do gráfico.

Com o número de variáveis e o número de árvores definidos foi possível identificar as variáveis de mais importância por meio do algoritmo Floresta Aleatória. Os resultados da função Floresta Aleatória obtidos foram: um erro *out-of bag* geral de 2,54% com 1.000 árvores e 4 variáveis, sendo para a classe *Fast* um erro de 2%, para a classe *Slow* um erro de 3% e para a classe *Catalytic* um erro de 0%.

As variáveis de mais importância foram obtidas por meio do método de permutação que avalia a diminuição média na exatidão e pelo índice de Gini que avalia a diminuição média na impureza do nó. Conforme o gráfico mostrado na Figura 15 e por meio da Tabela 4, é possível perceber que as variáveis *Tempo de residência médio no reator para o gás e arraste, Porc. de carbono em base seca livre de cinza na matéria-prima, Tamanho da partícula média do reator e Porc. de hidrogênio em base seca livre de cinza na matéria-prima* foram as variáveis com maior relevância. Portanto, essas foram as variáveis

Figura 14 - Representação gráfica da taxa de erro *Out of bag* em relação ao número de árvores (Batelada).

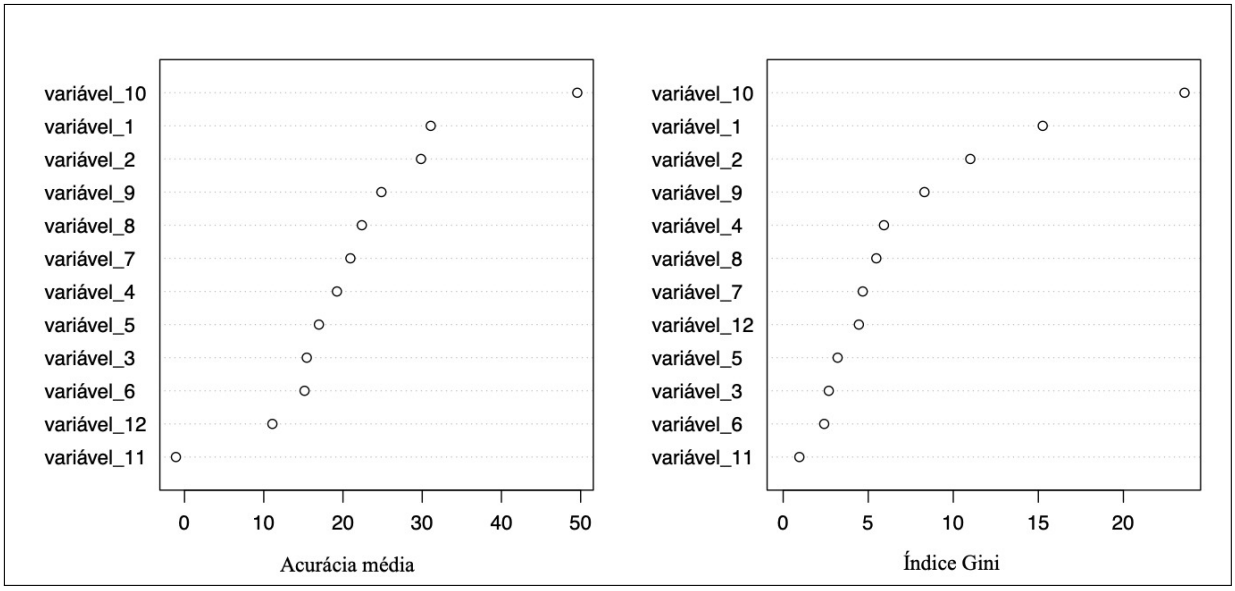


Legenda: Representação gráfica da taxa de erro *Out of bag* em relação ao número de árvores do regime tipo de reator Batelada.

Fonte: A autora, 2022.

utilizadas.

Figura 15 - Representação gráfica da importância de cada variável na classificação considerando a Acurácia e o ìndice de Gini (Batelada).



Legenda: Representação gráfica dos resultados do método de impureza para obter as variáveis de maior relevância do conjunto de dados de biomassa para o tipo de reator Batelada.

Fonte: A autora, 2022.

Para identificar o número ideal de grupos considerou-se agrupamentos de 2 até 8. A distância Euclidiana foi utilizada tanto para os métodos de agrupamento aplicáveis, quanto para as medidas de validação.

As medidas de validação interna utilizadas foram o valor da silhueta e o índice de Dunn. Vale ressaltar que o valor da silhueta traz informações sobre a "consistência" e a "separação" de cada grupo. Um método de agrupamento mais adequado aos dados deve apresentar um valor da silhueta próximo de 1, podendo variar de -1 a 1. Na Tabela 10 podemos ver os resultados obtidos por esta medida de validação.

As medidas de validação também foram exibidas graficamente e são ilustradas nos gráficos das Figuras 16 e 17. Vale lembrar que o índice de Dunn e a silhueta devem ser maximizadas. Desta forma, é possível perceber que o método hierárquico e *k-means*, com seis e sete agrupamentos, e *Diana*, com seis agrupamentos, obtiveram os melhores resultados para o índice Dunn. Já para a silhueta, o método hierárquico e *k-means*, com cinco agrupamentos, obtiveram os melhores desempenhos. Ou seja, quanto maior o valor do índice Dunn e silhueta, melhor será o desempenho. Pode-se observar também que o agrupamento baseado em modelo se destaca negativamente em todos os agrupamentos analisados, não funcionando bem em nenhuma das medidas. Independentemente do algo-

Tabela 4 - Variáveis de maior relevância (Batelada).

Variável	Erro Médio Quadrático	Índice de Gini
10	46,84	23,64
1	32,92	16,94
2	30,14	10,67
9	26,84	8,94

Legenda: Variáveis apontadas como relevantes e seus respectivos valores do regime tipo de reator Batelada.

Fonte: A autora, 2022.

Tabela 5 - Resultados da validação interna (Batelada).

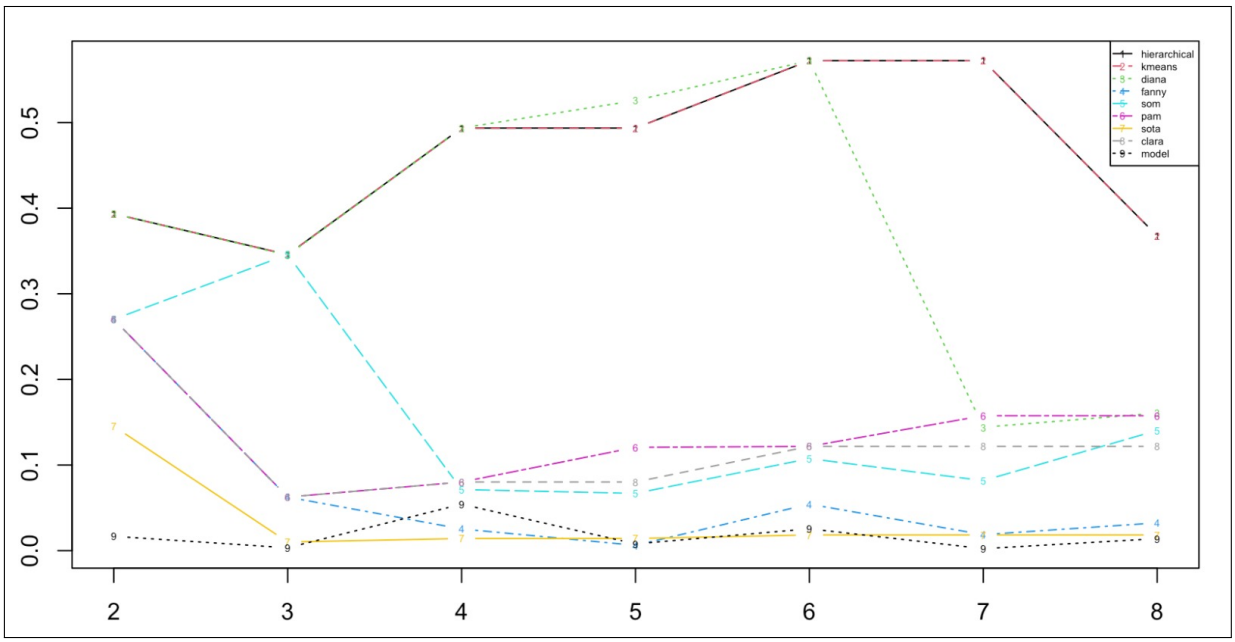
Método	Métrica	2	3	4	5	6	7	8
Hierárquico	Dunn	0,394	0,346	0,494	0,494	0,572	0,572	0,368
Hierárquico	Silhueta	0,597	0,681	0,685	0,690	0,661	0,653	0,653
k-means	Dunn	0,394	0,346	0,494	0,494	0,572	0,572	0,368
k-means	Silhueta	0,597	0,681	0,685	0,690	0,661	0,653	0,653
Diana	Dunn	0,394	0,345	0,494	0,526	0,572	0,144	0,161
Diana	Silhueta	0,597	0,466	0,685	0,655	0,661	0,637	0,645
Fanny	Dunn	0,270	0,063	0,025	0,006	0,054	0,019	0,033
Fanny	Silhueta	0,646	0,552	0,346	0,235	0,357	0,267	0,367
Som	Dunn	0,270	0,346	0,071	0,065	0,013	0,082	0,082
Som	Silhueta	0,646	0,681	0,600	0,567	0,396	0,460	0,482
Pam	Dunn	0,270	0,063	0,080	0,121	0,122	0,158	0,158
Pam	Silhueta	0,646	0,552	0,589	0,616	0,587	0,588	0,588
Sota	Dunn	0,145	0,010	0,014	0,014	0,018	0,018	0,018
Sota	Silhueta	0,631	0,554	0,556	0,555	0,559	0,564	0,566
Clara	Dunn	0,270	0,063	0,080	0,080	0,122	0,122	0,122
Clara	Silhueta	0,646	0,552	0,589	0,597	0,587	0,574	0,619
Model	Dunn	0,017	0,003	0,054	0,008	0,026	0,002	0,014
Model	Silhueta	0,357	0,438	0,429	0,276	0,493	0,422	0,411

Legenda: Resultados da validação interna com as métricas índice Dunn e silhueta para o regime tipo de reator Batelada.

Fonte: A autora, 2022.

ritmo de agrupamento, o número ideal de grupos parece ser seis ou mais para a métrica índice Dunn e cinco usando a métrica silhueta.

Figura 16 - Representação gráfica do Índice de Dunn em relação ao número de grupos (Batelada).



Legenda: Representação gráfica do Índice de Dunn para o regime tipo de reator Batelada. Eixo Y é o valor do Índice dunn e o eixo X é o número de agrupamentos (de 2 até 8).
Fonte: A autora, 2022.

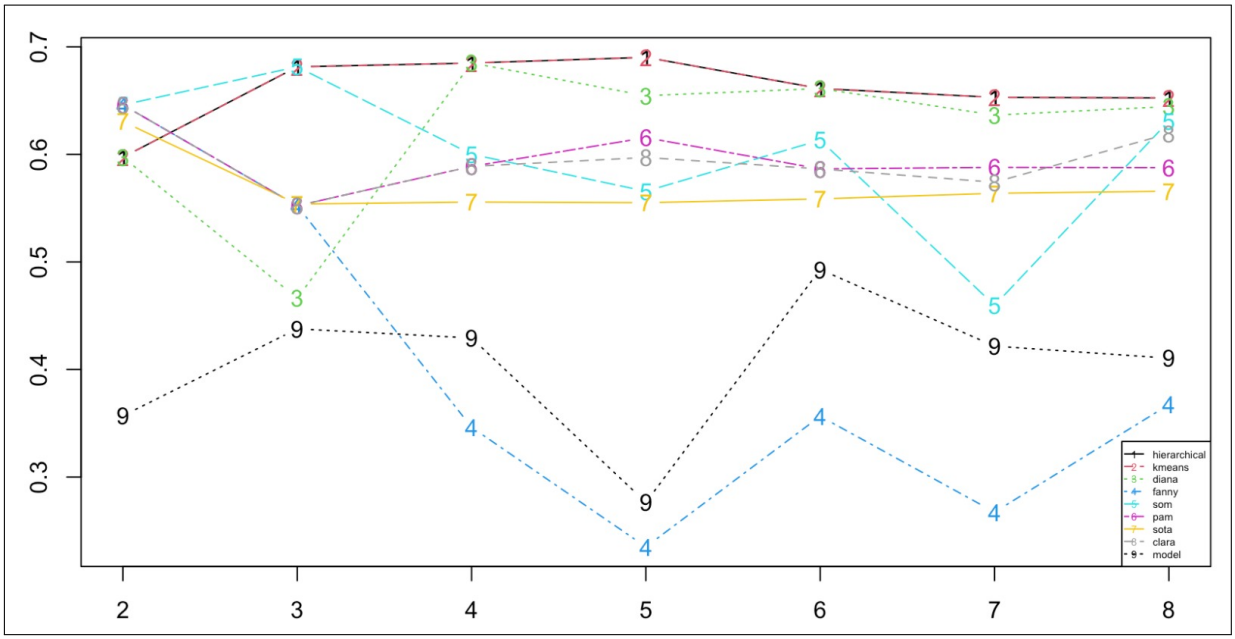
Por meio do gráfico de radar nas Figuras 18 e 19, é possível identificar facilmente esses resultados obtidos. No gráfico de radar, o centro é 0% e quanto mais distante as extremidades das linhas estiverem do centro, maior será o percentual de aceitação. Para a métrica de índice Dunn os agrupamentos Hierárquico, *k-means* e *Diana* obtiveram melhores desempenhos para os grupos 6 e 7, pois as respectivas linhas estão bem distantes do centro em comparação com os outros agrupamentos. Já para a métrica silhueta, é possível perceber que os agrupamentos Hierárquico e *k-means* obtiveram melhores desempenhos para o grupo 5.

Na tabela 6, são mostrados os resultados consolidados obtidos por meio da validação interna com as respectivas métricas, índice Dunn e silhueta:

Com isto, é possível concluir que o agrupamento hierárquico e *k-means* possuem desempenhos consistentes para duas das métricas da validação gerada, índice Dunn e silhueta. Já o valor de k, o mais indicado é que seja um valor maior do que 5.

Foi gerado também o gráfico do valor da silhueta, com base nos métodos hierárquico e *k-means*, para $k = 5$, conforme gráficos nas Figuras 20 e 21. A maioria dos valores são próximos de 1, o que sugere que a observação é bem compatível com o grupo atribuído

Figura 17 - Representação gráfica da Silhueta em relação ao número de grupos (Batelada).



Legenda: Representação gráfica da Silhueta para o regime tipo de reator Batelada. Eixo Y é o valor da Silhueta e o eixo X é o número de agrupamentos (de 2 até 8).

Fonte: A autora, 2022.

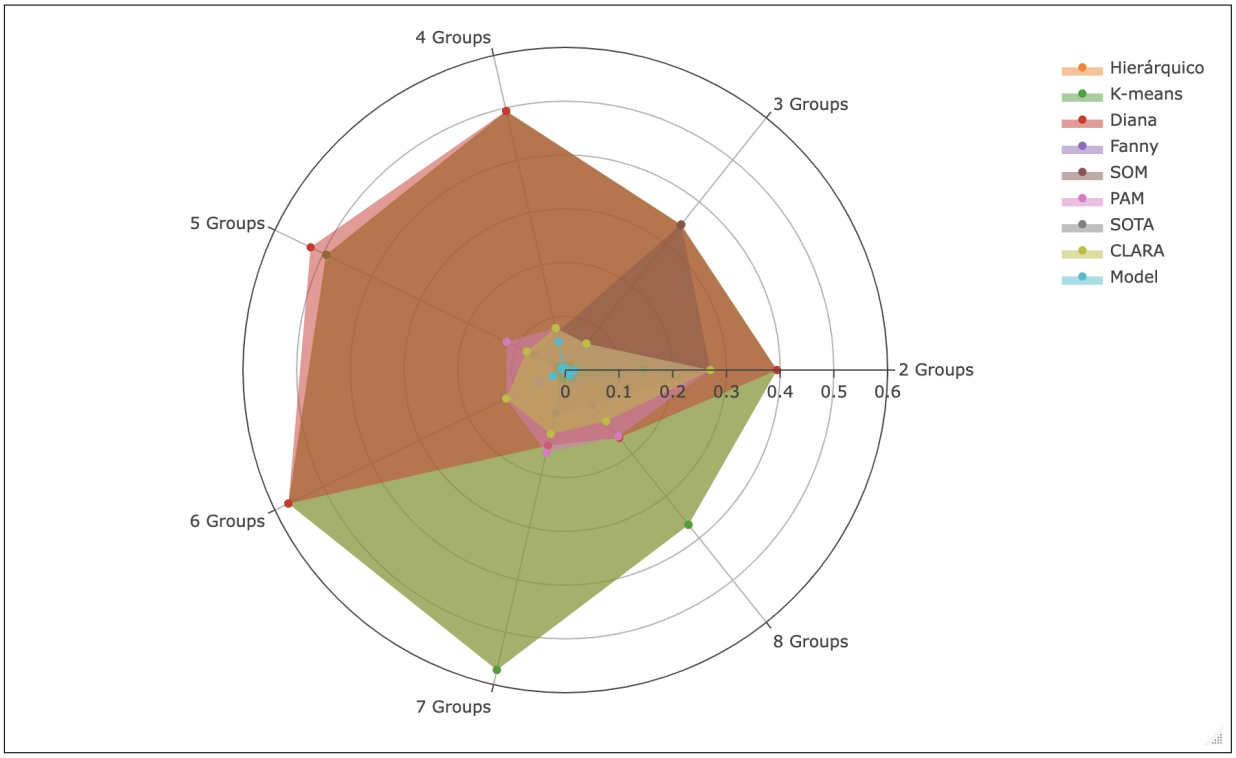
Tabela 6 - Resultados consolidados (Batelada).

Indicador	Interna	Valores de K
Dunn	Hierárquico/k-means/Diana	6 e 7
Silhueta	Hierárquico/k-means	5

Legenda: Resultados consolidados obtidos por meio da validação interna para o regime tipo de reator Batelada.

Fonte: A autora, 2022.

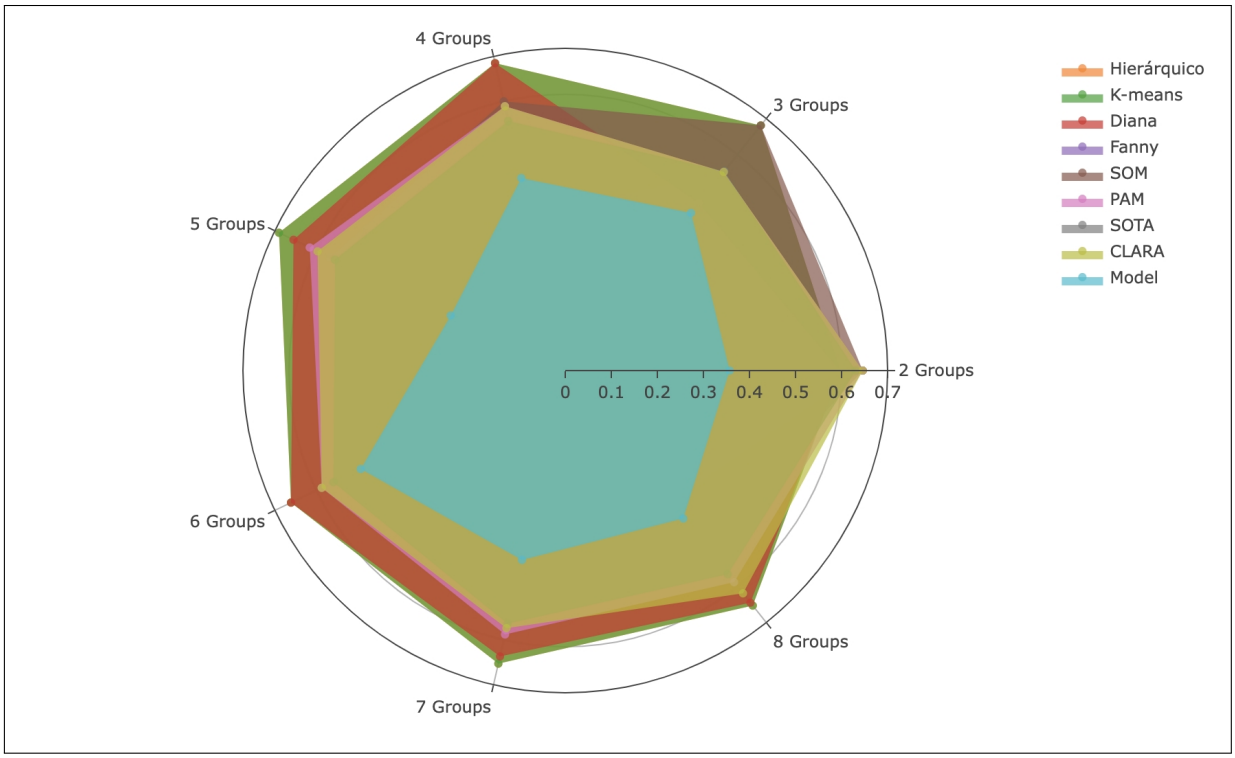
Figura 18 - Representação gráfica de radar do Índice de Dunn (Batelada).



Legenda: Representação gráfica de radar do Índice de Dunn para o regime tipo de reator Batelada. Eixo Y é o valor do Índice dunn e o eixo X é o número de agrupamentos (de 2 até 8).

Fonte: A autora, 2023.

Figura 19 - Representação gráfica de radar da silhueta (Batelada).

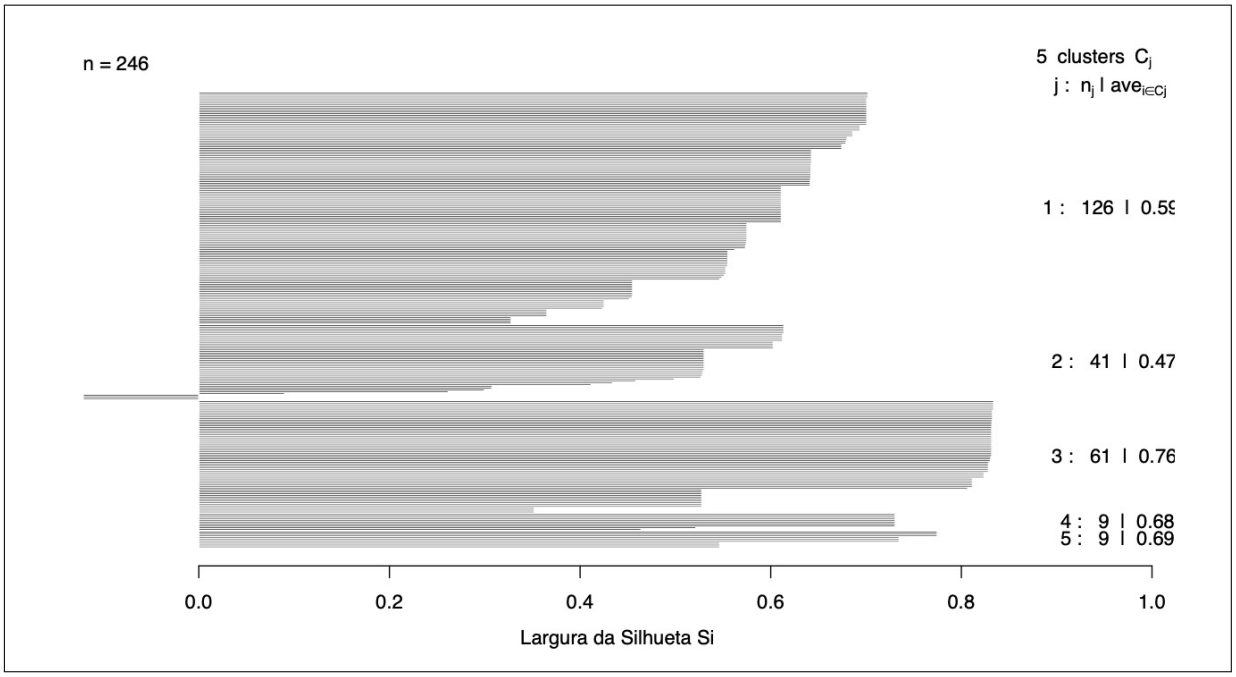


Legenda: Representação gráfica de radar da silhueta para o regime tipo de reator Batelada. Eixo Y é o valor da silhueta e o eixo X é o número de agrupamentos (de 2 até 8).

Fonte: A autora, 2023.

tanto para hierárquico quanto para *k-means*.

Figura 20 - Representação gráfica da silhueta (Batelada).



Legenda: Representação gráfica da silhueta para agrupamento hierárquico do regime tipo de reator Batelada.

Fonte: A autora, 2022.

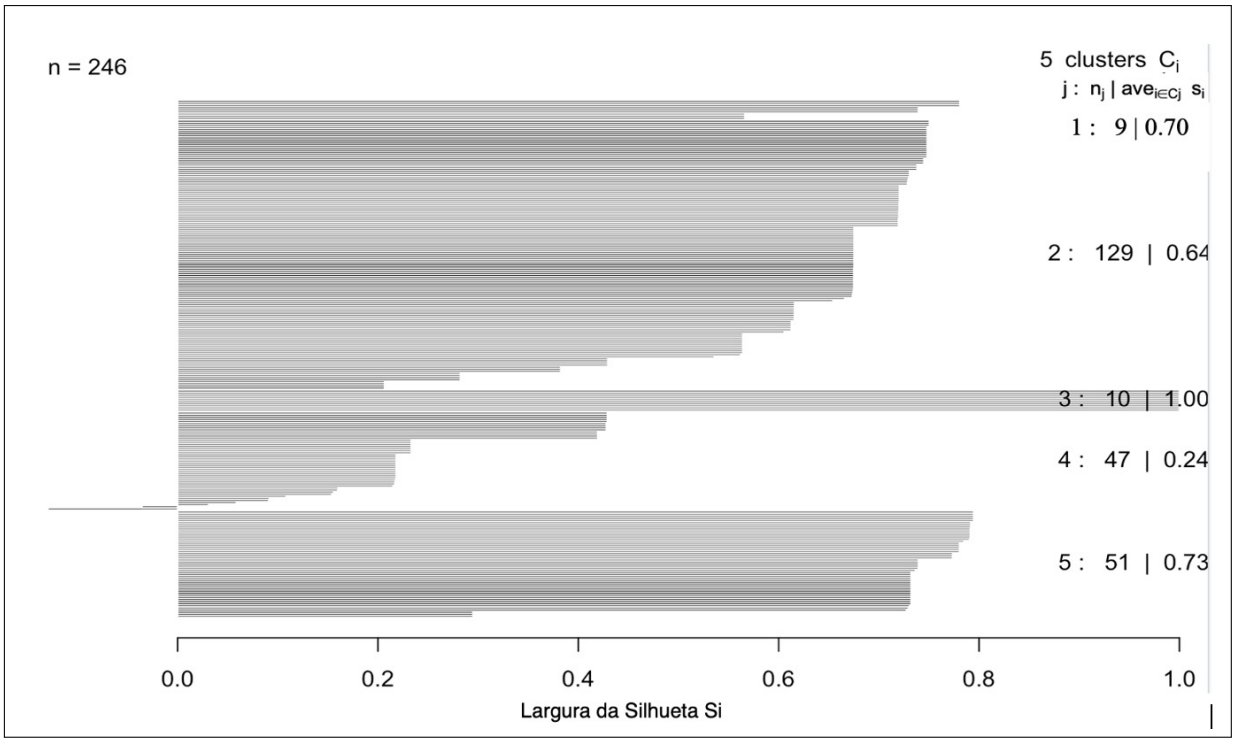
Diante do bom desempenho obtido na medida de validação para os métodos hierárquico e *k-means*, foram extraídos os resultados do agrupamento hierárquico, para traçar o dendrograma e visualizar as observações que são agrupadas nos vários níveis da topologia. O dendrograma está representado na Figura 22, pertencentes às classes funcionais *Fast*, *Slow* e *Catalytic* previamente rotuladas. Foram obtidos cinco grandes ramos ou agrupamentos que emergem do dendrograma, sendo eles bem distintos.

Em seguida, foram identificados os agrupamentos em que foram classificados os dados de biomassa para os métodos Hierárquico (Tabela 7) e *k-means* (Tabela 8). Em ambos os métodos, é possível identificar grupos puros em alguns dos valores de k, como é o caso do k=2, k=4 e k=5 para o método Hierárquico e k=2 e k=5 para o método *k-means*.

Redução de dimensionalidade

Para este método foi utilizado o banco de dados considerando apenas as quatro variáveis de maior relevância identificadas na fase de aplicação do algoritmo Floresta

Figura 21 - Representação gráfica da silhueta (Batelada).



Legenda: Representação gráfica da silhueta para agrupamento *k-means* do regime tipo de reator Batelada.

Fonte: A autora, 2022.

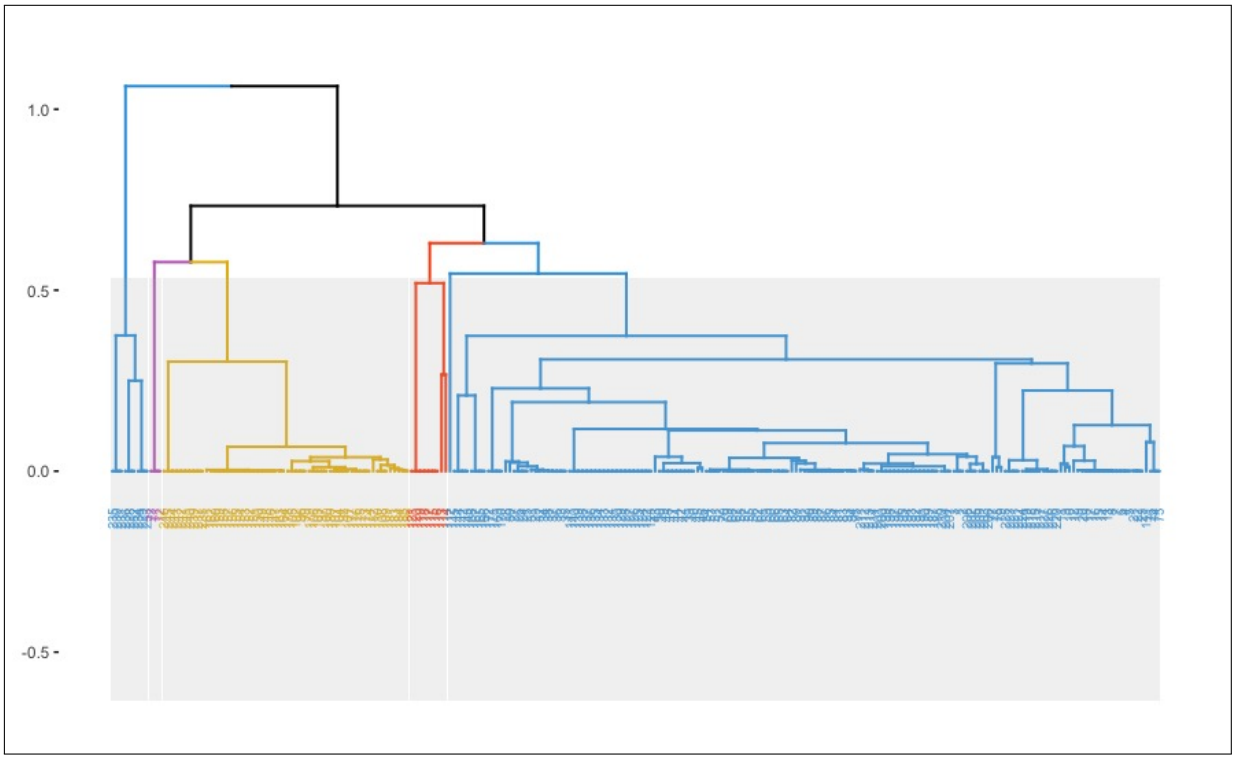
Tabela 7 - Resultados dos agrupamentos para o método hierárquico (Batelada).

Agrupamento	1	2	3	4	5
Catalytic	0	0	10	0	0
Fast	117	0	48	0	0
Slow	50	3	0	9	9

Legenda: Identificação dos agrupamentos em que foram classificados os dados de biomassa no agrupamento hierárquico do regime tipo de reator Batelada.

Fonte: A autora, 2022.

Figura 22 - Representação gráfica do dendrograma (Batelada).



Legenda: Representação gráfica do dendrograma para agrupamento hierárquico do regime tipo de reator Batelada.

Fonte: A autora, 2022.

Tabela 8 - Resultados dos agrupamentos para o método *k-means* (Batelada).

Agrupamento	1	2	3	4	5
Catalytic	0	0	0	10	0
Fast	66	0	51	48	0
Slow	22	8	4	3	34

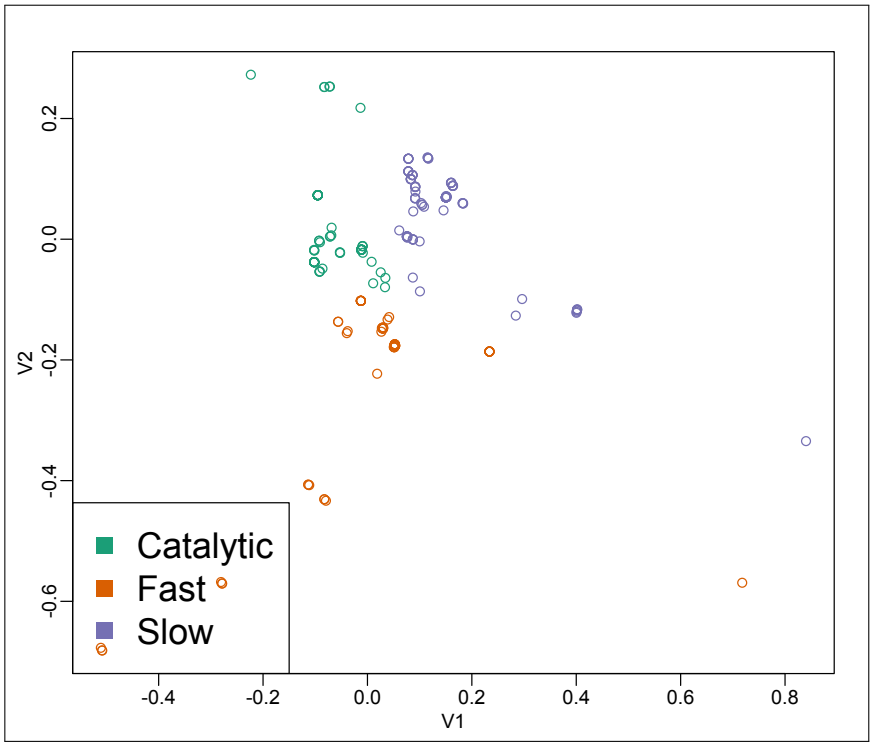
Legenda: Identificação dos agrupamentos em que foram classificados os dados de biomassa no agrupamento *k-means* do regime tipo de reator Batelada.

Fonte: A autora, 2022.

Aleatória, as quais foram: *Tempo de residência médio no reator para o gás e arraste, Porc. carbono em base seca livre de cinza na matéria-prima, Tamanho da partícula média no reator e Porc. hidrogênio em base seca livre de cinza na matéria-prima.*

Foi aplicada a redução de dimensionalidade utilizando o método de Escalonamento Multidimensional Local com suavização hiperbólica (XAVIER et al., 2018), com parâmetro $k = 38$. Já o parâmetro k da métrica Continuidade Local, foi escolhido o valor igual a 28. Com isto, obteve-se um resultado de 18 vizinhos preservados. Os valores de k foram atribuídos por meio de diversas tentativas até encontrar o melhor resultado. No gráfico da Figura 23 é possível observar essa projeção das classes e perceber que as classes estão divididas, entretanto as classes não são homogêneas.

Figura 23 - Representação gráfica da redução de dimensionalidade do plano R^4 para R^2 (Batelada)



Fonte: A autora, 2022.

Legenda: Representação gráfica da redução de dimensionalidade do plano R^4 para R^2 para o regime tipo de reator Batelada.

5.2 Tipo de reator Contínuo

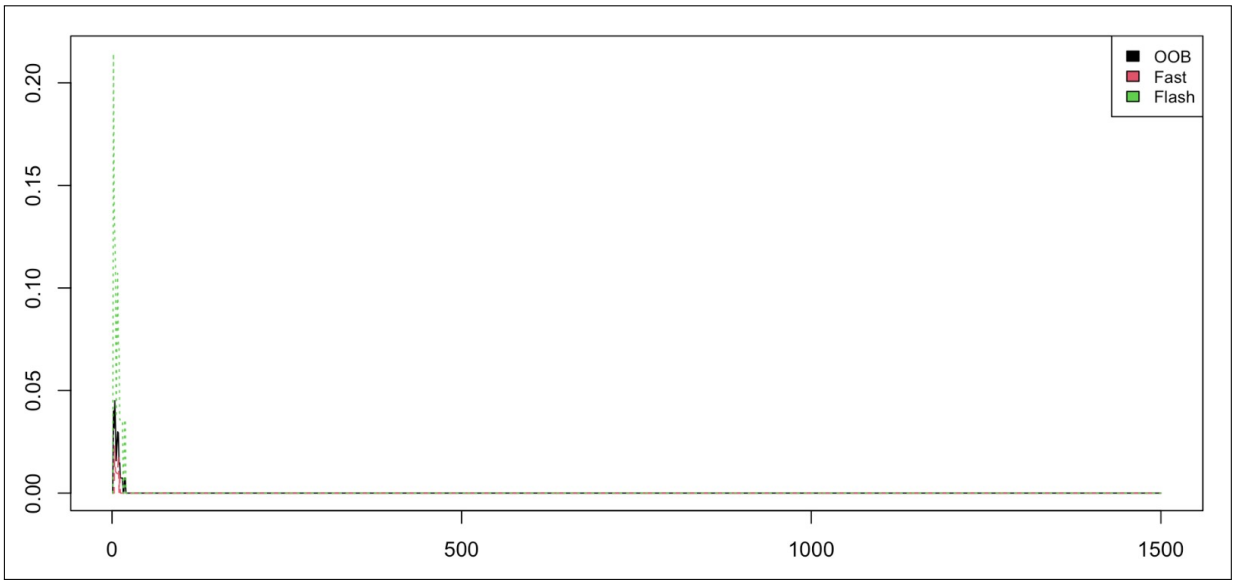
Para a amostra do regime de reator Contínuo, obteve-se um pico registrado em 1.500 árvores e com uma variável. Além disso, analisando os valores de exatidão e *kappa*,

foi obtido em torno de 98% para exatidão e em torno de 98% para o *kappa*, sendo o número de variáveis igual a um e o número de árvores igual a 1.500. Para a base de dados do reator Contínuo, foi excluído a classe de pirólise *Catalytic* pelo fato de possuir apenas uma amostra.

Os valores da exatidão balanceada foram: 98% para *Fast* e 100% para *Flash*. Foram obtidos também os valores de *F1-score*, sendo 98% para *Fast* e 100% para *Flash*.

Por meio do erro *Out of bag* (OOB) obtido no gráfico da Figura 24, é possível observar o erro geral na linha preta iniciando bem próximo de 0,05, em seguida decresce e se mantém com uma sazonalidade bem imperceptível a partir de 100 árvores no eixo x.

Figura 24 - Representação gráfica da taxa de erro *Out of bag* em relação ao número de árvores (Contínuo).



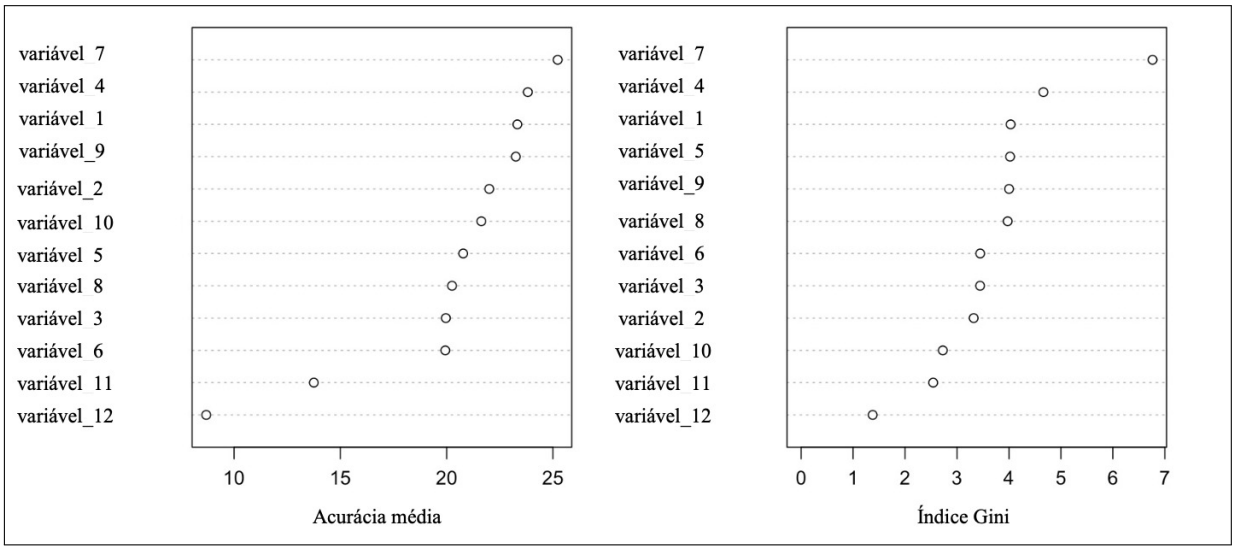
Legenda: Representação gráfica da taxa de erro *Out of bag* em relação ao número de árvores do regime tipo de reator Contínuo.
Fonte: A autora, 2022.

Os resultados da função Floresta Aleatória obtidos foram: um erro *Out of bag* geral de 0,01% com 1.500 árvores e uma variável.

Conforme o gráfico na Figura 25 e por meio da Tabela 9, é possível perceber que através do Índice de Gini as variáveis de mais importância obtidas foram: *Porc. cinzas em base seca da matéria-prima, Porc. carbono em base seca livre de cinza na matéria-prima e Porc. oxigênio em base seca livre de cinza na matéria-prima.*

Para o regime Contínuo também foi feita a variação no número de agrupamentos de 2 até 8 e foi usada a distância Euclidiana.

Figura 25 - Representação gráfica da importância de cada variável na classificação considerando a Acurácia e o ìndice de Gini(Contínuo).



Fonte: A autora, 2022.

Legenda: Representação gráfica dos resultados do método de impureza para obter as variáveis de maior relevância do conjunto de dados de Biomassa para o tipo de reator Contínuo.

Tabela 9 - Variáveis de maior relevância (Contínuo).

Variável	Erro Médio Quadrático	Índice de Gini
7	25,22	6,76
4	23,81	4,66
1	23,32	4,03

Legenda: Variáveis apontadas como relevantes na fase de aplicação do algoritmo Floresta Aleatória para o regime tipo de reator Contínuo.

Fonte: A autora, 2022.

Na Tabela 10 podemos ver os resultados obtidos para a medida de validação interna e nas Figuras 26 e 27, são exibidos graficamente o índice de Dunn e a silhueta. É possível perceber que o método hierárquico, *k-means* e *Diana* com sete e oito agrupamentos, obtiveram os melhores resultados para o índice Dunn. Já para a métrica silhueta, o método *Clara*, com sete agrupamentos, obteve o melhor desempenho.

Tabela 10 - Resultados da validação interna (Contínuo).

Método	Métrica	2	3	4	5	6	7	8
Hierárquico	Dunn	0,452	0,357	0,485	0,485	0,605	0,706	0,706
Hierárquico	Silhueta	0,437	0,540	0,517	0,463	0,574	0,650	0,647
k-means	Dunn	0,309	0,413	0,461	0,485	0,605	0,706	0,706
k-means	Silhueta	0,548	0,542	0,519	0,463	0,574	0,650	0,647
Diana	Dunn	0,313	0,332	0,483	0,506	0,506	0,706	0,550
Diana	Silhueta	0,396	0,429	0,423	0,561	0,557	0,650	0,688
Fanny	Dunn	0,309	0,197	0,082	0,082	0,054	0,091	0,100
Fanny	Silhueta	0,548	0,556	0,551	0,629	0,372	0,508	0,605
Som	Dunn	0,309	0,214	0,082	0,082	0,082	0,074	0,074
Som	Silhueta	0,548	0,457	0,551	0,629	0,680	0,544	0,538
Pam	Dunn	0,117	0,197	0,197	0,235	0,316	0,141	0,100
Pam	Silhueta	0,522	0,556	0,537	0,625	0,673	0,657	0,702
Sota	Dunn	0,309	0,290	0,069	0,069	0,069	0,133	NA
Sota	Silhueta	0,548	0,573	0,550	0,627	0,678	0,687	NA
Clara	Dunn	0,117	0,298	0,229	0,235	0,316	0,091	0,141
Clara	Silhueta	0,522	0,572	0,549	0,625	0,673	0,704	0,677
Model	Dunn	0,149	0,048	0,021	0,021	0,086	0,080	0,103
Model	Silhueta	0,303	0,263	0,319	0,431	0,619	0,600	0,576

Legenda: Resultados da validação interna com as métricas índice Dunn e silhueta para o regime tipo de reator Contínuo.

Fonte: A autora, 2022.

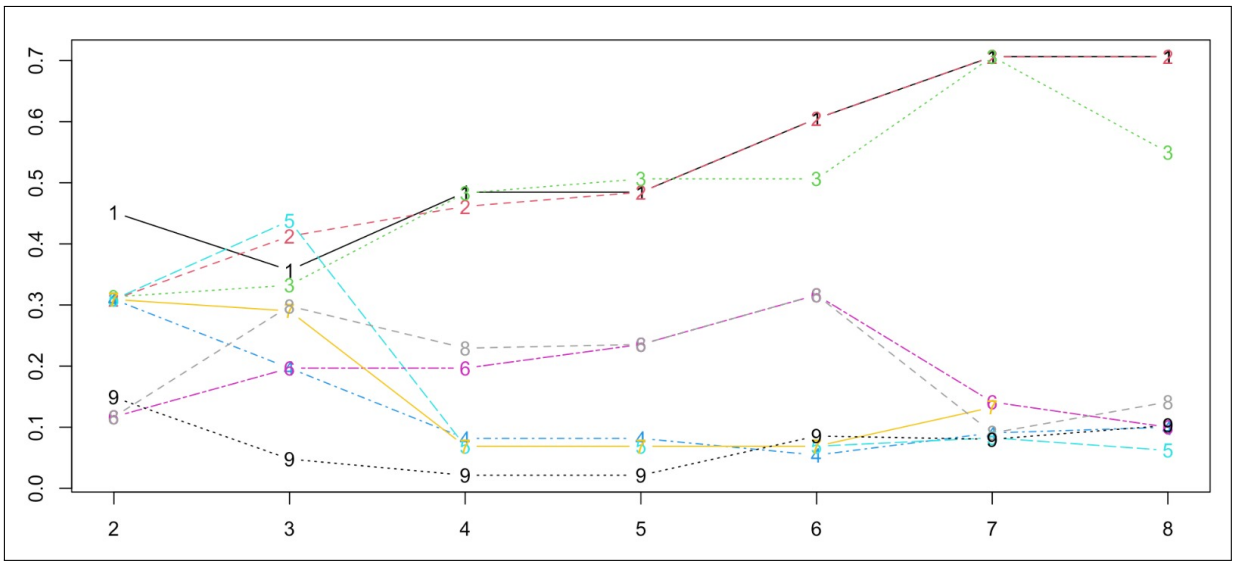
Assim como no tipo de reator Batelada, independente do algoritmo de agrupamento, o número ideal de grupos é sempre grande, no caso do reator contínuo, maior ou igual a sete.

Por meio do gráfico de radar nas Figuras 28 e 29, para a métrica de índice Dunn os agrupamentos Hierárquico, *k-means* e *Diana* obtiveram melhores resultados para os grupos 7 e 8, pois as respectivas linhas estão bem distantes do centro em comparação com os outros agrupamentos. Já para a métrica silhueta, é possível perceber que o agrupamento *Clara* obteve o melhor desempenho para o grupo igual a 7.

Na Tabela 11, são mostrados os resultados consolidados obtidos pela validação interna com as respectivas métricas, índice Dunn e silhueta.

Com isto, é possível concluir que, assim como o caso do regime de reator Batelada, para o Contínuo, o valor de k também é grande, maior ou igual a sete.

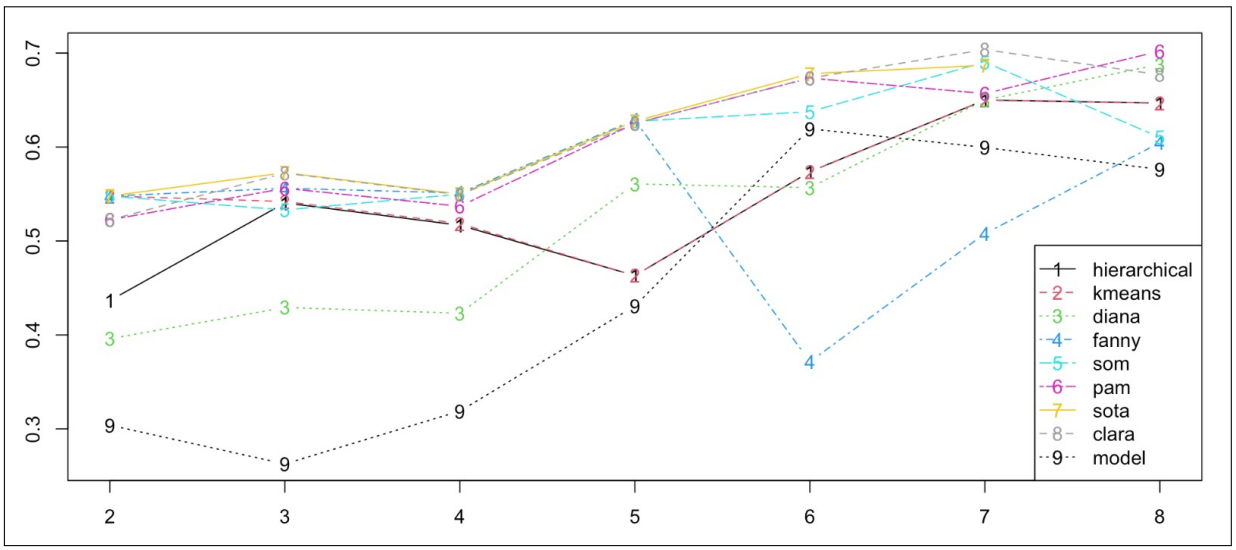
Figura 26 - Representação gráfica do Índice de Dunn em relação ao número de grupos (Contínuo).



Legenda: Representação gráfica do Índice de Dunn para o regime tipo de reator Contínuo. Eixo Y é o valor do Índice dunn e o eixo X é o número de agrupamentos (de 2 até 8).

Fonte: A autora, 2022.

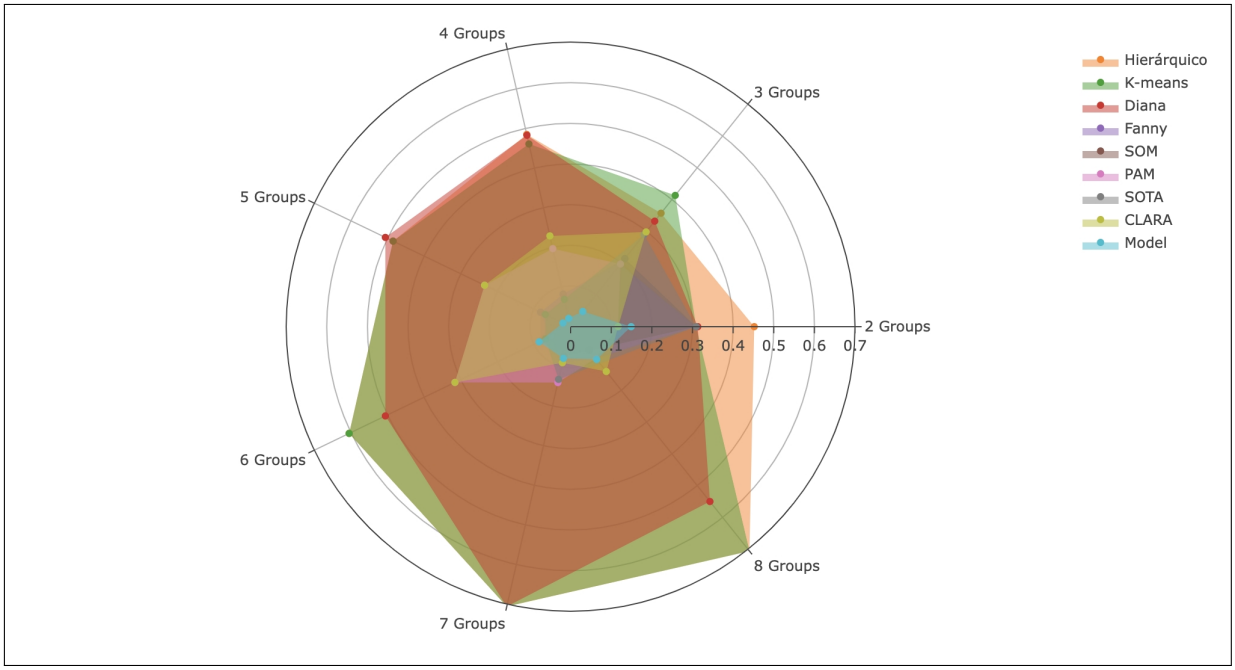
Figura 27 - Representação gráfica da silhueta em relação ao número de grupos (Contínuo).



Legenda: Representação gráfica da silhueta para o regime tipo de reator Contínuo. Eixo Y é o valor da silhueta e o eixo X é o número de agrupamentos (de 2 até 8).

Fonte: A autora, 2022.

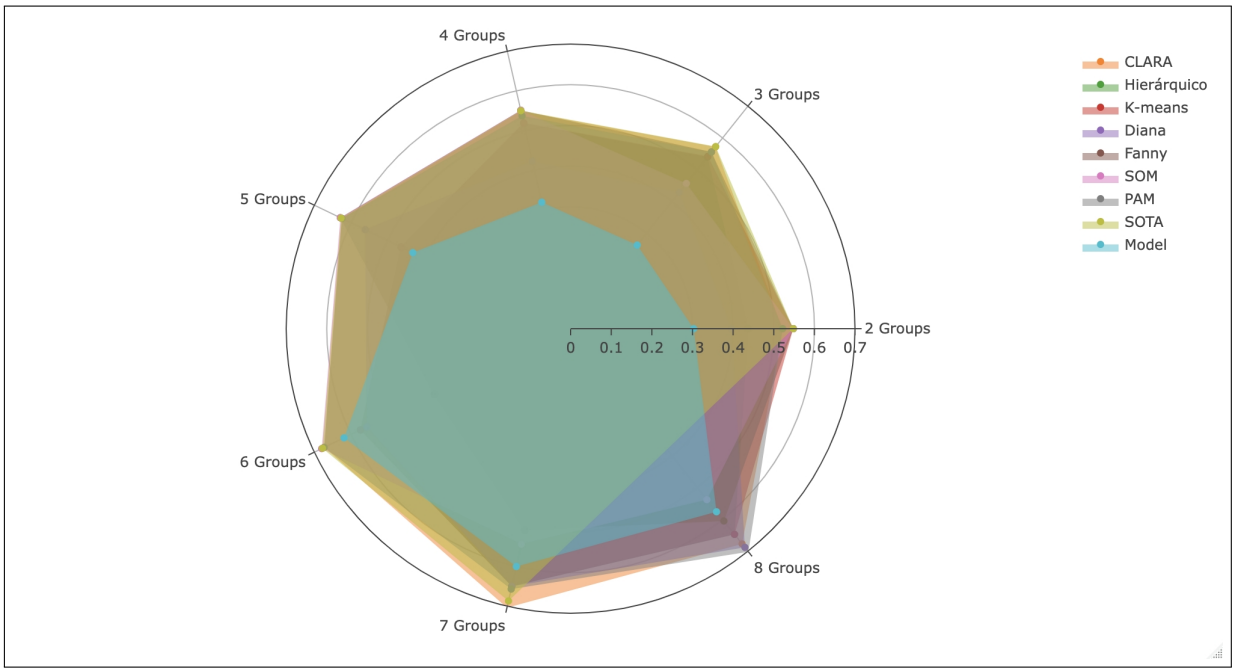
Figura 28 - Representação gráfica de radar do Índice de Dunn (Contínuo).



Legenda: Representação gráfica de radar do Índice de Dunn para o regime tipo de reator Contínuo.

Fonte: A autora, 2023.

Figura 29 - Representação gráfica de radar da silhueta (Contínuo).



Legenda: Representação gráfica de radar da silhueta para o regime tipo de reator Contínuo.

Fonte: A autora, 2023.

Tabela 11 - Resultados consolidados (Contínuo).

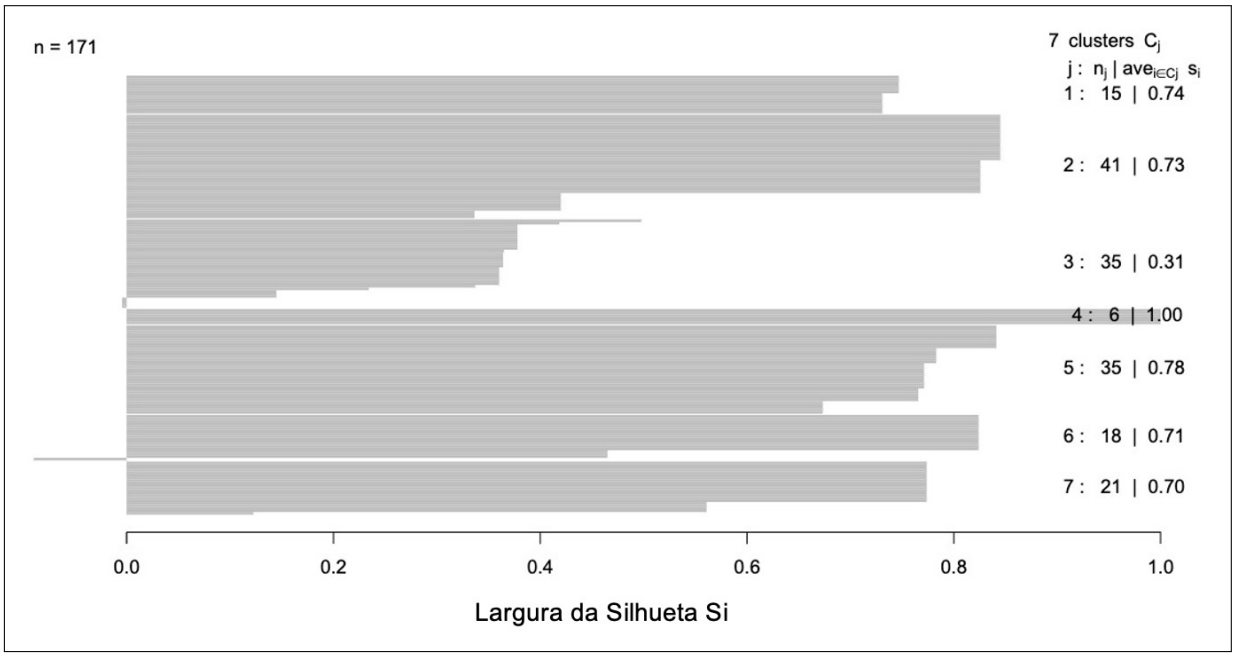
Indicador	Interna	Valores de K
Dunn	Hierárquico/k-means/Diana	7 e 8
Silhueta	Clara	7

Legenda: Resultados consolidados obtidos por meio da validação interna para o regime tipo de reator Contínuo.

Fonte: A autora, 2022.

Foi gerado também o gráfico da silhueta, com base nos métodos hierárquico e *k-means*, para $k = 7$, conforme gráficos nas Figuras 30 e 31. A maioria dos valores são próximos de 1, o que sugere que a observação é bem compatível com o agrupamento atribuído em ambos os métodos.

Figura 30 - Representação gráfica da silhueta (Contínuo).

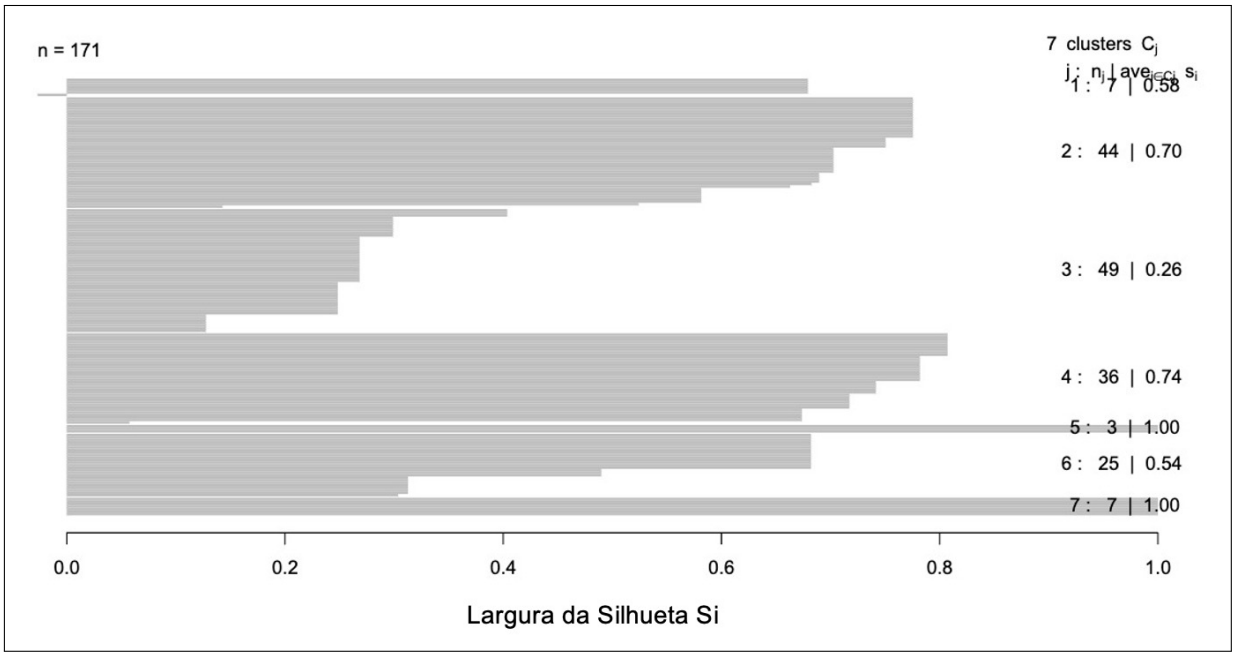


Legenda: Representação gráfica da silhueta para agrupamento hierárquico do regime tipo de reator Contínuo.

Fonte: A autora, 2022.

Diante do bom desempenho obtido na medida de validação para os métodos hierárquico e *k-means*, foram extraídos os resultados do agrupamento hierárquico, para traçar o dendrograma e visualizar as observações. O dendrograma está representado no gráfico da figura 32, pertencentes às classes funcionais *Fast* e *Flash* previamente rotuladas. Foram obtidos sete grandes ramos ou agrupamentos que emergem do dendrograma.

Figura 31 - Representação gráfica da silhueta (Contínuo).



Legenda: Representação gráfica da silhueta para agrupamento *k-means* do regime tipo de reator Contínuo.

Fonte: A autora, 2022.

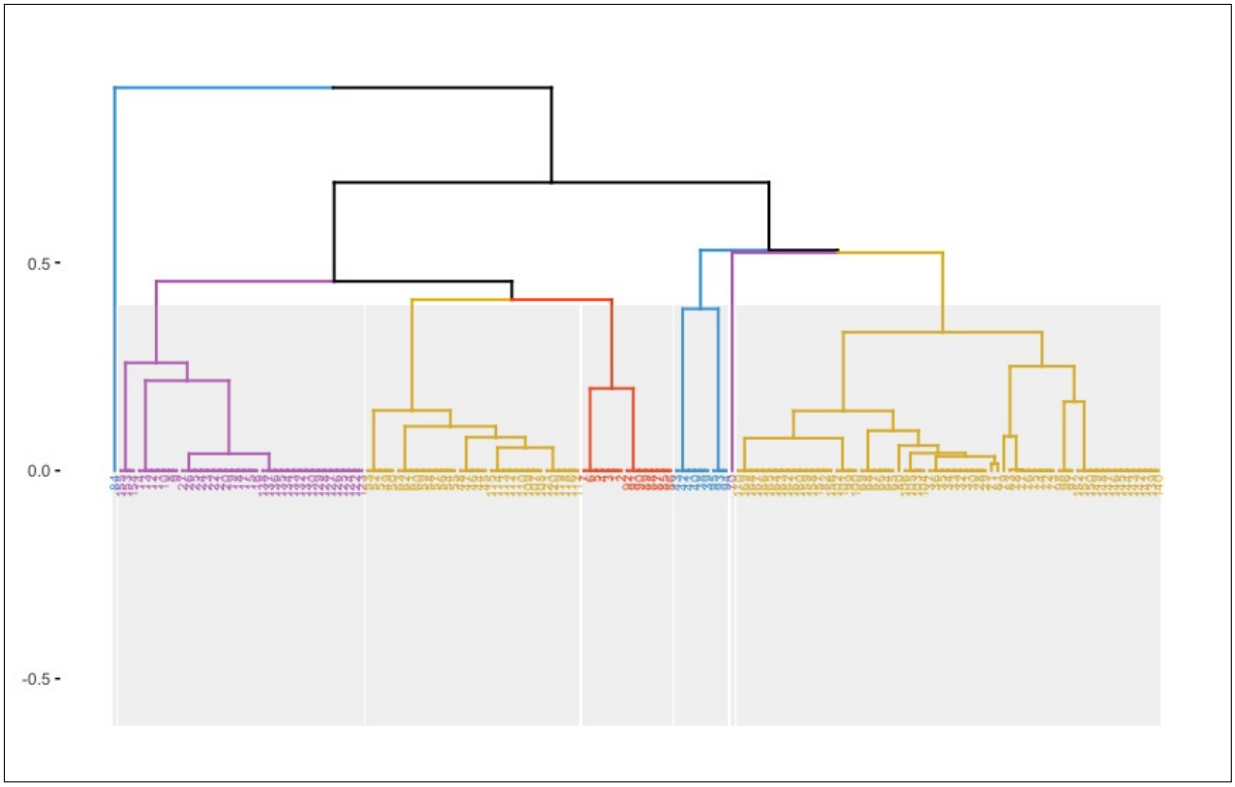
Em seguida, foram identificados os agrupamentos em que foram classificados os dados de biomassa para os métodos hierárquico (Tabela 12) e *k-means* (Tabela 13). Em ambos os métodos, é possível identificar grupos puros em todos os valores de k, tanto para o hierárquico quanto para o *k-means*.

Redução de Dimensionalidade

Para este método foi utilizado o banco de dados considerando apenas as três variáveis de maior relevância identificadas na fase de aplicação do algoritmo Floresta Aleatória, as quais foram: *Porc. cinzas em base seca da matéria-prima, Porc. carbono em base seca livre de cinza na matéria-prima e Porc. oxigênio em base seca livre de cinza na matéria-prima*

Foi aplicada a redução de dimensionalidade utilizando o método de Escalonamento Multidimensional Local com suavização hiperbólica (XAVIER et al., 2018), com parâmetro $k = 38$. Já o parâmetro k da métrica Continuidade Local, foi escolhido o valor igual a 18. Com isto, obteve-se um resultado de 11 vizinhos preservados. Os valores de k foram atribuídos por meio de diversas tentativas até encontrar o melhor resultado. No gráfico da Figura 33 é possível observar essa projeção das classes e perceber que as classes estão divididas, entretanto as classes não são homogêneas.

Figura 32 - Representação gráfica do dendrograma (Contínuo).



Legenda: Representação gráfica do dendrograma para agrupamento hierárquico do regime tipo de reator Contínuo.

Fonte: A autora, 2022.

Tabela 12 - Resultados dos agrupamentos para o método hierárquico (Contínuo).

Agrupamento	1	2	3	4	5	6	7
Fast	15	40	69	9	0	1	1
Flash	0	0	0	0	35	0	0

Legenda: Identificação dos agrupamentos em que foram classificados os dados de biomassa no agrupamento hierárquico do regime tipo de reator Contínuo.

Fonte: A autora, 2022.

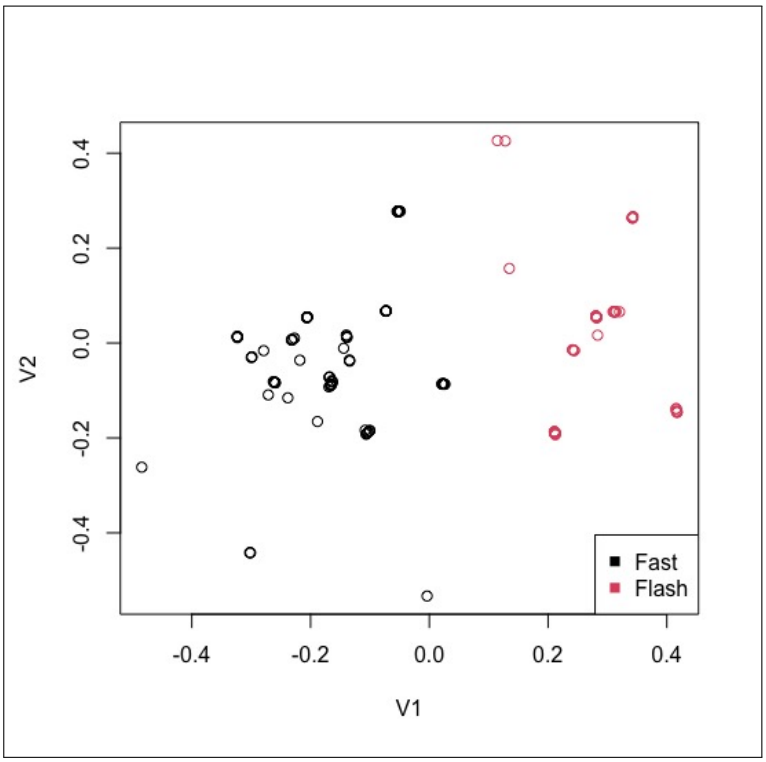
Tabela 13 - Resultados dos agrupamentos para o método *k-means* (Contínuo).

Agrupamento	1	2	3	4	5	6	7
Fast	18	3	21	1	25	37	30
Flash	0	0	0	0	0	0	35

Legenda: Identificação dos agrupamentos em que foram classificados os dados de biomassa no agrupamento *k-means* do regime tipo de reator Contínuo.

Fonte: A autora, 2022.

Figura 33 - Representação gráfica da redução de dimensionalidade do plano R^4 para R^2 (Contínuo).



Legenda: Representação gráfica da redução de dimensionalidade do plano R^4 para R^2 para o regime tipo de reator Contínuo.

Fonte: A autora, 2022.

CONCLUSÃO

Nesta dissertação foi tratado o problema de análise de dados de pirólise de biomassa para o tipo de reator Batelada e tipo de reator Contínuo. Para o tipo de reator Batelada, as amostras são agrupadas nas classes *Fast*, *Slow* e *Catalytic* e para o tipo de reator Contínuo, *Fast*, *Flash* e *Catalytic*, de acordo com as doze variáveis elementares, complementares e de processo.

Foi feita uma análise exploratória dos dados para cada uma das amostras para entender o perfil dos dados. Visto que as variáveis apresentaram diferenças de variâncias e assimetrias, foi realizada a padronização dos dados para cada uma das amostras separadamente.

Em seguida, foram aplicados diversos métodos de Inteligência Artificial, buscando melhor entender a classificação dos dados para os diferentes tipos de pirólise. Analisou-se o problema de classificação de dados de biomassa, com o objetivo de identificar as variáveis de maior relevância para a classificação do tipo de pirólise de biomassa. Para isto, foi necessário aplicar, previamente, o algoritmo de Floresta Aleatória para identificar as variáveis de maior relevância, obtendo uma exatidão em torno de 97% para a amostra do reator Batelada e 98% para a amostra do reator Contínuo na aplicação do algoritmo Floresta Aleatória.

Para a amostra do regime do tipo de reator Batelada, as variáveis de maior relevância identificadas, foram: *Tempo de residência médio no reator para o gás e arraste, Porc. carbono em base seca livre de cinza na matéria-prima, Tamanho da partícula média no reator e Porc. hidrogênio em base seca livre de cinza na matéria-prima.* Com este subconjunto, foram aplicados nove métodos de agrupamento. Através da medida de validação interna foi possível concluir que o agrupamento hierárquico e *k-means* possuem desempenhos consistentes para duas das métricas da validação interna geradas, índice Dunn e silhueta. Já o valor de k, o mais indicado é que seja um valor maior do que cinco, pois os grupos naturais formados pelos métodos de agrupamento, são bem distintos das classes associadas aos dados.

As mesmas aplicações foram realizadas para a amostra do regime do tipo de reator Contínuo, com as variáveis de maior relevância identificadas, as quais foram: *Porc. cinzas em base seca da matéria-prima, Porc. carbono em base seca livre de cinza na matéria-prima e Porc. oxigênio em base seca livre de cinza na matéria-prima*, considerando apenas as classes *Fast* e *Flash*. Foi possível concluir que o agrupamento hierárquico e o agrupamento *k-means* obtiveram bons resultados para a métrica do Índice Dunn e o agrupamento *Clara* obteve um bom resultado para a métrica da silhueta. E, assim como no caso do reator Batelada, o valor de K indicado é bem maior do que o número de grupos previamente classificados no banco de dados, sendo k igual ou maior do que sete como o

ideal para o reator Contínuo.

Por fim, foi aplicado o método de redução de dimensionalidade, considerando apenas as variáveis de maior relevância encontradas na fase de aplicação do algoritmo Floresta Aleatória. Foi aplicado o método de Escalonamento Multidimensional Local com suavização hiperbólica, com parâmetro $k = 38$. Para o tipo de reator Batelada, foi utilizado o parâmetro k do critério Continuidade Local igual a 28 e obteve-se 18 vizinhos preservados. Já para o tipo de reator Contínuo, foi utilizado o parâmetro k do critério Continuidade Local igual a 18 e obteve-se 11 vizinhos preservados. Os resultados do método de redução de dimensionalidade mostraram que as classes ficaram divididas, porém, as classes não são homogêneas.

Em suma, os resultados estatísticos obtidos mostram que apenas quatro variáveis para o reator Batelada, e apenas três variáveis para o reator Contínuo, são necessárias para classificar os tipos de pirólise. Conclui-se também a necessidade de dividir o conjunto de dados em um número maior de grupos de tipos de pirólise (mais do que as três classes previamente apresentadas no banco de dados), pois as classes previamente já rotuladas e fornecidas no banco de dados são muito limitadas para caracterizar o processo de pirólise.

REFERÊNCIAS

AZEVEDO, B. B.; ANZANELLO, M. J. Agrupamento de trabalhadores com perfis semelhantes de aprendizado apoiado em análise de componentes principais. *Gestão & Produção*, SciELO Brasil, v. 22, p. 35–52, 2015.

BACKER, E.; JAIN, A. K. A clustering performance measure based on fuzzy set decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, n. 1, p. 66–75, 1981.

BARBOSA, M.; CARNEIRO, T.; TAVARES, A. I. Métodos de classificação por árvores de decisão disciplina de projeto e análise de algoritmos. *UFOP–Universidade Federal de Ouro Preto Ouro Preto, Minas Gerais–MG*, 2012.

BATOOL, F.; HENNIG, C. Clustering with the average silhouette width. *Computational Statistics & Data Analysis*, Elsevier, v. 158, p. 107190, 2021.

BERTOLETTI, A. Z. et al. Técnica de validação cruzada para qualificação do ajuste das curvas tempo versus corrente dos elos fusíveis do tipo expulsão.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BRIAN, S. Cluster analysis 3rd ed. *Edward Arnold, London*, v. 169, 1993.

BROCK, G. et al. clvalid: An r package for cluster validation. *Journal of Statistical Software*, v. 25, p. 1–22, 2008.

BURMAN, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, Oxford University Press, v. 76, n. 3, p. 503–514, 1989.

CAO, H.; XIN, Y.; YUAN, Q. Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach. *Bioresource technology*, Elsevier, v. 202, p. 158–164, 2016.

CHEN, L.; BUJA, A. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, Taylor & Francis, v. 104, n. 485, p. 209–219, 2009.

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. *Genomics*, Elsevier, v. 99, n. 6, p. 323–329, 2012.

DUNN, J. C. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, Taylor & Francis, v. 4, n. 1, p. 95–104, 1974.

FADEL, A. C.; SEMAAN, G. da S.; BRITO, J. A. de M. Um estudo da aplicação de técnicas de combinação de agrupamentos. *Anais do XVII Simpósio de Pesquisa Operacional e Logística da Marinha*, v. 1, n. 1, p. 188–200, 2014.

GORDON, A. D. *Classification*. [S.l.]: CRC Press, 1999.

GUEDES, R. E.; LUNA, A. S.; TORRES, A. R. Operating parameters for bio-oil production in biomass pyrolysis: A review. *Journal of analytical and applied pyrolysis*, Elsevier, v. 129, p. 134–149, 2018.

HAIR, J. F. et al. *Análise multivariada de dados*. [S.l.]: Bookman editora, 2009.

HASTIE, T. et al. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2009. v. 2.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, Acm New York, NY, USA, v. 31, n. 3, p. 264–323, 1999.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112.

KAUFMAN, L.; ROUSSEEUW, P. Fuzzy analysis (program fanny). *Finding Groups in Data*, Wiley, p. 164–198, 1990.

LAM, D.; WUNSCH, D. Clustering. In: _____. [S.l.: s.n.], 2014. v. 1, p. 1115–1149. ISBN 9780123965028.

LANDAU, S.; STER, I. C. Cluster analysis: overview. *Á Á*, v. 11, n. x12, p. x1p, 2010.

LINDEN, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, n, v. 4, n. 4, p. 18–36, 2009.

LORENA, A. C.; CARVALHO, A. C. de. Introduçao aos classificadores de margens largas. *Sao Carlos-SP*, 2003.

MAHESH, B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, v. 9, p. 381–386, 2020.

MANLY, B. F.; ALBERTO, J. A. N. *Métodos estatísticos multivariados: uma introdução*. [S.l.]: Bookman Editora, 2008.

MÁRQUEZ, J. F.; SABATÉ, J. Gonzàlez i; CASADO, M. H. Bioinformática: interfaz web para estudiar el efecto de diferentes condiciones sobre la expresión de los genes. 2011.

MARTINS, M. P.; GUIMARÃES, L. N. F.; FONSECA, L. M. G. Classificador de texturas por redes neurais. In: *Anais do II Congresso Brasileiro de Computação, Itajaí-SC*. [S.l.: s.n.], 2002.

MEDEIROS, C.; COSTA, J. A. F. Uma comparação empírica de métodos de redução de dimensionalidade aplicados a visualização de dados. *Learning and Nonlinear Models-Revista da Sociedade Brasileira de Redes Neurais (SBRN)*, v. 6, n. 2, p. 81–110, 2008.

MERDUN, H. Modeling of pyrolysis product yields by artificial neural networks. *International Journal of Renewable Energy Research (IJRER)*, v. 8, n. 2, p. 1178–1188, 2018.

METZ, J. *Interpretação de clusters gerados por algoritmos de clustering hierárquico*. Tese (Doutorado) — Universidade de São Paulo, 2006.

MILARÉ, C. R. *Extração de conhecimento de redes neurais artificiais utilizando sistemas de aprendizado simbólico e algoritmos genéticos*. Tese (Doutorado) — Universidade de São Paulo, 2003.

MOSCATO, A. L. S. Análise exergética de uma caldeira de biomassa utilizando redes neurais artificiais. Universidade Estadual Paulista (UNESP), 2019.

NIETTO, P. R.; SAMPAIO, H. V. O uso do algoritmo de agrupamento hierárquico divisivo diana em uma rede de sensores sem fio aplicada à agricultura.

ÖZBAY, G.; KÖKTEN, E. S. Modeling of bio-oil production by pyrolysis of woody biomass: Artificial neural network approach. *Politeknik Dergisi*, v. 23, n. 4, p. 1255–1264, 2020.

PALMA, L. Agrupamento de dados: k-médias. *Universidade Federal do Recôncavo da Bahia Centro de Ciências Exatas e Tecnológicas*, 2018.

QUEIROZ, S. S. F.; PINTO, K. L. N. Extração de características e reconhecimento de padrões e objetos. *VETOR-Revista de Ciências Exatas e Engenharias*, v. 24, n. 2, p. 2–13, 2014.

RAJ, R. *Supervised, Unsupervised Learning and Semi-Supervised Learning with reallife usecase*. 2023. Urlhttps://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning.

RODRIGUES, F. S. Métodos de agrupamento na análise de dados de expressão gênica. Universidade Federal de São Carlos, 2009.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987.

SANTANA, F. B. D. Floresta aleatória para desenvolvimento de modelos multivariados de classificação e regressão em química analítica. [sn], 2020.

SCHUBERT, E.; ROUSSEEUW, P. J. Fast and eager k-medoids clustering: O(k) runtime improvement of the pam, clara, and clarans algorithms. *Information Systems*, v. 101, p. 101804, 2021. ISSN 0306-4379. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0306437921000557⟩.

SILVA, D. M. da; BRITO, J. A. de M.; OLIVEIRA, C. S. Um estudo computacional comparativo entre algoritmos de agrupamento e de detecção de comunidades.

SILVA, I. N. d.; SPATTI, D. H.; FLAUZINO, R. A. Redes neurais artificiais para engenharia e ciências aplicadas. 2010.

SOARES, I.; OLIVEIRA, C. S.; BRITO, J. A. de M. Um estudo do problema de detecção de comunidades em redes. *Sistemas & Gestão*, v. 9, n. 4, p. 566–574, 2014.

SOUTO, M. D. et al. Técnicas de aprendizado de máquina para problemas de biologia molecular. *Sociedade Brasileira de Computaçao*, v. 1, n. 2, 2003.

TEAM, R. C. *R Core Team R. R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria; 2016*. [S.l.]: ISBN 3–900051–07–0, URL http://www. R–project. org/. Available: http://www ..., 2016.

VIEIRA, G. E. G. et al. Biomassa: uma visão dos processos de pirólise. *Revista Liberato*, v. 15, n. 24, p. 167–178, 2014.

XAVIER, V. L. *Resolução do Problema de Agrupamento segundo o Critério de Minimização da Soma de Distâncias*. Tese (Doutorado) — M. Sc. Thesis—COPPE—UFRJ, Rio de Janeiro, 2012.

XAVIER, V. L. et al. Escalonamento multidimensional local: Uma abordagem via suavização hiperbólica. *Cadernos do IME-Série Estatística*, v. 44, p. 37, 2018.