Revista Internacional de Ciências,
Rio de Janeiro, v. 13, n. 02, p. 124 - 139, mai-ago, 2023
http://www.e-publicacoes.uerj.br/ojs/index.php/ric

e-ISSN 2316-7041

DOI: 10.12957/ric.2023.71261

Avaliação de Métodos de Agrupamentos em Dados de Biomassa Considerando os Diferentes Tipos de Pirólise

Sabrinna Rodrigues de Oliveira de Souza[1]; *Vinicius Layter Xavier*[1]; *Raquel Escrivani Guedes*[1]; *Alexandre Rodrigues Torres*[1]; *Aderval Severino Luna*[1]; *Marcello Montillo Provenza*[1]

✉sabrinnardol23@gmail.com

1. *Universidade do Estado do Rio de Janeiro, Brasil.*

Histórico do Artigo: O autor detém os direitos autorais deste artigo.

Recebido em: 16 de novembro de 2022 Aceito em: 18 de maio de 2022 Publicado em: 31 de agosto de 2022

Resumo: Este estudo aborda um problema de classificação de dados de Biomassa. Um dos objetivos é identificar as variáveis mais relevantes para a classificação do tipo de pirólise de Biomassa. Além disso, avaliar se as classes dos tipos de pirólise são suficientes para caracterizar esse processo químico. O algoritmo de Floresta Aleatória foi aplicado para identificar quais variáveis são relevantes no processo de classificação do tipo de pirólise, obtendo uma acurácia 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. Com as variáveis mais relevantes, os seguintes métodos de agrupamentos foram utilizados: k-means, pam, clara, diana, fanny, hierárquico, som, sota, model. Para avaliar os métodos de agrupamentos, foram utilizadas as medidas de validação interna com as métricas de índice Dunn e silhouette. As medidas de validação indicaram o agrupamento hierárquico e k-means com melhores resultados para número de grupos maior do que o número de classes de pirólise já existentes. Desta forma, o conjunto de dados deve ser dividido em um número maior de grupos de tipo de pirólise, pois considerar apenas as classes disponíveis é muito limitado para caracterizar o tipo de pirólise, uma vez que os algoritmos de classificação não supervisionado indicam o número de agrupamentos como maiores ou iguais a cinco.
Palavras-chave: Cluster, Inteligência Artificial, Biomassa, Programa R.

Evaluation Of Grouping Methods In Biomass Data Considering The Different Types Of Pyrolysis

Abstract: This study addresses a Biomass data classification problem. One of the objectives is to identify the most relevant variables for the classification of 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 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*. With the most relevant variables, the following clustering methods were used: k-means, pam, clear, diana, fanny, hierarchical, sound, sota, 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 number of groups greater than the number of existing pyrolysis classes. Thus, the dataset should be divided into a larger number of pyrolysis type groups, as considering only the available classes is 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, Artificial intelligence, Biomass, R program.

Evaluación De Métodos De Agrupación En Datos De Biomasa Considerando Los Diferentes Tipos De Pirólisis

Resumen: Este estudio aborda un problema de clasificación de datos sobre biomasa. Uno de los objetivos es identificar las variables relevantes para la clasificación del tipo de pirólisis de la Biomasa. Además, evaluar si las clases de tipo de pirólisis son suficientes para caracterizar este proceso químico. Se aplicó el algoritmo Random Forest para identificar qué variables son relevantes en el proceso de clasificación del tipo de pirólisis, obteniendo una precisión en torno al 97%. Se identificó que las variables importantes son: *Tiempo medio de residencia en el reactor para el gas y el portador, Porcentaje de carbono en base seca sin cenizas en la materia prima, Tamaño medio de las partículas en el reactor y Porcentaje de hidrógeno en base seca sin cenizas en la materia prima.* Con las variables más relevantes, se utilizaron los siguientes métodos de agrupación: k-means, pam, clear, diana, fanny, jerárquico, sound, sota, model. Para evaluar los métodos de agrupación, se utilizaron medidas de validación interna con las métricas de índice de Dunn y silueta. Las medidas de validación indicaron la agrupación jerárquica y k-means con mejores resultados para un número de grupos superior al número de clases de pirólisis existentes. Por lo tanto, el conjunto de datos debería dividirse en un mayor número de grupos de tipo de pirólisis, ya que considerar únicamente las clases disponibles es muy limitado para caracterizar el tipo de pirólisis, puesto que los algoritmos de clasificación no supervisada indican que el número de conglomerados es mayor o igual a cinco.
Palabras clave: Clúster, Inteligencia artificial, Biomasa, Programa R

INTRODUÇÃO

Biomassa é um termo utilizado para 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 combustíveis e produtos químicos e biomateriais. A pirólise é a decomposição térmica que ocorre em concentrações de 02 muito baixas e tem sido aceita como um método promissor em relação ao custo razoável e operação simples para conversão de biomassa (GUEDES et al., 2018).

O processo de pirólise é dividido em quatro tipos, dependendo das condições de operação utilizadas: pirólise Fast, Slow, Catalytic e Flash. A pirólise Slow ocorre a uma temperatura de processo mais baixa, menor taxa de aquecimento e maiores tempos 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 (GUEDES et al., 2018). A pirólise Catalityc serve para melhorar a qualidade do óleo produzido. Os tipos de pirólise são agrupados nas respectivas classes (Fast, Slow, Flash e Catalityc) de acordo com as variáveis químicas, representadas na tabela 1 na próxima seção.

Este banco de dados 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 (GUEDES et al., 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.

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 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 desenvolveu um modelo confiável de rede neural artificial para modelar o produto líquido da pirólise, considerando os tipos de pirólise fast e slow. Outro exemplo sobre redes neurais é o estudo do autor CAO et al. (2016) que desenvolveu modelos de Inteligência Artificial, baseados em redes neurais artificiais e máquina de vetor de suporte, para prever a distribuição de produtos e HHV 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. Este é o caso do estudo realizado por CAO et al. (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. Este trabalho tem como diferencial realizar o que não é encontrado na literatura sobre 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.

Neste estudo, é realizada uma aplicação do algoritmo de Floresta Aleatória no intuito de identificar as variáveis mais relevantes para a classificação do tipo de pirólise. Com as

variáveis que foram identificadas como mais eminentes, 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. A medida de validação indicou o agrupamento Hierárquico e k-means com os melhores resultados.

MATERIAL E MÉTODOS

O Banco de dados utilizado para este estudo é composto por dados referentes ao processo de biomassa, fornecido por (GUEDES *et al.*, 2018). Os tipos de pirólise são agrupados nas respectivas classes de acordo com as doze variáveis químicas descritas na tabela l:

Tabela l. Variáveis que agrupam os tipos de pirólise.

Variável	Descrição	Número da variável
p_c_mp	Porcentagem de carbono em base seca livre de cinza na matéria-prima	1
p_h_mp	Porcentagem de hidrogênio em base seca livre de cinza na matéria-prima	2
p_n_mp	Porcentagem de nitrogênio em base seca livre de cinza na matéria-prima	3
p_o_mp	Porcentagem de oxigênio em base seca livre de cinza na matéria-prima	4
p_umid_mp	Porcentagem de umidade na matéria-prima	5
p_cfix_mp	Porcentagem de carbono fixo em base seca da matéria-prima	6
p_cinz_mp	Porcentagem de cinzas em base seca da matéria-prima	7
p_vola_mp	Porcentagem de 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

Fonte: GUEDES *et al.*, 2018.

Para este estudo é considerado os casos de regime de operação do reator igual a "Batelada" e o tipo do reator igual a "Fixed bed", dos quais contemplam os tipos de pirólese: Fast, Slow e Catalytic.

Diante dos dados de Biomassa disponíveis, este estudo 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. O Aprendizado de Máquina é um

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

O método não supervisionado, é composto por dados não rotulados. No aprendizado não supervisionado, os dados não possuem uma classe equivalente. Nesse caso são analisados os dados fornecidos 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 E, no qual cada dado consiste somente de vetores x, não incluindo a informação sobre a classe y à qual ele pertence.

No aprendizado supervisionado são utilizados exemplos previamente rotulados, ou seja, é fornecido ao sistema um conjunto de dados $E = E_1, E_2, \dots, E_n$, sendo que cada dado $E_i \in E$ possui um rótulo associado (MILARÉ, 2003). Este rótulo define a classe a qual o dado pertence, ou seja, cada dado $E_i \in E$ é uma tupla: $E_i = (x_i, y_i)$ onde x_i é um vetor de valores que representam as características de E_i e y_i é o valor da classe desse dado (MILARÉ, 2003).

Uma das abordagens iniciais neste trabalho é o algoritmo de aprendizado supervisionado Floresta Aleatória. Este algoritmo é uma ferramenta popular de aprendizado de máquina baseada em árvore que é altamente adaptável a dados. Isso o torna particularmente atraente para análise de dados de alta dimensão.

Floresta Aleatória geralmente é uma coleção de centenas a milhares de árvores, onde cada árvore é cultivada usando uma amostra inicial dos dados originais (bootstrap) (CHEN; ISHWARAN, 2012).

A Floresta Aleatória tem sido tradicionalmente aplicada em configurações de classificação e regressão. A construção é descrita nas seguintes etapas:

1. Gera n árvores com amostras de bootstrap dos dados originais.
2. Cria uma árvore para cada conjunto de dados de bootstrap. Em cada nó da árvore, seleciona aleatoriamente um conjunto de variáveis para a divisão.
3. Agrega as informações do conjunto de árvores para previsão de novos dados com a votação da maioria para classificação.
4. Calcular um indicador de medida de erro out-of-bag (OOB) usando os dados que não estão na amostra bootstrap.

Os dados que não pertencem a amostra bootstrap, são chamados de ***out-of-bag*** (OOB), do qual é utilizado para estimar o desempenho do modelo. A partir destes dados, é possível medir a taxa de erro do modelo de predição (BREIMAN, 2001). No bootstrap, como as amostras são retiradas com reposição, algumas observações são repetidas e outras são deixadas de fora da amostra (KHAN *et. al.*, 2021).

É possível encontrar as variáveis de mais relevância por meio do método de permutação, do qual avalia a diminuição média na acurácia e pelo índice Gini, do qual avalia a diminuição média na impureza do nó (James *et al.,* 2013).

As análises de agrupamento ou clustering, técnica não supervisionada, 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. Desta forma temos dois princípios básicos da análise de agrupamento que são homogeneidade e separação. Sendo assim, quanto mais homogêneos são os elementos dentro de um grupo, mais separados ou diferentes são os grupos (XAVIER, 2012). 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 clusters e, essencialmente, buscam otimizar uma função objetivo, tendo uma avaliação da homogeneidade dentro do agrupamento (XAVIER, 2012).

Para este trabalho, foram utilizados nove algoritmos de agrupamentos, são eles: k-means, sota, clara, diana, pam, fanny, hierárquico, model. 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 cluster como entrada e usam informações intrínsecas nos dados para avaliar a qualidade do agrupamento (BROCK *et al.*, 2008). Para este estudo foi utilizada essa medida de validação.

Para a validação interna, utilizou-se o índice de Dunn e o valor da Silhouette. Estes métodos são ambos os exemplos de combinações não lineares da compacidade e separação (BROCK *et al.*, 2008).

Existem diversos índices de validação de cluster que podem ser utilizados junto com os métodos de agrupamento para encontrar um valor ótimo de k. Mas, o silhouette 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 de silhouette 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 serão mal agrupadas (ROUSSEEUW, 1987). Para a observação i, ela é definida como:

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (1)$$

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

$$a_i = \frac{1}{n(C(i))} \sum_{j \in C(i)} dist(i,j), \; b_i = min'_{c_k \in \varsigma \setminus C(i)} \sum_{j \in C_k} \frac{dist(i,j)}{nC_k} \quad (2)$$

onde c_i é o cluster contendo a observação i, dist (i, j) é a distância (por exemplo, Euclidiana, Manhattan etc.) entre as observações i e j, e n(C) é a cardinalidade do cluster C. O tamanho da silhueta é a média do valor da silhueta de cada observação. A largura da silhueta, portanto, está no intervalo [-1, 1], e deve ser maximizada (BROCK et al., 2008).

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

$$D = \frac{= min_{C_k, C_l \in \varsigma, C_k \neq C_l}(min_{i \in C_k, j \in C_l} dist(i,j))}{max_{C_m \in \varsigma} diam(C_m)} \quad (3)$$

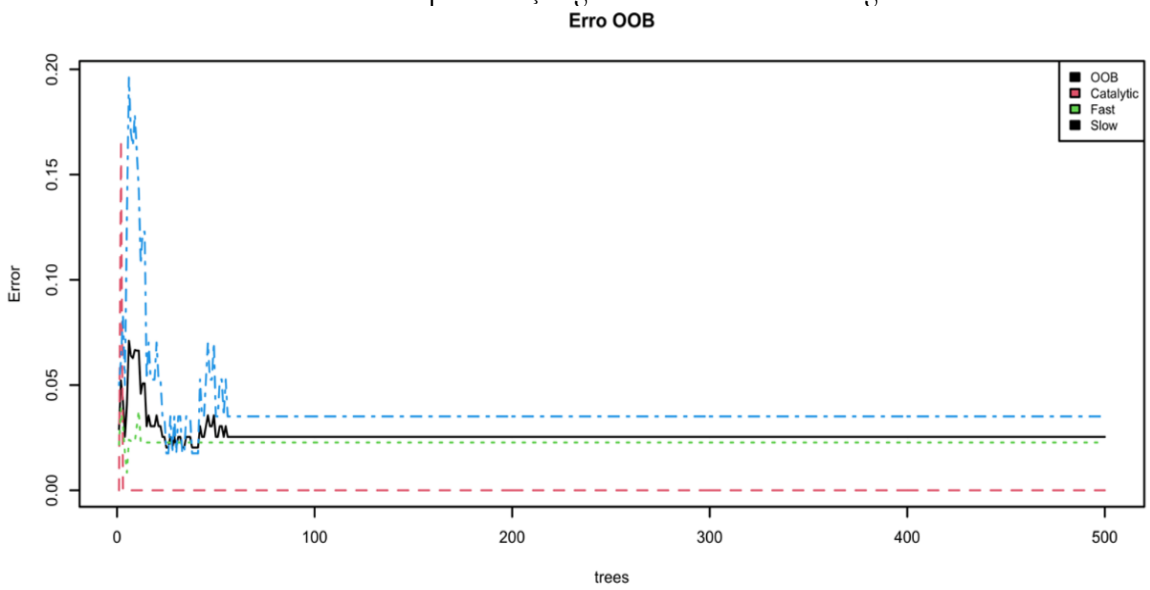
RESULTADOS E DISCUSSÃO

Em toda a análise de dados foi utilizado o programa R, livre e open source, Versão 1.2.5033 (R CORE TEAM, 2022). 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. Dado que o objetivo consiste em identificar grupos de tipos de pirólise com variáveis de processos semelhantes, foi realizada a leitura e normalização dos dados. 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 1000 árvores com quatro variáveis. Além disso, analisando os valores de acurácia e kappa, foram obtidos os valores em torno de 97% para acurácia e em torno de 95% para o Kappa, para número de variáveis igual a quatro e número de árvores igual a 1000.

Por meio do erro out-of-bag (OOB) obtido no gráfico 1, é possível observar o erro geral na linha preta. Este 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.

Gráfico 1. Representação gráfica do erro Out-of-bag

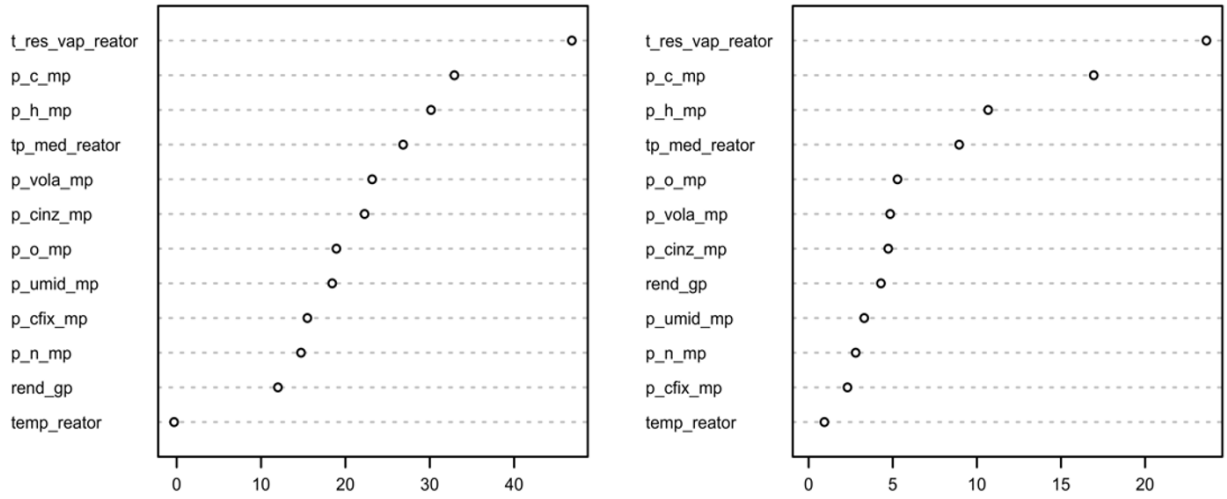


Fonte: os autores, 2022.

Com o número de variáveis e o número de árvores definido foi possível identificar as variáveis de mais importância por meio da função Floresta Aleatória. Os resultados da função Floresta Aleatória obtidos foram: um erro geral de 2,54% para o erro OOB com 1000 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 importância foram obtidas por meio do método de permutação e pelo índice de Gini. Conforme exibido no gráfico 2 e por meio dos valores obtidos na tabela 2, é possível perceber que através do Índice de Gini as variáveis *10, 1, 9 e 2* (enumeradas na tabela 1) foram as variáveis com mais importância. Portanto, essas foram as variáveis utilizadas para as análises de agrupamento.

Gráfico 2. Representação gráfica das variáveis de mais importância



Fonte: os autores, 2022.

Tabela 2. Resultado das variáveis de mais importância

Variáveis	Erro médio quadrático	Índice de Gini
Tempo de residência médio no reator para o gás (t-res-vap-reator)	46.840	23.646
Porcentagem de carbono em base seca livre de cinza na matéria-prima (p-c-mp)	32.917	16.947
Tamanho da partícula média do reator (tp-med-reator)	26.843	8.947
Porcentagem de hidrogênio em base seca livre de cinza na matéria-prima (p-h-mp)	30.147	10.670

Fonte: os autores, 2022.

Foram realizadas também as etapas de predição, as quais determinam uma classificação para o conjunto de dados teste. Todas as etapas nessa fase se assemelham às etapas utilizadas em outros algoritmos de problemas de classificação. Foi feita a predição e gerada a matriz de confusão. Foi obtida uma sensibilidade em torno de 92%, uma especificidade em torno de 96% e uma acurácia em torno de 97%.

Para identificar o número ideal de grupos para os dados analisados foi feita a variação no número de 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 silhouette e o índice de Dunn. Vale ressaltar que o valor da silhouette traz informações sobre a "consistência" e a

"separação" de cada grupo. Um método de agrupamento mais adequado a uma determinada distribuição deve apresentar um valor da silhouette próximo de 1, podendo variar de -1 a 1. Na tabela 3 podemos ver os resultados obtidos por esta medida de validação.

Tabela 3. Resultado da medida de validação interna (indíce dunn e silhouette)

Método de agrupamento	Métrica associada	Número de grupos (k)						
		2	3	4	5	6	7	8
Hierárquico	Dunn	0,394	0,346	0,494	0,494	0,572	0,572	0,368
	Silhouette	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
	Silhouette	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
	Silhouette	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
	Silhouette	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
	Silhouette	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
	Silhouette	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
	Silhouette	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
	Silhouette	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
	Silhouette	0,357	0,438	0,429	0,276	0,493	0,422	0,411

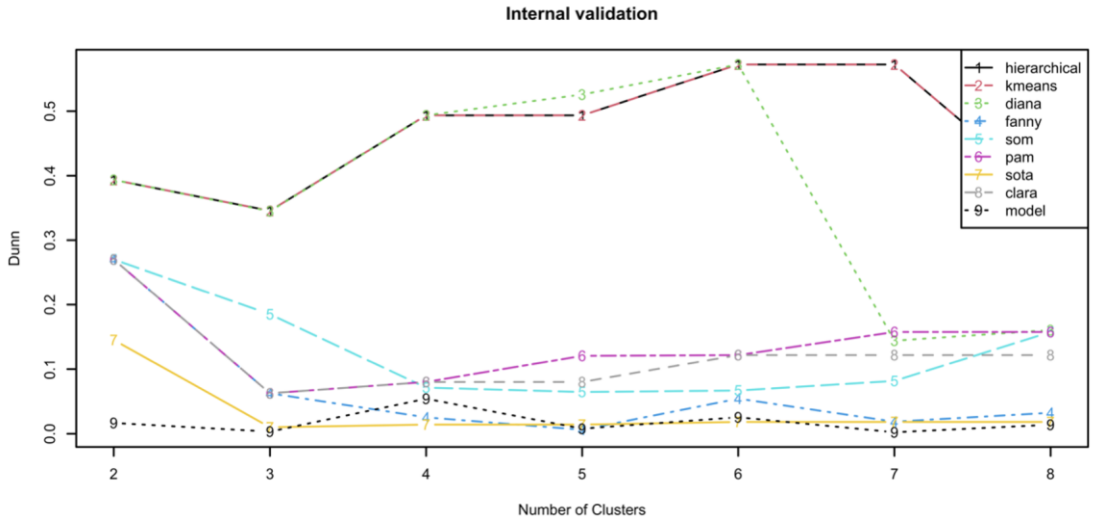
Fonte: os autores, 2022.

As medidas de validação também foram exibidas nos gráficos 3 e 4. Vale lembrar que o índice de Dunn e a Silhouette devem ser maximizadas. Desta forma, é possível perceber que o método hierárquico e k-means, com seis e sete clusters, e diana, com seis clusters, obtiveram

os melhores resultados para o índice Dunn. Já para o silhouette, o método hierárquico e k-means, com cinco clusters, obtiveram os melhores desempenhos. Ou seja, quanto maior o valor do índice Dunn e silhouette, melhor será o desempenho (BROCK et al., 2008).

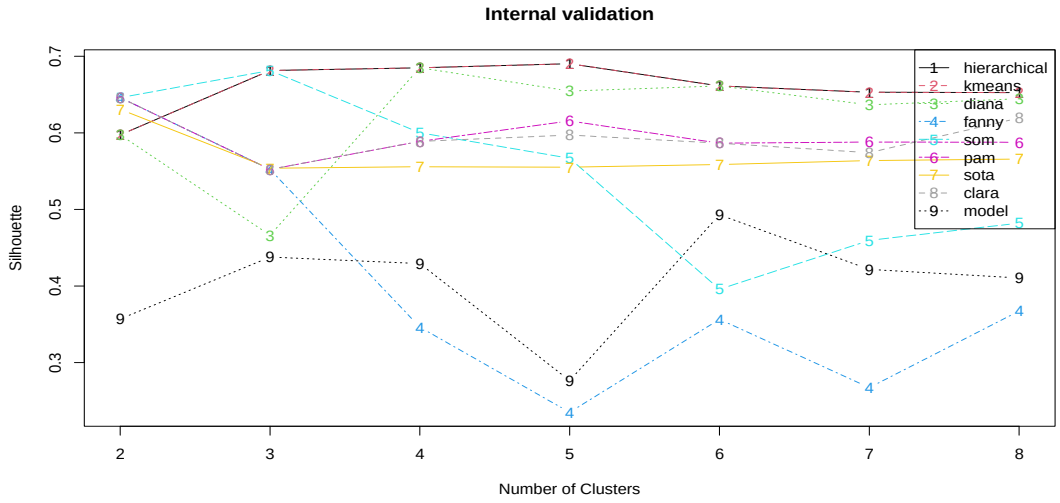
Pode-se observar também que o clustering baseado em modelo se destaca negativamente em todos os clusters analisados, não funcionando bem em nenhuma das medidas. Independentemente do algoritmo de agrupamento, o número ideal de grupos parece ser seis ou mais para o índice dunn e cinco usando o silhouete.

Gráfico 3. Representação gráfica do índice Dunn



Fonte: os autores, 2022.

Gráfico 4. Representação gráfica do silhouette



Fonte: os autores, 2022.

Na tabela 4, são mostrados os resultados consolidados obtidos por meio da validação interna com as respectivas métricas, índice Dunn e silhouette

Tabela 4. Resultados da validação Interna

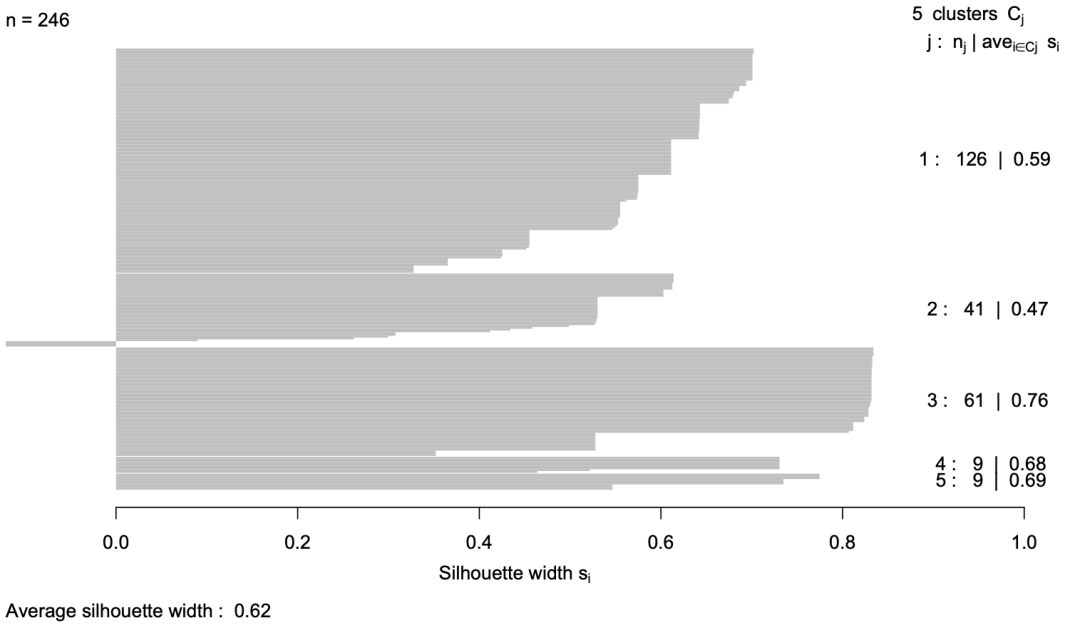
Indicador	Interna
Dunn	Hierárquico/k-means/Diana com k = 6 e K = 7
Silhouette	Hierárquico/k-means com k = 5

Fonte: os autores, 2022.

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

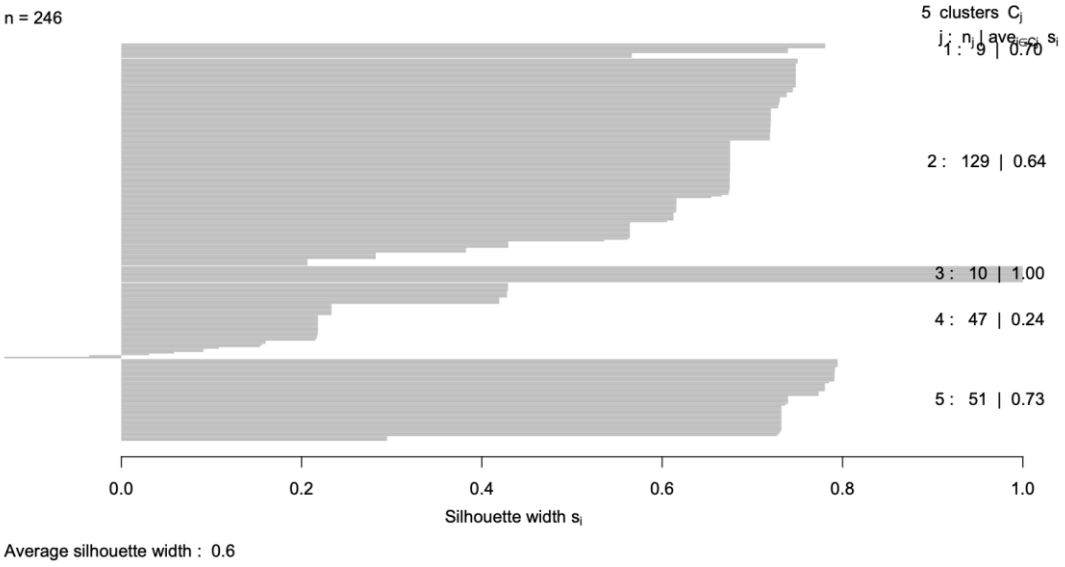
Foi gerado também o gráfico de silhouette, com base nos métodos hierárquico (gráfico 5) e k-means (gráfico 6), para k = 5. A maioria dos valores são próximos de 1, o que sugere que a observação é bem compatível com o cluster atribuído tanto para hierárquico quanto para k-means.

Gráfico 5. Representação gráfica do Silhouette para método hierárquico



Fonte: os autores, 2022.

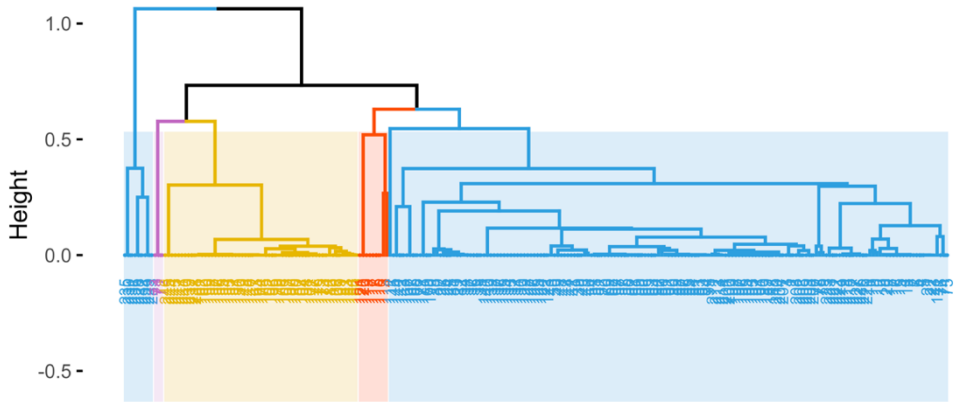
Gráfico 6. Representação gráfica do Silhouette para método k-means



Fonte: os autores, 2022.

Em seguida, foram aplicados os métodos Hierárquico e k-means, diante do bom desempenho obtido na medida de validação. Primeiro, 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 no gráfico 7, pertencentes às classes funcionais "Fast", “Slow” e “Catalytic”, previamente rotuladas. Foram obtidos cinco grandes ramos ou agrupamentos que emergem do dendrograma, sendo que eles são bem distintos, conforme gráfico 7:

Gráfico 7. Representação gráfica do algoritmo Hierárquico



Fonte: os autores, 2022.

Em seguida, foram identificados os agrupamentos em que foram classificados os dados de biomassa para os métodos Hierárquico (tabela 5) e k-means (tabela 6). Em ambos os métodos, é possível identificar grupos puros, 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.

Tabela 5. Identificação dos agrupamentos em que foram classificados os dados de biomassa para o método Hierárquico

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

Fonte: os autores, 2022.

Tabela 6. Identificação dos agrupamentos em que foram classificados os dados de biomassa para o método k-means

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

Fonte: os autores, 2022.

CONCLUSÃO

Com os resultados obtidos, foi possível concluir que o agrupamento hierárquico e k-means possuem desempenhos consistentes para duas das métricas da validação geradas, Dunn e silhouete. Já o valor de k, o mais indicado é que seja um valor maior que 5, pois os grupos naturais formados pelos métodos de agrupamento, são bem distintos das classes associadas aos dados.

Sendo assim, conclui-se que é necessário dividir o conjunto de dados em um número maior de grupos de tipos de pirólise, já que as classes previamente já rotuladas e fornecidas no banco de dados são muito limitadas para caracterizar o tipo de pirólise.

AGRADECIMENTOS

Este estudo foi feito com a colaboração de Vinicius Layter Xavier e Marcello Montillo Provenza nas aplicações das técnicas estatísticas e análise de resultados. Alexandre Rodrigues Torres, Aderval Severino Luna e Raquel Escrivani Guedes apoiaram na coleta, extração e manipulação dos dados, assim como entendimento dos processos químicos da biomassa. A redação do texto foi feita por Sabrinna Rodrigues de Oliveira de Souza, Vinicius Layter Xavier e Aderval Severino Luna. A revisão do texto foi realizada por todos os autores.

REFERÊNCIAS BIBLIOGRÁFICAS

BATOOL, FATIMA; HENNIG, CHRISTIAN. Clustering with the average silhouette width. **Computational Statistics & Data Analysis**, v. 158, p. 107190, jun. 2021.

BREIMAN, L. **Random forests**. Machine learning v. 45.1, p. 5-32, 2001.

BROCK, G.; PIHUR, V.; DATTA, S.; DATTA, S. clValid: An R Package for Cluster Validation. **Journal of Statistical Software**, [S. l.], v. 25, n. 4, p. 1–22, mar. 2008.

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

CHEN, Xi; ISHWARAN, HEMANT. Random forests for genomic data analysis. **Genomics**, v. 99, n. 6, p. 323–329, abr. 2012.

DE SOUTO, M. C. P., LORENA, A. C., DELBEM, A. C. B., & DE CARVALHO, A. C. P. L. F. Técnicas de aprendizado de máquina para problemas de biologia molecular. **Sociedade Brasileira de Computação,** v.1, n. 2, 2003.

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

JAMES, GARETH, DANIELA WITTEN, TREVOR HASTIE & ROBERT TIBSHIRANI. **An introduction to statistical learning: with applications in R**. 2 ed., New York: Springer, 2013.

KHAN, ZARDAD; GUL, NAZ; FAIZ, NOSHEEN; GUL, ASMA, WERNER ADLER AND BERTHOLD LAUSEN, Optimal Trees Selection for Classification via Out-of-Bag Assessment and Sub-Bagging. ***IEEE Access***, vol. 9, pp. 28591-28607, fev. 2021.

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

MILARÉ, CLAUDIA REGINA. **Extração de conhecimento de redes neurais artificiais utilizando sistemas de aprendizado simbólico e algoritmos genéticos**. 2003. Tese (Doutorado em Ciências da Computação e Matemática Computacional) — Universidade de São Paulo, São Paulo.

MOSCATO, ANDRÉ LUIZ SALVAT. **Análise exergética de uma caldeira de biomassa utilizando redes neurais artificiais**. 2019. Tese (Doutorado em Engenharia Mecânica) – Universidade Estadual Paulista, São Paulo.

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

R CORE TEAM. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**. Vienna, Austria, 2022. Disponível em <https://www.R-project.org>.

ROUSSEEUW PJ. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 56, nov. 1987.

XAVIER, VINICIUS LAYTER. **Resolução do Problema de Agrupamento segundo o Critério de Minimização da Soma de Distâncias**. 2012. Dissertação (Mestrado em Engenharia de Sistemas e Computação) – Universidade Federal do Rio de Janeiro, Rio de Janeiro.