MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Especialização em Ciência de Dados

Perfis Físico-Químicos e de Polpação de Clones de Eucaliptos: Uma Análise Baseada em Técnicas Não Supervisionadas

Felipe Guerra Carneiro

João Monlevade, MG
2024

Felipe Guerra Carneiro

Perfis Físico-Químicos e de Polpação de Clones de Eucaliptos: Uma Análise Baseada em Técnicas Não Supervisionadas

Trabalho de conclusão de curso apresentado ao curso de Ciência de Dados do Instituto de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto, como parte dos requisitos necessários para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Carlos Henrique Gomes Ferreira

João Monlevade, MG

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

C289p Carneiro, Felipe Guerra.
Perfis físico-químicos e de polpação de clones de eucaliptos [manuscrito]: uma análise baseada em técnicas não supervisionadas. / Felipe Guerra Carneiro. - 2024.
42 f.: il.: , gráf., tab..

Orientador: Prof. Dr. Carlos Henrique Gomes Ferreira.
Produção Científica (Especialização). Universidade Federal de Ouro Preto. Departamento de Engenharia de Produção.

1. Análise multivariada. 2. Análise por agrupamento. 3. Celulose. 4. Estatística matemática - Algoritmos. 5. Eucalipto. 6. Polpa de madeira - Branqueamento. I. Ferreira, Carlos Henrique Gomes. II. Universidade Federal de Ouro Preto. III. Título.

CDU 519.2:004.021

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Felipe Guerra Carneiro

PERFIS FÍSICO-QUÍMICOS E DE POLPAÇÃO DE CLONES DE EUCALIPTOS:
UMA ANÁLISE BASEADA EM TÉCNICAS NÃO SUPERVISIONADAS

Trabalho de conclusão de curso apresentado ao curso de Especialização em Ciência de Dados
da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Especialista em Ciência de Dados

Aprovada em 09 de Maio de 2024

Membros da banca

Dr. - Carlos Henrique Gomes Ferreira - Orientador(a) Universidade Federal de Ouro Preto
Me. - Ronaldo Neves Ribeiro - Celulose Nipo-Brasileira S/A
Dr. - Thiago Augusto de Oliveira Silva - Universidade Federal de Ouro Preto

Carlos Henrique Gomes Ferreira, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 08/07/2024

Documento assinado eletronicamente por **Carlos Henrique Gomes Ferreira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 08/07/2024, às 13:14, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0732974** e o código CRC **DB553D39**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

Agradecimentos

A Deus pelo sentido da vida, à minha família pelo amor diário e aos meus colegas de trabalho pela caminhada.

“Com a sabedoria se edifica a casa, e com o entendimento ela se estabelece", Provérbios 24.3

Resumo

Este trabalho de conclusão de curso investigou a aplicabilidade de técnicas de agrupamento não supervisionado para classificar clones de eucalipto, considerando características químicas, físicas e do processo de polpação e branqueamento KRAFT. A principal matéria-prima para a produção de celulose branqueada no Brasil, o eucalipto, apresenta uma composição complexa que desafia a otimização dos processos industriais devido à sua variabilidade genética e ambiental. O objetivo geral deste estudo foi aplicar e avaliar a eficácia das técnicas de agrupamento não supervisionado para melhor entender as relações entre as propriedades dos clones de eucalipto e a qualidade da celulose produzida. Especificamente, o trabalho focou em selecionar e aplicar técnicas de agrupamento adequadas para a caracterização de clones e analisar os agrupamentos formados para identificar padrões e relações relevantes. Utilizou-se o algoritmo *HDBSCAN* combinado com a técnica de redução de dimensionalidade *UMAP* para explorar os padrões nos dados multidimensionais. Esta combinação mostrou-se particularmente eficaz em identificar grupos com características homogêneas e distintas entre os clones analisados. Adicionalmente, implementou-se um índice de pontuação ('*score*') baseado em critérios pré-definidos para classificar os grupos quanto ao seu potencial para produção de celulose. Por fim, avaliou-se as variáveis mais discriminativas do agrupamento obtido. Os resultados indicaram que o agrupamento não supervisionado, complementado pela análise do índice de Gini, oferece insights valiosos sobre a variabilidade da matéria-prima, que podem ser utilizados para otimizar o processo de seleção de clones e orientar o abastecimento segregado de madeira.

Palavras-chave: Eucaliptos, Celulose KRAFT, Agrupamento Não Supervisionado, *HDBSCAN*, *UMAP*, Análise Multivariada, Seleção de Clones, Índice de Gini.

Abstract

This capstone project investigated the applicability of unsupervised clustering techniques for classifying Eucalyptus clones, considering chemical, physical, and Kraft pulping and bleaching process characteristics. Eucalyptus, the primary raw material for the production of bleached pulp in Brazil, presents a complex composition that challenges the optimization of industrial processes due to its genetic and environmental variability. The general objective of this study was to apply and evaluate the effectiveness of unsupervised clustering techniques to better understand the relationships between the properties of Eucalyptus clones and the quality of the pulp produced. Specifically, the project focused on selecting and applying suitable clustering techniques for characterizing clones and analyzing the clusters formed to identify relevant patterns and relationships. The HDBSCAN algorithm, combined with the UMAP dimensionality reduction technique, was used to explore patterns in the multidimensional data. This combination proved particularly effective in identifying groups with homogeneous and distinct characteristics among the analyzed clones. Additionally, a scoring index ('score') based on predefined criteria was implemented to classify the groups according to their potential for producing high-quality pulp. The analysis of Gini indices for the considered variables revealed that certain characteristics, such as the concentration of specific chemical elements, have high discriminatory power, significantly contributing to the differentiation between groups. The results indicated that unsupervised clustering, complemented by Gini index analysis, provides valuable insights into the raw material variability, which can be utilized to optimize the clone selection process and guide the segregated wood supply.

Keywords: Eucalyptus, Kraft Pulp, Unsupervised Clustering, HDBSCAN, UMAP, Multivariate Analysis, Clone Selection, Gini Index.

Sumário

1 Introdução

A celulose é uma commoditie utilizada tradicionalmente na fabricação de papéis para impressão e escrita, papel moeda, papel filtro, papel higiênico, lenços faciais, toalhas de papel, cartões e embalagens. Sua aplicação ainda se estende à indústria alimentícia, como estabilizante em sorvetes, e à produção de explosivos na forma de nitrocelulose. Recentemente, a celulose tem encontrado novos usos em tecnologias avançadas, incluindo a fabricação de nanocelulose para eletrônicos flexíveis, biomedicina e fibras de tecidos. Subprodutos do processo de produção, como a lignina e extrativos, também apresentam outros usos possíveis. A lignina, por exemplo, pode ser utilizada como biocombustível, aditivo em concretos e materiais compósitos. Essa versatilidade da celulose e seus subprodutos demonstra a potencialidade desta indústria e justifica o crescimento do setor (CAMPOS; FOELKEL, 2016).

O Brasil é reconhecido mundialmente como grande produtor e exportador de celulose, especialmente branqueada de fibras curtas. Em 2022, o Brasil produziu 22 milhões de toneladas de celulose de fibras curtas, das quais mais de 75% foram destinadas à exportação, mantendo a posição de maior exportador de celulose no mercado mundial, com 8,4 bilhões de dólares. No setor de árvores plantadas, a celulose representa 59% do valor exportado (Indústria Brasileira de Árvores (Ibá), 2023).

Os excelentes números apresentados podem ser explicados pelo constante investimento em pesquisa e desenvolvimento, tanto na indústria quanto na base florestal, bem como pelas condições edafoclimáticas do Brasil, associadas à constante evolução dos processos de melhoramento genético e manejo florestal. A principal matéria-prima para produção de polpa celulósica KRAFT branqueada de fibras curtas no Brasil é o eucalipto. A madeira representa 50 a 60 por cento do custo total da produção de celulose (MARINHO *et al.*, 2017; DEPEC-BRADESCO, 2017; EUROPE, 2011).

Durante muito tempo foram desprendidos esforços para alavancar as produtividades dos plantios de eucalipto com foco nos aspectos silviculturais. Nos últimos anos, uma atenção especial também tem sido dada às características da qualidade da madeira (aspectos físicos, químicos e anatômicos) e mais recentemente aos efeitos destas características no processo de produção de celulose (COLODETTE; MOUNTEER; GOMES, 2012; GOMIDE; NETO; REGAZZI, 2010; LEITE, 2006).

A madeira de eucalipto é constituída basicamente de carboidratos (celulose e hemiceluloses), lignina e extrativos. Para obtenção da polpa celulósica KRAFT, a madeira é submetida a um processo de polpação, sob temperatura e pressão, em uma solução alcalina chamada de licor de cozimento. Esse licor de cozimento é um reagente aquoso composto majoritariamente de hidróxido de sódio e sulfato de sódio. A quantidade de álcali (substâncias com características básicas) presentes no licor, que efetivamente participa do processo de cozimento, é denominada álcali efetivo (AE) (GOUVêA *et al.*, 2009; GOMIDE *et al.*, 2005).

Durante a polpação KRAFT busca-se remover a lignina e extrativos, preservando ao máximo os carboidratos. A eficiência da deslignificação nesta fase é expressa pelo número KAPPA, que representa o conteúdo de lignina residual presente na polpa. Outros indicadores da eficiência do processo são o rendimento depurado (relação entre o peso seco de polpa depurada e peso seco da madeira) e o álcali residual (diferença entre álcali efetivo inicial e álcali efetivo final). O produto da polpação é uma massa de fibras individualizadas de coloração marrom, principalmente devido à lignina residual ainda presente nas fibras. A etapa subsequente é o branqueamento, que consiste em remover os compostos cromóforos que conferem cor residual à polpa ao final do cozimento, bem como alterar ou remover os compostos responsáveis pela reversão de alvura. Para obtenção do produto acabado, a massa passa pelos processos de secagem e enfardamento (CARVALHO; SILVA; COLODETTE, 2014; SENAI. Serviço Nacional de Aprendizagem Industrial, 2013; MARANESI, 2010).

Do ponto de vista econômico, a qualidade da madeira é fator chave para o rendimento do processo, consumo de químicos e eficiência do uso e geração de energia. Por outro lado, a grande variabilidade dos componentes da madeira, sejam estes entre espécies, na mesma espécie ou ainda dentro da mesma árvore, torna o processo de produção de celulose e papel desafiador, especialmente devido ao grande número de materiais genéticos existentes na base florestal, necessários para o abastecimento do processo industrial (CARVALHO; SILVA; COLODETTE, 2014; ALMEIDA, 2010; TRUGILHO *et al.*, 2004).

Por outro lado, a análise dessas características pode ser desafiadora devido à sua natureza multivariada. Cada amostra de madeira possui um conjunto de características, cujas relações podem ser complexas e não lineares. Neste sentido, a análise multivariada torna-se ferramenta essencial para a classificação da madeira, superando as limitações da análise univariada, onde uma avaliação conjunta dos atributos da madeira proporciona uma visão abrangente da qualidade do material. Isso leva a uma classificação mais precisa e confiável, otimizando o processo ao identificar os atributos mais relevantes para seleção da madeira em diferentes aplicações (FEARON *et al.*, 2020; FARDIM; DURáN, 2004).

Neste contexto, o uso de algoritmos de agrupamento não supervisionado torna-se ferramenta valiosa, com potencial para revelar padrões nos dados e agrupar amostras de madeira com características semelhantes. Isso pode ajudar a entender melhor a relação entre essas características e a qualidade da celulose produzida (TRUGILHO *et al.*, 2004; CAIXETA *et al.*, 2003).

1.1 Objetivo geral

O objetivo deste trabalho é aplicar técnicas de agrupamento não supervisionado em características químicas, físicas e de polpação KRAFT, provenientes das análises laboratoriais de clones de eucalipto.

1.1.1 Objetivos específicos

- Avaliar técnicas de agrupamento não supervisionado para agrupamento de clones de eucalipto;
- Analisar os agrupamentos formados e a relação entre as características da madeira, buscando identificar padrões e relações relevantes;

1.2 Contribuições

A hipótese deste trabalho é que a aplicação de técnicas de agrupamento não supervisionado em características químicas, físicas e de polpação KRAFT de clones de eucalipto pode revelar padrões e relações significativas. Esses padrões e relações podem contribuir para uma melhor compreensão das propriedades da madeira e a qualidade da polpa produzida. Além disso, espera-se que essas descobertas possam auxiliar na otimização do processo de produção de celulose, levando a processos mais eficientes e à criação de novos produtos.

1.3 Organização do Trabalho

O restante deste documento está organizado da seguinte forma: Capítulo 2 realiza uma revisão da literatura sobre as características da madeira de eucalipto, com foco em suas propriedades químicas, físicas e de polpação KRAFT. Aborda as principais técnicas de agrupamento não supervisionado utilizadas no trabalho. No Capítulo 3 são descritos os passos realizados nos agrupamentos dos materiais genéticos analisados, por meio de aprendizagem não supervisionada, envolvendo desde a seleção e o pré-processamento de dados ao agrupamento e avaliação. O Capítulo 4 apresenta os resultados e métricas dos agrupamentos obtidos no conjunto de dados da empresa. Descreve os agrupamentos formados, detalhando suas características e comparando-os com o método atualmente utilizado para classificação destes materiais. O Capítulo 5 resume os principais resultados do trabalho e responde à pergunta de pesquisa. Confirma ou refuta a hipótese inicial e destaca as contribuições do estudo para a área de conhecimento. O capítulo também apresenta as aplicações práticas dos resultados, reforçando a relevância e propostas de trabalhos futuros.

2 Revisão da Literatura

2.1 Características dos Clones de Eucalipto

A caracterização tecnológica de clones de eucalipto é um processo complexo que envolve o estudo das propriedades físicas, químicas e anatômicas, bem como das propriedades resultantes da polpação, branqueamento e produção de papel. Estas características podem variar dependendo da região, da idade, do tempo pós-corte, das condições de crescimento e do processamento da madeira. Pode ainda variar entre espécies ou ainda na própria árvore. A caracterização da madeira pode envolver uma série de análises laboratoriais como: densidade, frações de carboidratos, lignina, extrativos, minerais, anatômicas, além de análises mecânicas, de polpação e do licor de cozimento (CARVALHO; SILVA; COLODETTE, 2014; GOMIDE *et al.*, 2005). Na sequência, serão abordadas as principais características de interesse neste estudo.

2.1.1 Características físicas da madeira

A Densidade Básica (DB) representa quanto de material lenhoso, em peso (massa), se tem por volume. Esta informação tem sido muito utilizada para expressar a qualidade da madeira, sobretudo pela facilidade de sua determinação. Alguns estudos mostram boa correlação da DB com consumo específico (NETO, 2012), por outro lado o uso da DB tem sido muito questionado pois apresenta baixa correlação com outras características como rendimento da polpação (MOKFIENSKI *et al.*, 2008; GOMIDE; NETO; REGAZZI, 2010; NETO, 2012). No processo industrial, busca-se trabalhar com densidades uniformes uma vez que a velocidade de impregnação é diretamente afetada por ela. Segundo Lanna *et al.* (2001), madeiras de baixa densidade (442 kg/m³) favoreceram o rendimento de celulose quando comparadas com madeiras consideradas de alta densidade (520 kg/m³), para um mesmo número KAPPA. Densidades mais elevadas também favorecem a produtividade dos digestores, pois maior volume do digestor estará preenchido com madeira. Por sua vez, densidades mais baixas favorecem a impregnação, demandando menores cargas alcalinas, com reflexos na qualidade e rendimento da polpa, bem como menores cargas de sólidos para recuperação. Por outro lado, o aumento do consumo específico (m³ de madeira/t de celulose) torna-se uma grande desvantagem, uma vez que o volume do digestor estará preenchido com menos massa (NETO, 2012; VENTORIM *et al.*, 2009; MOKFIENSKI *et al.*, 2008).

2.1.2 Características Químicas da Madeira

- **Carboidratos**: Segundo Morais (2008), a análise generalizada dos carboidratos tratando-os somente como celulose e hemicelulose é mais útil que separadamente (arabinose, galactose, glicose, xilose e manose) uma vez que ao final do processo de polpação e branqueamento o produto será basicamente composto por celulose e hemicelulose, sendo estes bons parâmetros para conhecimento da matéria-prima (COLODETTE; GOMES, 2015).

- **Hemiceluloses**: As hemiceluloses desempenham um importante papel no rendimento do processo de produção, na qualidade da polpa e em propriedades do papel. Em geral, madeiras com baixo teor de lignina e um alto teor de carboidratos exigirão condições menos severas de polpação e conduzirão a um alto rendimento gravimétrico (SANTOS, 2005; COLODETTE; GOMES, 2015).

- **Extrativos**: Os extrativos podem causar incrustações ao longo do processo de produção de celulose, reduzir a qualidade da polpa e do papel por meio da formação de pitch, além de consumirem reagentes de cozimento (ALVES *et al.*, 2011; COLODETTE; GOMES, 2015).

- **Lignina**: Para o processo de polpação, os materiais mais indicados são aqueles com menor teor de lignina total e proporcionalmente maior relação de lignina solúvel frente a insolúvel. Estudos indicam que quanto maior o teor de lignina solúvel, maior a relação siringila/guaiacila (S/G) da lignina e, por sua vez, maior a velocidade de reação, uma vez que a estrutura siringila da lignina é mais reativa, tornando o processo mais seletivo, reduzindo a demanda de álcali e melhorando o rendimento (ALVES *et al.*, 2011; GOMIDE *et al.*, 2005; COLODETTE; GOMES, 2015).

- **Relação S/G**: Uma característica importante da madeira está relacionada à estrutura da lignina presente na madeira, dentre elas destacam-se as estruturas siringil e guaiacil, especialmente em madeira de folhosa. As estruturas do tipo siringil são mais reativas devido à sua característica menos condensada, proporcionando melhor deslignificação no cozimento KRAFT. Desta forma, quanto maior a relação S/G, melhor a deslignificação e menor carga de álcali (GOMIDE *et al.*, 2005; COLODETTE; GOMES, 2015).

- **Cinza e metais**: A madeira contém uma variedade de minerais que são absorvidos pela árvore durante seu crescimento. Esses minerais, como cálcio, potássio, magnésio e outros, podem afetar várias propriedades da madeira e definir seu desenvolvimento. No processo de produção de celulose, os minerais, muitas vezes referidos como cinzas, podem influenciar a eficiência dos reagentes químicos usados no processamento da madeira e na qualidade da polpa produzida. Os metais atuam como catalisadores nas reações do peróxido de hidrogênio, resultando em uma menor eficiência dos reagentes e exigindo maiores cargas dos mesmos. Além disso, a presença de metais na polpa pode levar ao aumento dos valores de reversão de alvura. Assim, altos teores de cinzas e metais na madeira não são desejáveis no processo industrial, embora sejam essenciais ao desenvolvimento da planta (FIGUEIREDO, 2019; COLODETTE; GOMES, 2015).

2.1.3 Características de Polpação KRAFT

O objetivo do cozimento é solubilizar a lignina e preservar os carboidratos obtendo-se uma massa de fibras individualizadas, no entanto, o licor de cozimento responsável por este processo também degrada os carboidratos afetando negativamente o rendimento. Sendo a madeira o principal componente de custo do processo, a otimização do rendimento é um fator atraente do ponto de vista econômico (GOMIDE; NETO; REGAZZI, 2010; COLODETTE; GOMES, 2015).

- **Número KAPPA**: O número KAPPA é um indicador do teor de lignina residual ou da capacidade de branqueamento da polpa e é medido como número de mililitros da solução de permanganato de potássio 0,1N, consumido sob condições especificadas, por um grama de massa celulósica. É importante ressaltar que esta medida pode sofrer interferência da presença dos ácidos hexenurônicos, pequenas frações de carbonila além da lignina residual. Colodette, Mounteer e Gomes (2012) observa que uma faixa de KAPPA ideal está entre 16 e 18, sendo o KAPPA objetivo ideal altamente dependente do tipo de madeira e do processo de produção. Analisando isoladamente o número KAPPA, o ideal seria selecionar os materiais que apresentem o menor número KAPPA com a maior redução deste por unidade de AE (COLODETTE; GOMES, 2015).

- **Rejeito**: O teor de rejeito está diretamente associado à eficiência do cozimento. Altas concentrações de álcali resultam em rendimentos menores e menor teor de rejeito; por outro lado, baixas concentrações resultam em aumento do teor de rejeito (ALMEIDA, 2003; COLODETTE; GOMES, 2015).

- **Rendimento Depurado (RD)**: Por meio do RD se mede a eficiência do processo de cozimento, permitindo avaliar quanto de celulose (massa de fibras individualizadas) foi possível extrair da madeira alimentada no digestor. O processo KRAFT é considerado um processo de baixo rendimento, apresentando valores em torno de 50 por cento. Em geral, processos agressivos de cozimento (carga alcalina e temperaturas elevadas) ou uma má impregnação contribuem para redução do rendimento. O rendimento também é influenciado pelo teor de carboidratos, lignina e extrativos da madeira (GOMIDE *et al.*, 2005; NETO, 2012; COLODETTE; GOMES, 2015).

- **Álcali Efetivo Residual (AER)**: O álcali efetivo residual (AER) é um bom indicador da eficiência do cozimento, embora não exista uma relação linear; até certo ponto, quanto maior o AE, maior o AER. A falta de residual ao final do cozimento indica que o material pode não ter sido devidamente deslignificado, além de possibilitar redeposição de lignina; por outro lado, o excesso de AER é indicativo de baixos rendimentos devido a degradações dos carboidratos por meio de condições mais severas das reações. Do ponto de vista da maximização do rendimento (por meio da melhor conservação dos carboidratos), as polpações devem buscar residual baixo e bem controlado de álcali. Nos processos industriais, estes limites normalmente são manipulados para ajustar a impregnação, devido aos tempos de retenção inferiores aos projetados para os equipamentos. O AER também contribui para proteção anódica da estrutura dos digestores, o que em alguns casos pode gerar um excesso de carga alcalina, que não está especificamente relacionado ao cozimento da madeira (ABRANCHES, 2017; COLODETTE; GOMES, 2015).

- **pH do Licor Negro**: O pH do licor negro pode ser um bom indicativo do consumo de álcali efetivo durante o processo de cozimento. Para uma mesma carga de AE, quanto maior o pH do licor, menor o consumo de álcali no processo de cozimento (FIGUEIREDO, 2019; COLODETTE; GOMES, 2015).

- **Ácidos Hexenurônicos (HexA)**: Os ácidos hexenurônicos são formados durante a polpação alcalina, a partir dos ácidos 4-O-metilglicurônicos presentes nas xilanas, e são influenciados por condições como álcali ativo, sulfidez e temperatura. Embora presentes na polpa, são indesejáveis no branqueamento, pois consomem reagentes químicos e aumentam a capacidade de quelar metais, levando a uma maior reversão de alvura. Além disso, é importante ressaltar que o HexA é um componente da polpa KRAFT que se liga à lignina residual, consumindo permanganato de potássio durante a titulação do KAPPA. Isso pode levar a uma superestimação do número KAPPA, indicando um teor de lignina mais alto do que o real (VENTORIM *et al.*, 2009; GOMIDE *et al.*, 2005; COLODETTE; GOMES, 2015).

- **Viscosidade**: A viscosidade é uma forma indireta de medir a qualidade da polpa, sobretudo considerando o aspecto relacionado à preservação das cadeias de polissacarídeos, o que afeta as propriedades físicas e mecânicas das polpas. Dessa forma, é um indicativo de seletividade do processo (FIGUEIREDO, 2019; COLODETTE; GOMES, 2015).

- **Teor de Sólidos do Licor Preto**: O conceito de sólidos refere-se à quantidade de matéria orgânica e inorgânica presente no licor após o processo de polpação. Trata-se de um subproduto do processo de cozimento na produção de celulose. O teor de sólidos é crucial, pois afeta a energia gerada durante a queima desse material na caldeira de recuperação. Uma variação no teor de sólidos pode impactar toda a matriz energética da fábrica. Na polpação, teores mais altos podem significar um processo menos seletivo; já na recuperação, pode gerar problemas para a caldeira, especialmente se a capacidade de queima representar um gargalo para o aumento de produção (ALMEIDA, 2010).

2.1.4 Características do Branqueamento KRAFT

O branqueamento da polpa KRAFT visa remover a lignina residual após o cozimento. A dosagem de reagentes como oxigênio, dióxido de cloro ClO_2 e peróxido de hidrogênio H_2O_2 deve ser controlada, assim como a pressão sob a qual estes são aplicados. A consistência da polpa e a carga de álcali influenciam diretamente a eficácia do branqueamento e a estabilidade da alvura. O pH do meio regula a estabilidade química e eficiência das reações e o pré-tratamento com oxigênio (pré-O_2) reduz a carga química requerida na recuperação, alinhando economia operacional com menor impacto ambiental (COLODETTE; GOMES, 2015; GOMIDE; NETO; REGAZZI, 2010).

- **Consumo Total de ClO_2**: A quantificação do dióxido de cloro total é crucial para avaliar a eficiência do branqueamento. Este indicador reflete quão efetivamente o ClO_2 está sendo utilizado para remover a lignina, otimizando o uso de reagentes químicos e minimizando emissões e resíduos.

- **Consumo de O_2**: O uso de oxigênio no estágio inicial de branqueamento indica a eficiência da deslignificação. Monitorar o O_2 consumido por tonelada ajuda a otimizar a quantidade de cloro necessário nos estágios subsequentes.

- **Alvura**: A alvura alcançada no estágio EP, medida em porcentagem ISO, é um indicador crítico da qualidade do branqueamento. Esta variável não apenas reflete a eficácia dos agentes branqueadores utilizados, mas também ajuda a determinar a adequação da polpa para aplicações específicas onde a alvura é prioritária.

2.2 Aprendizado de Máquina

O aprendizado de máquina é uma área da inteligência artificial que estuda formas de se identificar padrões em dados e surgiu da necessidade de se processar grandes volumes de informações sem a necessidade de uma programação explícita e direta para obtenção dos resultados. Os algoritmos de aprendizado de máquina se popularizaram sobretudo após resultados satisfatórios em áreas como diagnósticos médicos, detecções de fraudes, reconhecimento de imagens, controles de veículos, entre outros. O processo para extração de conhecimento através do aprendizado de máquina passa pela coleta de dados, tratamento dos dados, seleção de modelos, treinamentos, testes, ajustes de parâmetros e aplicação propriamente dita. Podemos dividir a área de aprendizado de máquina em dois ramos: preditivo, que envolve o aprendizado supervisionado, e o descritivo, que aborda o aprendizado não supervisionado (LUDERMIR, 2021).

2.2.1 Aprendizado Supervisionado

No aprendizado supervisionado, os algoritmos são desenvolvidos para identificar padrões em conjuntos de dados rotulados (modelo) e assim predizer novos valores. A tarefa de predição pode estar associada ao objetivo de executar uma regressão ou uma classificação (LUDERMIR, 2021).

2.2.2 Aprendizado Não Supervisionado

Nas tarefas de aprendizado não supervisionado, quando o objetivo estiver associado à extração de padrões do conjunto de dados não rotulados, ou seja, que não possuem atributos de referência (alvo) para guiar os algoritmos, estes por sua vez passam a buscar características relevantes nos dados que permitam gerar agrupamentos distintos (LUDERMIR, 2021).

2.2.2.1 Tarefa de Agrupamento

A tarefa de realizar agrupamentos também é conhecida como clusterização. Esta técnica de aprendizado não supervisionado consiste em gerar grupos considerando as dimensões do conjunto de dados de forma que os elementos de um *cluster* compartilhem propriedades comuns que os diferenciem dos elementos dos outros *clusters*. Ou seja, este tipo de abordagem busca encontrar centros de grupos próximos e determinar os melhores agrupamentos considerando apenas as informações que os descrevem e seus relacionamentos. Dessa forma, os elementos de um cluster devem apresentar similaridades intra-grupos e dissimilaridades inter-grupos (MOORI; MARCONDES; ÁVILA, 2002; MONTEIRO, 2019).

Médias de Similaridade: Para a determinação dos grupos, é necessário estabelecer medidas de similaridade ou dissimilaridade entre seus elementos. Estas medidas serão aplicadas para a formação dos *clusters*. Para cada tipo de dado (binário, nominal, ordinal ou contínuo), deve-se considerar as medidas apropriadas. No caso de dados contínuos, a proximidade dos elementos é tipicamente medida por métricas de distância como: distância Euclidiana, Manhattan, Minkowski, Camberra ou baseada em correlações e cosseno (CASTRO; FERRARI, 2017).

2.2.2.2 Métodos de Agrupamento Não Supervisionado

- **KMeans**: é um método de agrupamento particional que divide um conjunto de dados em k grupos com base na similaridade entre os seus elementos. Ele é um dos métodos mais populares para análise de dados e aprendizado de máquina, devido à sua simplicidade e eficiência. O algoritmo recebe como entrada o número de grupos desejados (k) e seleciona aleatoriamente k elementos do grupo para serem os centroides iniciais. Cada elemento é atribuído ao grupo do centroide mais próximo, de acordo com uma distância definida (tipicamente Euclidiana). A média dos objetos em cada grupo é calculada e atualiza os valores dos centroides em cada iteração até que não haja mais mudanças no centroide ou até atingir o número máximo de iterações (SINAGA; YANG, 2020; TIRELLI *et al.*, 2022).

- **Agrupamento Hierárquico**: é um algoritmo de agrupamento não supervisionado que cria uma estrutura de árvore hierárquica, também conhecida como dendrograma, para organizar os elementos do conjunto de dados em grupos. Este método pode ser classificado em aglomerativo, quando cada elemento começa como um cluster individual e, em seguida, combina os clusters mais próximos de acordo com uma medida de similaridade até que todos os objetos estejam em um único cluster, ou divisivo, quando todos os elementos começam em um único cluster e, em seguida, divide os clusters em clusters menores de forma recursiva até que cada elemento esteja em um cluster individual. O método hierárquico aglomerativo é o mais utilizado. Ele calcula a matriz de proximidade que contém a distância entre cada par de elementos no conjunto de dados e, em seguida, encontra os clusters mais próximos. A cada iteração, o par de clusters com a menor distância entre si é encontrado. Em seguida, combina os clusters mais próximos em um único cluster e atualiza a matriz de proximidade para refletir a nova estrutura de clusters. Isso se repete até que todos os elementos estejam em um único cluster (MONTEIRO, 2019; CASTRO; FERRARI, 2017).

- **HDBSCAN**: É um avançado método de agrupamento baseado em densidade, ideal para superar as restrições do número pré-definido de clusters, característica de muitos algoritmos de agrupamento. O HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) é uma extensão do DBSCAN que considera a densidade variável dos clusters. Este algoritmo identifica regiões de alta densidade que são separadas por regiões de baixa densidade e não exige a especificação de um raio (ε) e o número mínimo de pontos (MinPts) para a formação de um cluster. Em vez disso, o HDBSCAN opera com um parâmetro de *min_cluster_size*, que é intuitivamente mais fácil de configurar e fornece maior flexibilidade na identificação de clusters de diferentes densidades. O algoritmo começa com uma estimativa de densidade de cada ponto e constrói uma hierarquia de clusters, que é então condensada para formar os clusters finais, filtrando-se os pontos que não se enquadram em nenhuma categoria significativa como 'ruído'. Esta abordagem torna o HDBSCAN particularmente útil para dados com clusters de variadas densidades e formas, proporcionando uma segmentação mais precisa e adaptativa em comparação ao DBSCAN (RONCORONI *et al.*, 2023; STEWART; AL-KHASSAWENEH, 2022).

Na Tabela 1 é apresentado um comparativo entre algoritmos de agrupamento KMeans, Aglomerado Clustering e HDBSCAN.

Tabela 1 – Comparação entre Algoritmos de Agrupamento

Característica	**KMeans**	**Agglomerative Clustering**	**HDBSCAN**
Tipo de Algoritmo	Particional	Hierárquico	Hierárquico Baseado em densidade
Necessidade de Especificar Nº de Clusters	Sim (k)	Sim (indiretamente através do critério de ligação e threshold)	Não (parâmetros de densidade e mínimo de amostras)
Escalabilidade	Alta para média	Baixa para média	Média a alta
Sensibilidade a Outliers	Alta	Média	Baixa
Complexidade de Tempo	O(n*k*i) (n = número de pontos, k = clusters, i = iterações)	$O(n^2)$ a $O(n^3)$, dependendo da implementação	O(n*log(n)) a $O(n^2)$
Adequação	Melhor para clusters globulares e bem separados	Bom para dados que naturalmente formam hierarquias	Excelente para dados com variações de densidade
Flexibilidade de Formas de Clusters	Limitado a esferas	Pode formar clusters de várias formas	Pode formar clusters de várias formas
Implementações Comuns	Scikit-learn, MATLAB, R	Scikit-learn, MATLAB, R	Python (HDBSCAN library)

Fonte: Adaptado de Moori, Marcondes e Ávila (2002).

2.2.2.3 Avaliação dos Agrupamentos

A avaliação dos agrupamentos é um passo crucial na análise de dados não supervisionada. Ela permite avaliar a qualidade dos clusters formados por um algoritmo, determinar o número ideal de clusters, permitindo comparações mais precisas entre os agrupamentos formados e os dados originais. O coeficiente de Silhueta é um indicador que permite avaliar a qualidade dos clusters com base na proximidade dos elementos dentro e entre os grupos. Trata-se de uma medida que varia entre -1 e 1, onde elementos próximos de 1 indicam boa alocação, e -1 indicam alocação inadequada. Nos gráficos de Silhueta é possível identificar as distâncias de cada elemento em cada grupo, bem como o valor médio do coeficiente para a quantidade de grupos formada. O Método Elbow normalmente é utilizado para identificar a quantidade ideal de clusters em agrupamentos baseados em distâncias. Ele determina o número ideal de clusters com base na soma dos erros quadráticos intragrupos. Na análise do gráfico de "cotovelo"é possível identificar o ponto de inflexão que compreende a soma dos erros quadráticos versus a quantidade de clusters avaliada, indicando a quantidade ideal de grupos para aquele conjunto de dados (MOORI; MARCONDES; ÁVILA, 2002).

Na Tabela 2 é apresentado um resumo de algumas métricas de validação de agrupamentos incluindo Índice Dun e Índice de Davies-Bouldin.

Tabela 2 – Métodos de Validação de Agrupamento

Método	**Definição**	**Vantagem**	**Desvantagem**	**Métrica**
Coeficiente de Silhueta	Mede a qualidade da alocação de elementos em clusters.	Simples de entender e interpretar. Robusto a outliers. Pode ser aplicado a qualquer tipo de dados.	Pode ser sensível à escolha da métrica de dissimilaridade. Não leva em consideração o tamanho dos clusters.	Dissimilaridade entre elementos e centroides
Método Elbow	Estima o número ideal de clusters com base na soma dos erros quadráticos intragrupos (SSE).	Simples de implementar. Visualmente intuitivo.	Pode ser sensível à escolha da métrica de dissimilaridade. Pode não ser adequado para datasets com clusters de tamanhos diferentes.	Soma dos erros quadráticos intragrupos (SSE)
Índice Dunn	Mede a coesão intracluster e a separação intercluster.	Compensa a sensibilidade a outliers.	Pode ser computacionalmente caro.	Razão entre a menor distância intercluster e a maior distância intracluster
Índice de Davies-Bouldin	Mede a dispersão dentro dos clusters e a similaridade entre clusters.	Menos sensível a outliers do que o índice Dunn.	Pode ser difícil de interpretar.	Razão entre a soma das dispersões intracluster e a menor distância intercluster

Fonte: Adaptado de Moori, Marcondes e Ávila (2002).

2.3 Normalização e Padronização dos Dados

Normalização e padronização são técnicas importantes aplicadas no pré-processamento dos dados em análise estatística e aprendizado de máquina. A normalização, frequentemente implementada através do escalonamento Min-Max, ajusta os dados para um intervalo comum entre 0 e 1, facilitando algoritmos que dependem da magnitude dos dados. Já a padronização, realizada pelo cálculo Z-Score, transforma os dados para terem média zero e desvio padrão unitário, adequada para métodos que pressupõem uma distribuição normal. Muitos algoritmos de aprendizado de máquina requerem normalização ou padronização dos dados e estes podem influenciar significativamente a performance dos modelos (ALBON, 2018).

2.4 Imputação de Dados

Tratar valores ausentes é crucial em aprendizado de máquina, pois esses valores podem afetar significativamente a performance dos modelos. Uma abordagem comum é a substituição pela média, que deve ser usada com cautela devido a possíveis outliers ou assimetria dos dados. Alternativamente, a mediana ou métodos mais sofisticados, como o KNN-Imputer, podem ser mais apropriados. O KNN-Imputer utiliza os k-vizinhos mais próximos para estimar os valores ausentes, baseando-se na premissa de que padrões semelhantes possuem valores semelhantes (ALBON, 2018).

2.5 Redução de Dimensionalidade (RD)

A Redução de Dimensionalidade é uma técnica poderosa para analisar conjuntos de dados multidimensionais. Ela é capaz de transformar um conjunto de dados com muitas características em um conjunto com características reduzidas chamadas de dimensões, preservando informações importantes referentes às variáveis originais. A RD contribui significativamente para reduzir a demanda computacional por meio da melhoria da eficiência do processamento e armazenamento de dados, especialmente para conjuntos de dados maiores. Além disso, contribui para melhor visualização dos dados e para redução de problemas relacionados a multicolinearidade e overfitting em modelos (LOVRIć *et al.*, 2021).

As técnicas de RD podem ser divididas em lineares e não lineares. Entre as técnicas lineares, temos a Análise de Componentes Principais (PCA), uma técnica clássica que identifica as direções de maior variância nos dados. Ela transforma o conjunto de dados original em um novo conjunto de variáveis, chamadas componentes principais, que são descorrelacionadas e ordenadas de forma que as primeiras componentes retenham a maior parte da variância presente no conjunto de dados original. A Análise Discriminante Linear (LDA) é uma técnica supervisionada que maximiza a separabilidade entre as classes conhecidas e, por isso, é normalmente utilizada em problemas de classificação. Já entre as técnicas não lineares, temos o t-SNE, uma técnica que preserva a estrutura local dos dados, sendo eficaz na visualização de dados de alta dimensionalidade em espaços de duas ou três dimensões, e o UMAP, uma técnica recente que preserva tanto a estrutura local quanto global dos dados, sendo mais eficiente para trabalhar em conjuntos de dados maiores e mais complexos (JOSWIAK *et al.*, 2019; FERNANDES; FILHO, 2019).

2.6 Caracterização dos Agrupamentos

O Índice de Gini e o Intervalo de Confiança são ferramentas estatísticas valiosas na caracterização de agrupamentos não supervisionados. O Índice de Gini quantifica a desigualdade ou a pureza dentro dos grupos formados, sendo próximo de 1 para variáveis que efetivamente diferenciam os grupos e próximo de 0 para aquelas com pouca diferenciação. Já o Intervalo de Confiança é uma faixa de valores que, com um certo nível de confiança, contém o verdadeiro valor de um parâmetro populacional. Em outras palavras, se você calcular o IC várias vezes, espera-se que em 95 por cento das vezes, para um IC de 95 por cento, o intervalo contenha o valor real do parâmetro. Da mesma maneira, o nível de confiança é a probabilidade de que o IC contenha o verdadeiro valor do parâmetro. O nível de 95 por cento é o mais comum, mas pode ser ajustado de acordo com a necessidade do estudo (FREIRE, 2021). Um IC estreito indica uma estimativa mais precisa, enquanto um IC largo indica uma estimativa menos precisa. É importante lembrar que o IC é uma estimativa probabilística e que pressupõe normalidade dos dados. Isso significa que não há garantia de que o verdadeiro valor do parâmetro esteja dentro do intervalo (PATINO; FERREIRA, 2015).

2.7 Trabalhos Relacionados

Nesta seção são apresentados alguns estudos relacionados à redução de dimensionalidade e agrupamento não supervisionados que serviram de referência para o desenvolvimento deste trabalho.

Anjos *et al.* (2015) apresenta um estudo que busca construir modelos matemáticos para prever a relação quantitativa entre a densidade e várias propriedades mecânicas e ópticas do papel. Foi empregada a Análise de Componentes Principais (PCA) combinada com técnicas de regressão multivariável para estabelecer modelos de previsão separados para cada espécie, possibilitando verificar a relação entre a densidade do papel e suas diversas propriedades, como resistência à tração, índice de rasgamento e opacidade.

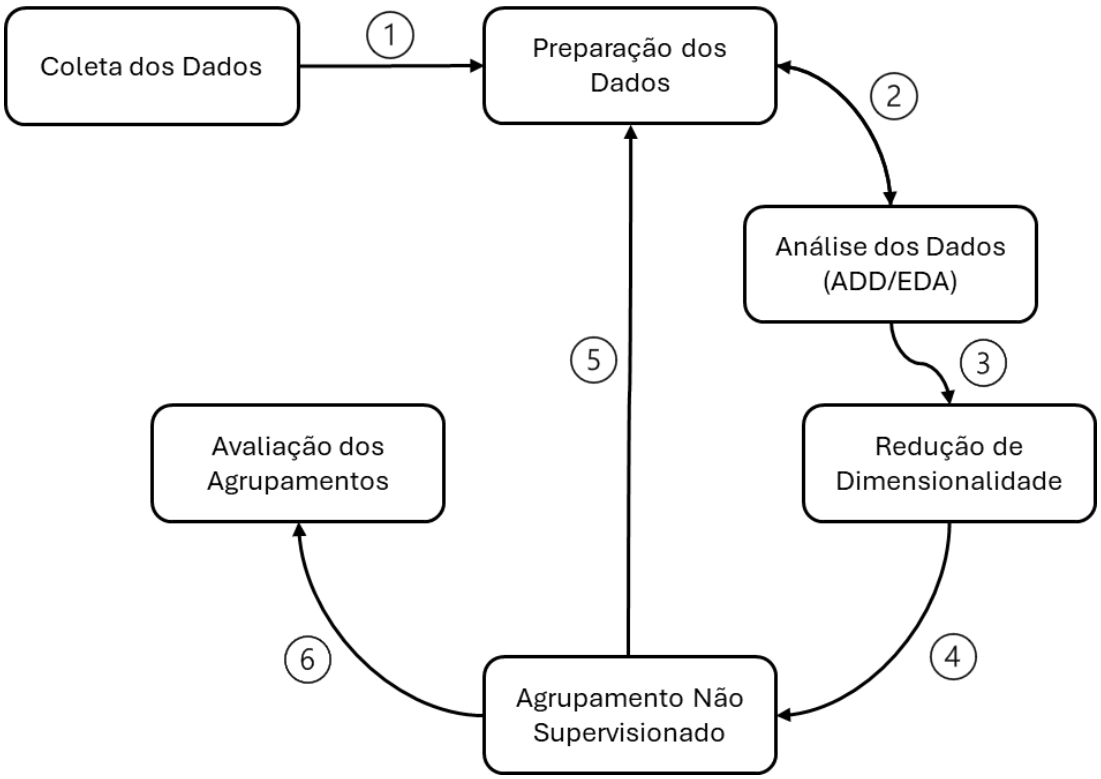
Trugilho *et al.* (2004) utilizaram o método de otimização de Tocher para resolver o problema de classificação de clones de Eucalyptus sp. com base em três características da polpa KRAFT: rendimento depurado, viscosidade da polpa e álcali ativo. O método de Tocher, aplicado à análise de agrupamento, divide os clones em grupos homogêneos com base em suas características. Este estudo demonstrou a eficiência do método de Tocher na classificação dos clones, formando cinco grupos distintos com alto, médio e baixo potencial para produção de polpa celulósica. As características que mais contribuíram para a divergência entre os clones foram o rendimento depurado (58,1 por cento) e a viscosidade da polpa celulósica (35,0 por cento). O método de Tocher se mostrou uma ferramenta valiosa para a seleção de clones com alto potencial para produção de celulose, facilitando o trabalho de melhoramento genético e otimizando a produção de polpa.

Da mesma maneira, Caixeta *et al.* (2003) exploraram formas de classificar e selecionar genótipos superiores de eucalipto por meio das propriedades da madeira. Os autores utilizaram a redução de dimensionalidade aplicando diagnóstico de multicolinearidade na matriz de correlação fenotípica, identificando e descartando variáveis que provocavam forte multicolinearidade. Essa etapa foi baseada nos fatores de inflação da variância, magnitude dos autovalores e valor singular. Também foi utilizada a análise de agrupamento, empregando distâncias euclidianas padronizadas como medida de dissimilaridade e o método de otimização de Tocher, resultando na formação de 11 grupos distintos entre os 44 genótipos de eucalipto estudados.

3 Metodologia

Este capítulo aborda a metodologia utilizada para investigar o agrupamento de clones de eucalipto com base nas características da madeira e de seus respectivos processamentos, conforme ilustrado na Figura 1. A pesquisa utilizou algoritmos de aprendizado de máquina não supervisionados para analisar variáveis químicas, físicas e do processo KRAFT, em especial da polpação, do licor de cozimento e do respectivo branqueamento, gerados a partir de análises laboratoriais. A metodologia abrange a preparação dos dados, análise descritiva e exploratória de dados (ADD/EDA), técnicas de redução de dimensionalidade e a aplicação de algoritmos de agrupamento.

Figura 1 – Esquema metodológico baseado no processo CRISP-DM[1]



Fonte: Adaptado de Plumed *et al.* (2019)

[1] CRISP-DM é um método abrangente que fornece diretrizes claras e detalhadas sobre o ciclo de vida de projetos de mineração de dados, incluindo fases de entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação.

3.1 Coleta e Preparação dos Dados

Os dados analisados neste estudo foram extraídos do trabalho de caracterização completa de clones de eucalipto, utilizados no processo de produção de celulose KRAFT de uma fábrica do leste de Minas Gerais. Desta base de caracterização completa, foram selecionados 58 tratamentos (clones de duas regiões em idade de corte) contendo 4 amostras cada. O subconjunto de dados resultou em 229 registros e 48 atributos, sendo 26 variáveis físico-químicas da madeira, 9 de polpação e 13 relativas ao processo de branqueamento conforme descrito na Tabela 3. Ao todo, foram avaliados 58 materiais genéticos diferentes (clones), provenientes de 4 espécies, sendo uma pura e 3 híbridos. Estes dados foram recuperados do sistema LIMS (Laboratory Information Management System), responsável por gerenciar informações laboratoriais e posteriormente processados na linguagem Python versão 3.10.12 no ambiente de desenvolvimento Google Colaboratory Pro.

Tabela 3 – Características analisadas

#	Característica	Descrição	Unidade	Grupo Característica
1	F_DB	DB_DISCO (Cor 5discos)	kg/m³	Física
2	Q_CEL	Celulose	%	Química
3	Q_HEM	Hemicelulose	%	Química
4	Q_HOLO	Celulose+Hemicelulose	%	Química
5	Q_GLIC	Glicose	%	Química
6	Q_ARA	Arabinose	%	Química
7	Q_GAL	Galactose	%	Química
8	Q_XIL	Xilose	%	Química
9	Q_MAN	Manose	%	Química
10	Q_LIG_INS	Lignina insolúvel	%	Química
11	Q_LIG_SOL	Lignina solúvel	%	Química
12	Q_LIG_TOT	Lignina total	%	Química
13	Q_EXT	Extrativos	%	Química
14	Q_REL_SG	Relação S/G	#	Química
15	Q_N	N	%	Química
16	Q_P	P	%	Química
17	Q_K	K	%	Química
18	Q_NA	Na	%	Química
19	Q_CA	Ca	%	Química
20	Q_MG	Mg	%	Química
21	Q_CU	Cu	mg/Kg	Química
22	Q_ZN	Zn	mg/Kg	Química
23	Q_FE	Fe	mg/Kg	Química

Continua na próxima página

Tabela 3 – continuação da página anterior

#	Característica	Descrição	Unidade	Grupo Característica
24	Q_MN	Mn	mg/Kg	Química
25	Q_B	B	mg/Kg	Química
26	Q_S	S	%	Química
27	P_KAPPA	#Kappa	#	Polpação
28	P_RB	RD Bruto	%	Polpação
29	P_REJ	Rejeito	%	Polpação
30	P_RD	Rendimento Depurado	%	Polpação
31	L_ARES	AE residual (NaOH)	g/L	Polpação
32	L_SOLID	Licor preto Sólidos	%	Polpação
33	L_CALCI	Perda por calcinação	%	Polpação
34	L_PH	pH	(vazio)	Polpação
35	L_PCALO	Poder calorífico	(cal.g-1)	Polpação
36	B_CLO2TOT	Consumo Total ClO2	Kg/Tsa	Branqueamento
37	B_CLO2_KAPPA	ClO2/#Kappa	(vazio)	Branqueamento
38	B_OO_HEXA	Estágio_OO_HexA	mmol/kg	Branqueamento
39	B_OO_VISCO	Estágio_OO_Viscosidade	mPa·s	Branqueamento
40	B_OO_KP_ENT	Estágio_OO_Kappa entrada	(vazio)	Branqueamento
41	B_OO_KP_SAI	Estágio_OO_#kappa saída	(vazio)	Branqueamento
42	B_OO_ALVU	Estágio_OO_Alvura	%ISO	Branqueamento
43	B_OO_O2	Estágio_OO_Consumo O2	Kg/Tsa	Branqueamento
44	B_D0_CLO2	Estágio_D0_Consumo ClO2	Kg/Tsa	Branqueamento
45	B_D1_CLO2	Estágio_D1_Consumo ClO2	Kg/Tsa	Branqueamento
46	B_EP_mKAPPA	Estágio_EP_µKappa	(vazio)	Branqueamento
47	B_EP_ALVU	Estágio_EP_Alvura A.D.	%ISO	Branqueamento
48	B_P_VISC	Estágio_P_Viscosidade	mPa·s	Branqueamento

Fonte: Autor (2024).

Para manipulação, análise e visualização dos dados, utilizou-se a importação de bibliotecas essenciais como Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, entre outras[2]. A preparação dos dados envolveu a remoção de variáveis consideradas desnecessárias para a análise, tratamento de valores ausentes e testes com variáveis categóricas utilizando técnicas de One-Hot Encoding. Este processo assegurou que o conjunto de dados estivesse limpo e preparado para análises subsequentes.

[2] (MCKINNEY, 2012)

3.2 Análise Exploratória de Dados (EDA)

Durante a EDA, técnicas visuais e estatísticas foram aplicadas para examinar as propriedades dos dados. Utilizou-se a biblioteca Sweetviz para geração de relatórios analíticos detalhados, facilitando a visualização de distribuições e correlações. Box plots e histogramas foram gerados para cada variável, permitindo avaliar a normalidade, a distribuição dos dados e a identificação de variáveis com presença de outliers.

3.3 Redução de Dimensionalidade

Para enfrentar desafios associados à alta dimensionalidade do conjunto de dados, procedeu-se à normalização (Z-score) dos dados e aplicaram-se as técnicas de PCA, UMAP e t-SNE. No contexto do estudo, o PCA foi utilizado para observar se a estrutura linear dos dados seria suficiente para captar as variações significativas entre os clones de eucalipto e facilitar a identificação de agrupamentos potenciais. O uso do UMAP teve como objetivo explorar se a abordagem não linear revelaria agrupamentos mais claros ou estruturas de dados que não eram aparentes com o PCA, proporcionando uma representação de dados mais adequada para a análise de agrupamento subsequente. Já a aplicação do t-SNE visou investigar se esta técnica proporcionaria visualizações mais intuitivas e discriminativas dos agrupamentos naturais dos dados, comparativamente ao PCA e ao UMAP, especialmente para a visualização e interpretação dos dados em duas ou três dimensões.

3.4 Algoritmos de Agrupamento

Diversos algoritmos de agrupamento foram aplicados ao conjunto de dados transformado:

- **KMeans**: Utilizou-se o método do cotovelo e a análise de silhueta para determinar o número ótimo de clusters.
- **AgglomerativeClustering**: Empregou-se o linkage de Ward para a construção de dendrogramas, auxiliando na visualização e decisão do número de clusters.
- **HDBSCAN**: Este algoritmo baseado em densidade foi aplicado para identificar clusters de forma automática, adaptando-se melhor a variações de densidade quando comparado ao DBSCAN, sem a necessidade de especificar o número de clusters.

Cada algoritmo foi avaliado com base no coeficiente de silhueta, índice de Davies-Bouldin e Calinski-Harabasz. Também foram utilizadas visualizações das projeções dos dados nos espaços de dimensão reduzida. Ao final, uma tabela foi gerada para facilitar as comparações e identificar a configuração do método de melhor desempenho.

3.5 Avaliação dos Clusters e Interpretação dos Resultados

Após a execução dos algoritmos de agrupamento, procedeu-se à avaliação dos grupos formados. Esta análise contou com a opinião de especialistas no que se refere à alocação dos clones em cada cluster, bem como a análise dos perfis gerados e dos intervalos de confiança. Além destas análises, procedeu-se à avaliação da variável de pontuação "score", calculada com o objetivo de facilitar a análise quantitativa e qualitativa de cada agrupamento, de acordo com os pesos atribuídos para cada variável de interesse no processo. Por fim, foram listadas as cinco características que mais contribuíram para a heterogeneidade do agrupamento para cada grupo de características (todas as características, características físico-químicas, minerais, polpação e branqueamento), utilizando o índice de Gini calculado sobre os valores médios dos clusters para cada variável.

4 Resultados

A Tabela 4 exibe a análise descritiva dos dados processados neste trabalho. Essa análise oferece uma visão geral das características dos dados, incluindo medidas de posição, tendência central e dispersão, o que permite orientar o processamento subsequente de tratamento e agrupamento dos dados. A média é uma medida de posição que indica a concentração dos pontos importantes para determinar a posição dos agrupamentos. O desvio padrão e o coeficiente de variação são indicadores da variabilidade dos dados. Valores altos nesses indicadores sugerem uma grande dispersão, que pode influenciar na formação de grupos homogêneos e sinalizar a necessidade de normalizar os dados. A mediana e os quartis são úteis para compreender a distribuição dos dados, atenuando o efeito de valores atípicos (outliers) que podem alterar a média. Isso proporciona um entendimento mais robusto da distribuição dos dados, que é resistente a valores extremos, essencial para identificar e corrigir anomalias antes do agrupamento. Além disso, a proporção de dados ausentes é um aspecto crítico para julgar a completude dos dados. Uma alta proporção de dados ausentes pode revelar falhas na coleta de dados ou características menos significativas ou sensíveis para a pesquisa. Neste trabalho, optou-se por proceder à imputação dos dados ausentes.

Tabela 4 – Resumo Estatístico dos Dados

#	Variável	Qtd	Méd	DP	Mín	25%	50%	75%	Máx	CV%	Nulo%
1	IDADE	229	6,59	0,6	5,6	6	6,6	7,1	7,9	9,1	0
2	F_DB	215	500,54	31,48	425,89	479,7	499,81	518,89	581,87	6,3	6,1
3	Q_CEL	225	47,3	2,04	40,9	46,2	47,25	48,7	52,48	4,3	1,7
4	Q_HEM	225	13,11	1,11	9,14	12,4	13,03	13,8	16,9	8,5	1,7
5	Q_HOLO	225	60,42	2,18	53,7	59,1	60,6	62,1	67,94	3,6	1,7
6	Q_ARA	225	0,22	0,06	0,1	0,2	0,2	0,24	0,4	25,2	1,7
7	Q_GAL	225	1,07	0,31	0,26	0,9	1	1,2	2,12	28,6	1,7
8	Q_XIL	225	10,87	1,11	7,73	10,2	10,83	11,5	14,5	10,2	1,7
9	Q_MAN	225	0,95	0,23	0,4	0,8	0,93	1,1	1,56	24,6	1,7
10	Q_LIG_INS	226	24,82	1,48	21,03	23,92	24,89	25,65	32,03	6	1,3
11	Q_LIG_SOL	226	3,49	1,03	1,38	2,75	3,55	4,17	6,32	29,6	1,3
12	Q_LIG_TOT	226	28,71	1,33	25,98	27,73	28,78	29,55	35,63	4,6	1,3
13	Q_EXT	226	1,84	0,68	0,8	1,3	1,69	2,27	3,96	36,9	1,3
14	Q_REL_SG	226	3,14	0,39	2,03	2,86	3,12	3,43	3,98	12,6	1,3
15	Q_N	201	0,09	0,03	0,01	0,05	0,1	0,11	0,18	38,9	12,2
16	Q_P	200	0,005	0,004	0	0,002	0,004	0,01	0,02	80,8	12,7
17	Q_K	201	0,04	0,02	0,02	0,03	0,04	0,05	0,11	39,5	12,2
18	Q_NA	201	0,01	0,004	0,002	0,01	0,01	0,01	0,03	38,5	12,2

Continua na próxima página

Tabela 4 – continuação da página anterior

#	Variável	Qtd	Méd	DP	Mín	25%	50%	75%	Máx	CV%	Nulo%
19	Q_CA	201	0,03	0,02	0,003	0,02	0,03	0,04	0,12	66,7	12,2
20	Q_MG	201	0,01	0,004	0	0,01	0,01	0,01	0,02	42,8	12,2
21	Q_CU	197	3,18	3,07	0	1,14	2,15	3,54	13,31	96,6	14
22	Q_ZN	197	3,27	2,14	0,04	1,56	2,78	4,66	11,05	65,4	14
23	Q_FE	187	13,1	11,5	0,49	5,98	10,88	16,16	95,94	87,8	18,3
24	Q_MN	198	14,72	14,21	0	7,09	10,44	17,72	79,57	96,6	13,5
25	Q_B	197	1,96	0,9	0	1,57	1,79	2,08	7,79	46	14
26	Q_S	200	0,01	0,01	0	0,003	0,01	0,01	0,03	97,8	12,7
27	P_KAPPA	229	20,55	3,29	13,37	18,57	20,26	22,23	31,86	16	0
28	P_RB	229	54,5	1,65	48,62	53,37	54,39	55,6	59,04	3	0
29	P_REJ	229	0,54	0,36	0,07	0,3	0,45	0,7	2,05	65,8	0
30	P_RD	229	53,95	1,68	48,13	52,71	53,93	55,2	58,85	3,1	0
31	L_ARES	229	2,24	1,46	0	1,24	2,21	3,18	6,97	65,1	0
32	L_SOLID	229	12,51	0,9	10,1	11,92	12,64	13,24	14,27	7,2	0
33	L_CALCI	229	62,08	2,65	57,82	60,37	61,39	63,12	75,16	4,3	0
34	L_PH	229	12,65	0,53	10,7	12,38	12,67	12,95	14,05	4,2	0
35	L_PCALO	229	3601,99	201,57	3195,75	3436,07	3517,24	3780,69	4043,5	5,6	0
36	B_CLO2TOT	228	14,82	3,38	8,88	12,44	14,02	16,91	27,09	22,8	0,4
37	B_CLO2_KAPPA	228	0,72	0,11	0,42	0,65	0,71	0,79	1,26	15	0,4
38	B_OO_HEXA	228	58,74	9,19	33,92	53,83	59,01	64,05	93,9	15,6	0,4
39	B_OO_VISCO	228	48,5	9,48	23,13	41,61	47,78	54,08	77,33	19,6	0,4
40	B_OO_KP_ENT	228	20,75	3,25	13,85	18,79	20,39	22,28	31,86	15,7	0,4
41	B_OO_KP_SAI	228	15,24	2,36	10,8	13,55	14,95	16,7	24,5	15,5	0,4
42	B_OO_ALVU	228	40,79	3,8	27,51	38,52	41,48	43,7	49,41	9,3	0,4
43	B_OO_O2	228	22,79	3,55	15,24	20,67	22,43	24,49	35,05	15,6	0,4
44	B_D0_CLO2	228	10,44	1,61	7,39	9,29	10,26	11,43	16,77	15,4	0,4
45	B_D1_CLO2	228	5,22	2,58	0,87	3,19	4,66	6,71	14,07	49,4	0,4
46	B_EP_mKAPPA	227	3,36	0,41	2,32	3,1	3,33	3,6	4,52	12,1	0,9
47	B_EP_ALVU	227	75,15	3,25	63,36	73,5	75,88	77,6	80,8	4,3	0,9
48	B_P_VISC	228	29,9	6,13	15,7	25,68	29,45	34,08	47,85	20,5	0,4

Fonte: Autor (2024).

A idade média das amostras analisadas foi de 6,59 anos, com um desvio padrão de 0,6 anos, indicando uma variação moderada dentro do conjunto de dados (CV=9,1%), embora esta variação seja aceitável idealmente deveria apresentar variações ainda menores uma vez que afeta o comportamento de outras variáveis, no entanto isolar estes efeitos ainda se mostra desafiador em trabalhos desta natureza. A F_DB, densidade básica do material, apresentou uma média de 500,54 kg/m³ e um coeficiente de variação de 6,3%, indicando baixa variabilidade, a presença de 6,1% de dados nulos confirma necessidade de trabalhar imputação de dados.

Quanto aos componentes químicos, a celulose (Q_CEL) e a hemicelulose (Q_HEM) tiveram médias de 47,3% e 13,11%, respectivamente, com baixos coeficientes de variação (CVs) (4,3% e 8,5%). Os extrativos (Q_EXT) e a relação S/G (Q_REL_SG) apresentaram maior variabilidade, com coeficientes de variação de 36,9% e 12,6%, respectivamente. Estes são indicativos de variação significativa e que pode contribuir para as diferenças entre os grupos. Já Q_ARA (Arabinose) e Q_GAL (Galactose) destacam-se com altos CVs de 25.2% e 28.6%, respectivamente, apontando para uma variabilidade significativa para estas variáveis.

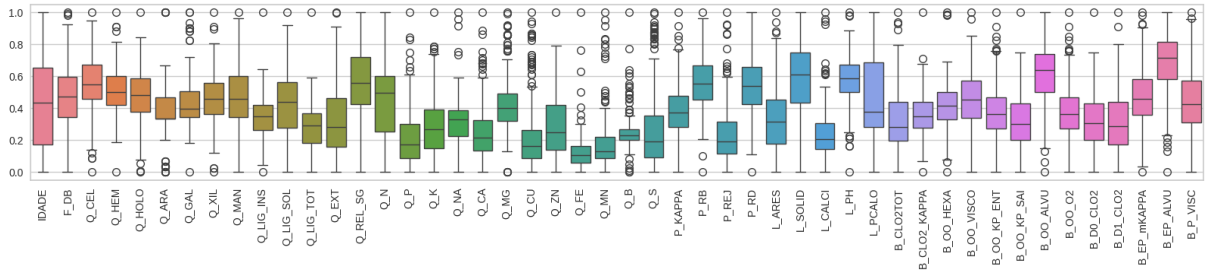
Quanto aos mineiras Q_CU (Cobre) e Q_FE (Ferro) apresentaram uma variabilidade alta com CVs de 96.6% e 87.8%, respectivamente. Estas variáveis também apresentaram porcentagem significativa de dados nulos (14.0% e 18.3%), sendo necessário aplicação de imputação de dados.

Nas variáveis de polpação o P_KAPPA, que reflete lignina residual da polpa, apresentou um desvio padrão de 3.29 e um CV de 16.0%, refletindo variações moderadas no conjunto analisado, o L_ARES (AE residual) com um CV de 65.1% e P_REJ (Rejeito) com um CV de 65.8% apresentaram alta variabilidade nos dados analisados.

Por fim B_CLO2TOT (Consumo Total de ClO2) e B_OO_VISCO (Viscosidade) apresentaram CVs de 22.8% e 19.6%, respectivamente, o que demonstra uma variação moderadas entre os materiais analisados, já o B_EP_ALVU (Alvura após estágio EP) mostra uma variabilidade mais controlada com CV de 4.3%.

Na Figura 2 é possível avaliar a distribuição dos dados. Variáveis como Q_LIG_TOT, Q_EXT e Q_REL_SG apresentaram menor dispersão e são mais centralizadas, sugerindo maior homogeneidade, em contraste, Q_CU e Q_FE exibem alta variabilidade e sugestão de potenciais outliers. Uma investigação mais detalhada por meio da estratificação dos dados em categorias (ESPECIE e REGIAO) dirimiu esta dúvida. Além disso a análise da posição da mediana dentro da caixa no boxplot oferece indícios sobre a assimetria dos dados, assim podemos observar que várias características apresentam indícios de uma distribuição assimétrica. Para verificar cada caso, foi realizado uma análise visual por meio de histogramas associados a boxplox e gráficos quartil-quartil (Q-Q) conforme exemplos nas Figuras 3 e 4

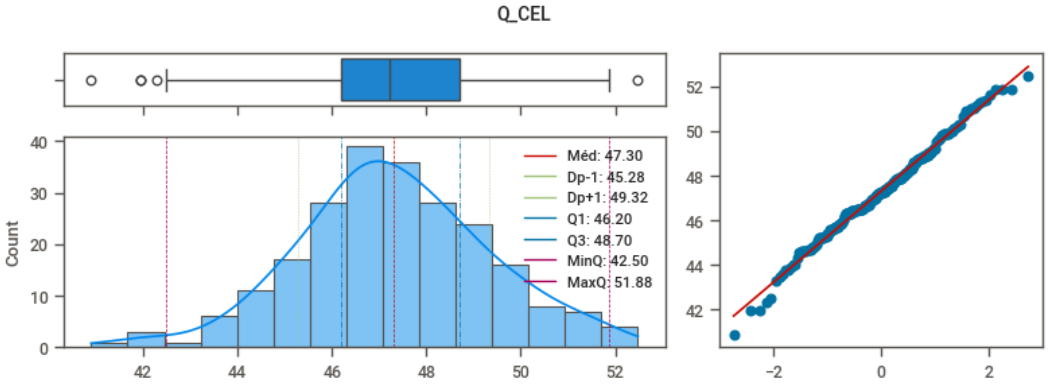
Figura 2 – Boxplot normalizado das características avaliadas



Fonte: Autor (2024).

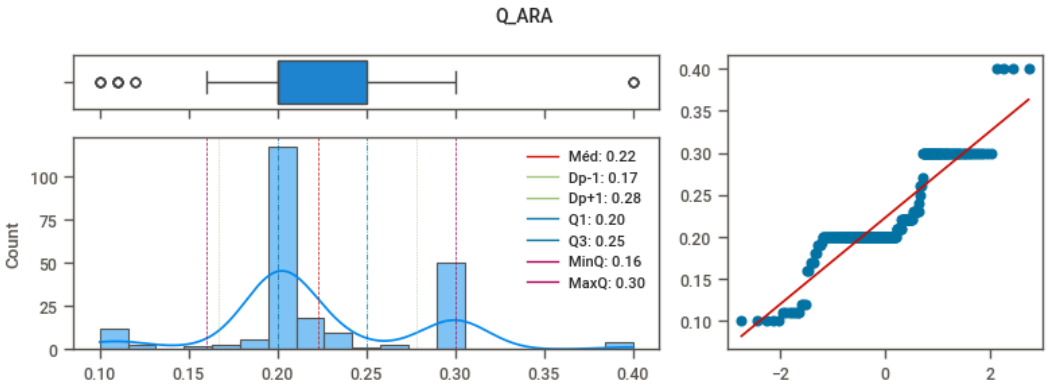
Além do gráfico de Q-Q Plot foi realizado o teste de Shapiro-Wilk para as variáveis avaliadas. No caso de Q_CEL o teste apresentou valor de 0,993 para um p-value igual a 0,368 sendo possível interpretar que provavelmente Q_CEL segue uma distribuição normal (aceita H0 quando p_valor > 0,05). Já Q_ARA apresentou valor de 0,832 para um p-value igual a 0,0 sendo possível interpretar que Q_ARAL não segue uma distribuição normal (rejeitar H0 quando p_valor <= 0,05).

Figura 3 – Análise da Q_CEL utilizando BoxPlot, Histogramas e Q-Q Plot



Fonte: Autor (2024).

Figura 4 – Análise da Q_ARA utilizando BoxPlot, Histogramas e Q-Q Plot



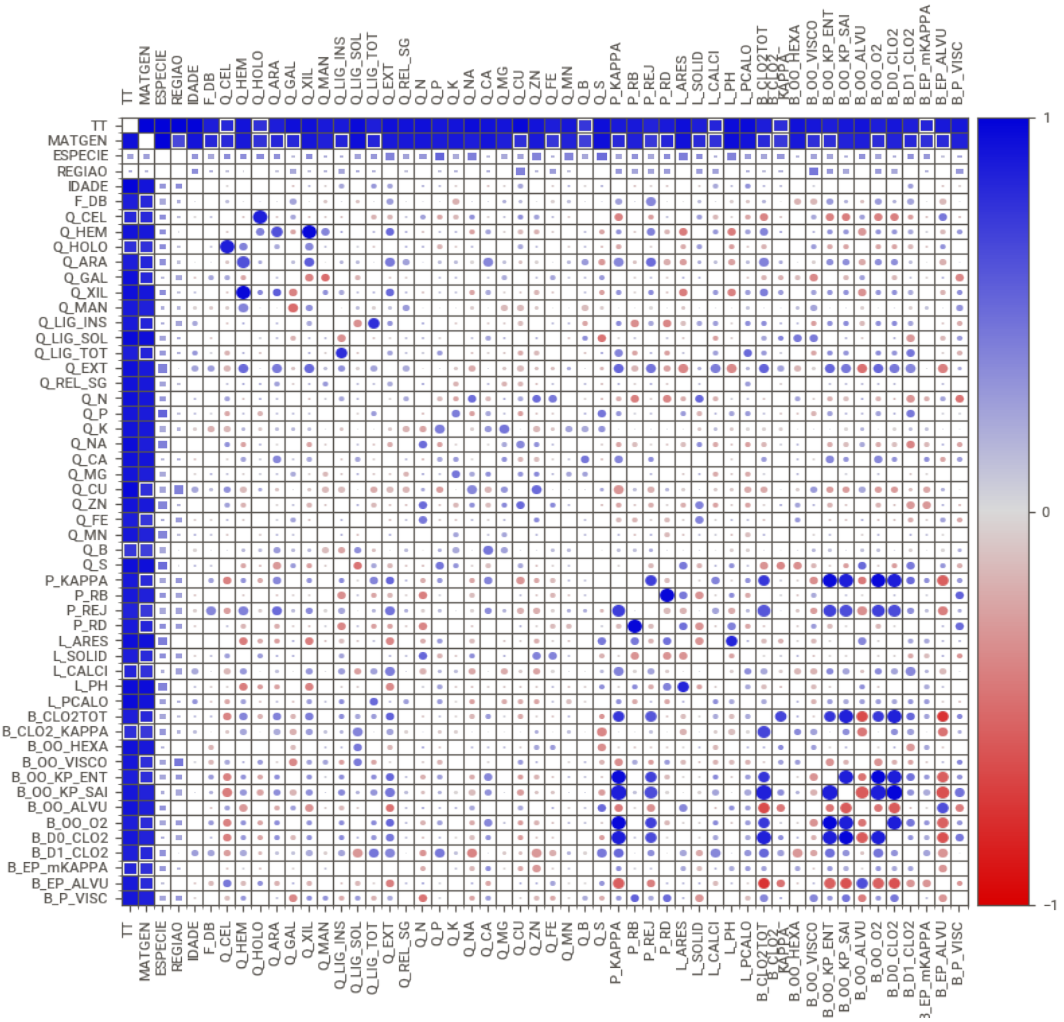
Fonte: Autor (2024).

Na Figura 5 é apresentado a matriz de correlação das variáveis analisadas. Os "quadrados"na matriz representam associações categóricas, utilizando métricas como o coeficiente de incerteza e a razão de correlação, ambos variando de 0 a 1. O coeficiente de incerteza é particularmente notável por ser assimétrico, significando que o valor derivado dos rótulos de linha indica o quanto de informação é fornecida aos rótulos no topo da matriz. Esta métrica ajuda a entender a quantidade de incerteza que uma variável pode remover da outra, destacando assim a dependência direcional entre variáveis categóricas, por outro lado, os "círculos"na matriz indicam correlações numéricas simétricas, calculadas pelo coeficiente de correlação de Pearson, que varia entre -1 e 1. Valores próximos de 1 ou -1 indicam uma forte correlação positiva ou negativa, respectivamente, enquanto valores próximos de 0 sugerem nenhuma correlação linear. Essa representação gráfica é útil para rapidamente identificar e interpretar as relações entre múltiplas variáveis, sejam elas categóricas ou numéricas.

As variáveis ESPECIE e REGIAO apresentaram associações fracas com TT, coeficientes de 0,22 e 0,17, respectivamente. A variável ESPECIE apresentou uma associação desprezível com REGIAO (0,01). As associações numéricas de ESPECIE com outras variáveis como Q_P (0,51) e Q_S (0,47) são moderadas. Em relação à variável REGIAO, observa-se uma associação forte com MATGEN (0,65) e uma associação fraca com TT (0,17). B_OO_VISCO (0,43) e Q_CU (0,37) apresentaram as associações numéricas mais relevantes com REGIAO, embora consideradas fracas, indicando pouca influência da REGIAO sobre esses fatores.

Com relação as variáveis numéricas chama atenção algumas correlações, Q_HOLO tem uma correlação muito forte com Q_CEL (r = 0,86), indicando uma relação quase direta entre essas variáveis. Da mesma forma, Q_XIL e Q_HEM também mostram uma correlação muito forte (r = 0,97), sugerindo uma alta dependência linear entre elas. P_KAPPA com B_OO_O2 (r = 0,97) e B_OO_KP_ENT (r = 0,96), e Q_LIG_INS com Q_LIG_TOT (r = 0,78), todas indicaram fortes relações lineares.

Figura 5 – Matriz de correlação das características analisadas



Fonte: Autor (2024).

Embora algumas variáveis apresentem forte indicação de multicolinearidades a aplicação de técnicas de redução de dimensionalidade permitiu trabalhar com estas variáveis sem a necessidade de exclusão das mesma e sem efeito negativo nos algoritmos de agrupamento.

4.1 Análise de Agrupamento

4.1.1 Identificação e validação:

Para proceder as análises de agrupamento primeiramente foi realizado a padronização dos dados por meio da biblioteca StandardScaler do sklearn. Com os dados normalizados procedeu-se a incorporação dos dados categóricos ESPECIE e REGIAO por meio da tranformação OneHotEncoder do pacote sklearn.

Para avaliar os agrupamentos foram testados três técnicas de redução de dimensionalidade (PCA, UMAP e T-SNE). Para cada técnica de redução de dimensionalidade foram testados três métodos de agrupamento não supervisionado (KMeans, Agglomerattive Clustering e HDBSCAN). Para obtenção dos parâmetros descritos na Tabela 5 foram realizadas rodadas de testes buscando obter o maior coeficiente de silhueta.

A Análise de Componentes Principais (PCA) aplicada com KMeans resultou em um escore de silhueta de 0,215 com 6 clusters, indicando um agrupamento moderadamente coeso. A técnica de redução de dimensionalidade UMAP, com a definição de 10 componentes produziu grupos distintos e coesos para todos os métodos de agrupamento testados.

T-SNE, configurado para três componentes e posteriormente agrupado usando KMeans, obteve um escore de silhueta de 0,351 com 7 clusters, apresentando uma eficácia razoável na formação de grupos claramente separados.

A aplicação do HDBSCAN após as reduções dimensionais destacou o potencial do UMAP combinado com HDBSCAN, onde o número mínimo de pontos por cluster foi definido como 4, resultando em 5 clusters com um escore de silhueta de 0,748. Esta combinação provou ser a mais eficaz, indicando a presença de clusters densamente agrupados que são bem capturados por este método.

Tabela 5 – Parâmetros de Técnicas de Redução de Dimensionalidade e de Agrupamentos

PCA	**UMAP**	**T-SNE**
n_components=0.8 random_state=1 variância explicada = 0.81 qtd_componentes = 14	n_components=10 random_state=30 n_neighbors=15 metric='euclidiano' min_dist=0.05	n_components=3 random_state=2024
PCA-KMeans n_clusters=6 init='KMeans++' n_init=10 max_iter=300 tol=1e-04 random_state=2024 escore de silhueta = 0.215	**UMAP-KMeans** n_clusters=5 init='KMeans++' n_init=10 max_iter=300 tol=1e-04 random_state=2024 escore de silhueta =0.748	**TSNE-KMeans** n_clusters=7 init='KMeans++' n_init=10 max_iter=300 tol=1e-04 random_state=2024 escore de silhueta =0.351
PCA-Aglomerativo n_clusters=6 linkage='ward' escore de silhueta = 0.2012	**UMAP-Aglomerativo** n_clusters=5 linkage='ward' escore de silhueta = 0.748	**TSNE-Aglomerativo** n_clusters=2 linkage='ward' escore de silhueta = 0.302
PCA-HDBSCAN min_cluster_size=7	**UMAP-HDBSCAN** min_cluster_size=4	**TSNE-HDBSCAN** min_cluster_size=5

Continua na próxima página

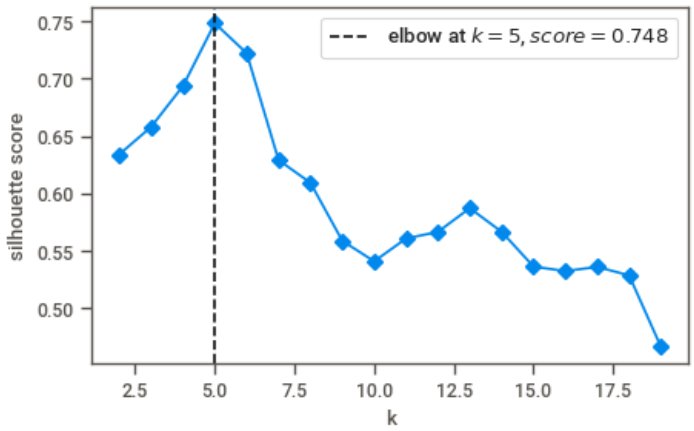
Tabela 5 – continuação da página anterior

PCA	UMAP	T-SNE
min_samples=3	min_samples=3	min_samples=3
cluster_selection_epsilon=0.1	cluster_selection_epsilon=1.0	cluster_selection_epsilon=0.1
n_clusters=4	n_clusters=5	n_clusters=3
escore de silhueta = 0.1262	escore de silhueta = 0.748	escore de silhueta = 0.200

Fonte: Autor (2024).

Com exceção do UMAP que já determina a quantidade de agrupamentos, os demais métodos testados possibilitaram avaliação visual da curva de cotovelo e do gráfico de avaliação de silhueta conforme Figuras 6 e 7.

Figura 6 – Gráfico do cotovelo para o escore de silhueta do agrupamento KMeans

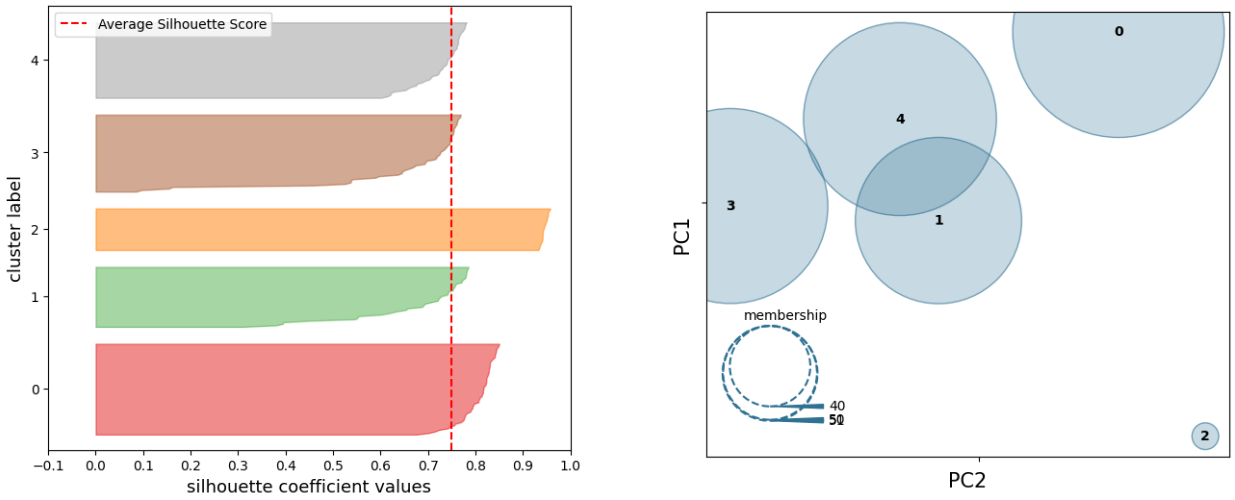


Fonte: Autor (2024).

A Figura 6 revela um ponto de inflexão ótimo em k=5 com um escore de silhueta de 0,748 para o agrupamento KMeans. Este resultado sugere que a divisão do conjunto de dados em cinco clusters é a configuração mais apropriada para maximizar a distinção entre os grupos, proporcionando uma coesão interna ideal dentro dos clusters e uma separação adequada entre eles neste ponto.

A análise dos resultados do agrupamento KMeans com 229 amostras distribuídas em 5 centros revelou uma configuração de cluster eficiente, como pode ser visto no gráfico de silhueta à esquerda da Figura 7. A linha tracejada vermelha indica a pontuação média de silhueta, que cruza o eixo y em torno de um valor alto, sugerindo um agrupamento adequado. Os coeficientes de silhueta estendem-se principalmente acima do valor médio, com poucas amostras de sobreposição "negativa", indicando que a maioria dos pontos foi alocada nos grupos apropriados ou seja mais próximos dos centros de cada grupo.

Figura 7 – Gráfico de silhueta e mapa de distância intercluster



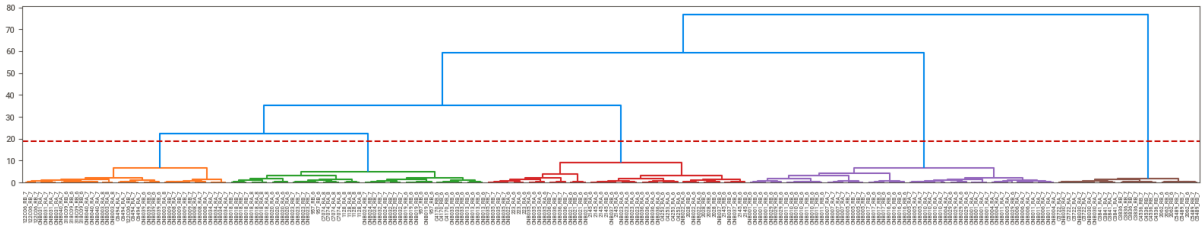
Fonte: Autor (2024).

A distribuição dos coeficientes de silhueta mostra que o cluster 0 possui a maior largura de silhueta, sugerindo que é o grupo mais coeso e separado. Por outro lado, o cluster 1 apresenta a mais ampla gama de valores de silhueta, indicando uma variabilidade maior dentro do grupo, o que pode significar a necessidade de uma subdivisão.

O mapa de distância intercluster à direita da Figura 7, ilustra a separação entre os clusters. Os tamanhos relativos dos círculos representam a distância entre os centros dos clusters, com clusters maiores sugerindo uma maior separação. A posição e sobreposição dos círculos indicam que, embora haja uma distinção clara entre os clusters, existe uma interseção entre eles, especialmente entre os clusters 3 e 4, que podem compartilhar algumas características comuns.

É importante observar que tal análise visual não se aplica diretamente aos métodos de agrupamento aglomerativo (Agglomerattive Clustering) e método baseado em densidade (HDBSCAN). Estes algoritmos operam sob premissas distintas que não são capturadas de maneira adequada por meio dos métodos de visualização utilizados para KMeans. Para estes casos foram avaliados os agrupamentos conforme métricas descritas na Tabela 6 e análises das projeções dos agrupamentos conforme Figuras 8 e 9.

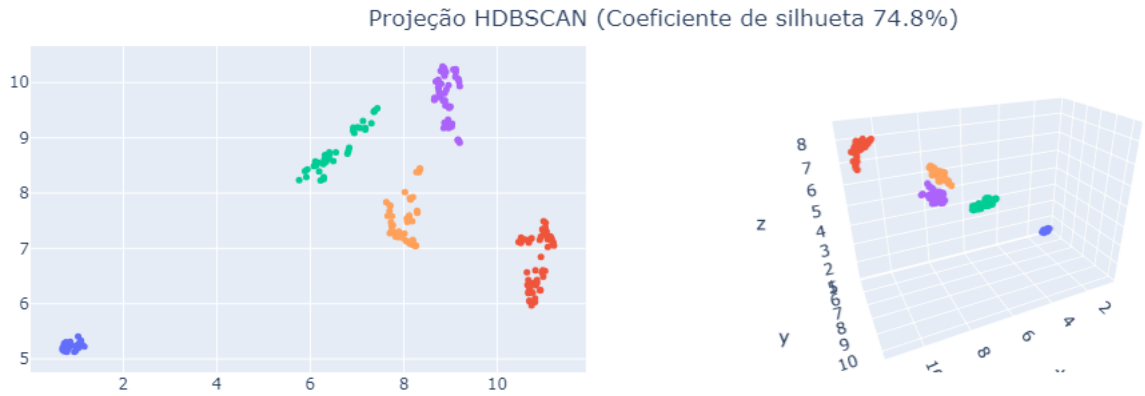
Figura 8 – Dendograma gerado com redução UMAP e linkage Ward para avaliação de métodos hierárquicos



Fonte: Autor (2024).

Na Figuras 9 é possível analisar visualmente o resultado do agrupamento gerado pelo HDBSCAN, reduzido por UMAP, por meio da projeção dos grupos nas componentes 1 e 2 do gráfico à esquerda e uma visão tridimensional do gráfico da projeção dos componentes à direita da figura.

Figura 9 – Projeção dos clusters reduzidos com UMAP e agrupados com HDBSCAN



Fonte: Autor (2024).

Para comparar as diferentes técnicas de redução de dimensionalidade associadas ao métodos de agrupamento não supervisionados foi utilizado as métricas descritas na Tabela 6

Tabela 6 – Comparação de Algoritmos de Clustering sob Diferentes Técnicas de Redução de Dimensionalidade

Técnica/Índice	Índice Silhueta	Índice Davies-Bouldin	Índice Calinski-Harabasz
PCA - KMeans	0.21	1.58	37.29
PCA - Agglomerative Clustering	0.2	1.88	39.1
PCA - HDBSCAN	0.13	2.61	24.69
UMAP - KMeans	0.75	0.34	1852.59
UMAP - Agglomerative Clustering	0.75	0.34	1852.59
UMAP - HDBSCAN	0.75	0.34	1852.59
t-SNE - KMeans	0.35	1.00	126.59
t-SNE - Agglomerative Clustering	0.30	1.35	106.53
t-SNE - HDBSCAN	0.20	2.97	61.09

Fonte: Autor (2024).

Os resultados apresentados na Tabela 6 possibilitam avaliar o desempenho de cada abordagem. Para a redução com PCA verifica-se que o KMeans registrou um Índice Silhueta de 0,21, apontando para uma qualidade moderada nos agrupamentos formados. O Índice Davies-Bouldin de 1,58, embora não indique clusters muito densos ou bem separados, sugere uma organização razoável. O índice Calinski-Harabasz foi de 37,29, reforçando a interpretação de que os clusters não são extremamente definidos. Por sua vez, o Agglomerative Clustering apresentou uma ligeira melhoria no índice Silhueta com 0,2, enquanto o Davies-Bouldin aumentou para 1,88, e o Calinski-Harabasz para 39,1, indicando que os grupos podem ter uma separação mais nítida. Já o HDBSCAN, associado à PCA, mostrou uma diminuição em todas as métricas sinalizando uma estrutura de agrupamento ineficiente.

Utilizando a redução UMAP, o desempenho dos algoritmos KMeans, o Agglomerative Clustering e o HDBSCAN compartilharam índice Silhueta de 0,75. Isso sugere separação eficiente dos grupos formados, o índice Davies-Bouldin de 0,34, indica agrupamentos compactos e bem delimitados. O índice Calinski-Harabasz registrou 1852,59 para as três combinações, o que denota uma distinção clara e consistência interna dos grupos.

Já no t-SNE, tanto o KMeans quanto o Agglomerative Clustering mantiveram um desempenho modesto, com índices Silhueta de 0,35 e 0,30, respectivamente, indicando uma definição razoável dos clusters. O Davies-Bouldin para o KMeans foi de 1,00 e para o Agglomerative Clustering, 1,35, sugerindo clusters menos densos em comparação com o UMAP. Os índices Calinski-Harabasz foram de 126,59 para o KMeans e 106,53 para o Agglomerative Clustering, o que pode indicar uma eficiência inferior na formação dos clusters. O HDBSCAN, em combinação com o t-SNE, enfrentou desafios significativos, com as menores pontuações em todas as métricas: um índice Silhueta de 0,20, Davies-Bouldin de 2,97 e Calinski-Harabasz de 61,09, evidenciando uma formação de grupos ineficaz.

4.1.2 Caracterização dos grupos:

Todos os agrupamentos testados (KMeans, Agglomerative Clustering e HDBSCAN) com a redução de dimensionalidade UMAP produziram uma alocação de elementos iguais entre seus respectivos clusters conforme descrito na Tabela 6. O método HDBSCAN foi selecionado devido ao fato de identificar clusters com base em densidade sem a necessidade de especificar o número de grupos expondo os padrões naturais dos dados, outra característica considerada é a baixa sensibilidade a outliers e capacidade de manipular clusters de diferentes tamanhos e formas proporcionam uma flexibilidade na análise, tudo isso foi considerado ideal tendo em vista a complexa estrutura do conjunto de dados multidimensional avaliado.

Na Tabela 7 é apresentados os elementos que compõem cada cluster gerado pelo UMAP com HDBSCAN. Neste agrupamento podemos observar uma distribuição homogênea dos elementos, o cluster 0 (zero) apresentou a menor alocação de elementos com 7 no total e o cluster 1 (um) o maior número, totalizando 15 elementos no grupo.

Além disso é possivel verificar que nem a espécie (GRA, URO_GRA, URO) nem a região (RA, RB) são fatores determinantes exclusivos para a alocação de elementos em cada cluster. A distribuição mista de espécies e a presença de elementos de ambas as regiões em vários clusters indicam uma influência limitada destas categorias na formação dos grupos. Este fato susinta a hipótese de que outros atributos multidimensionais, possivelmente mais sutis e complexos, estão influenciando o agrupamento.

Com exceção do CNB023 todos os clones que possuem análises em diferentes regiões (alta e baixa) foram agrupados no mesmo clustes, este é um fato intrigante tendo em vista a prática de se observar resultados superiores em clones de região alta em relação aos da baixa.

Tabela 7 – Clusters Identificados pelo UMAP com HDBSCAN e a Quantidade de Elementos

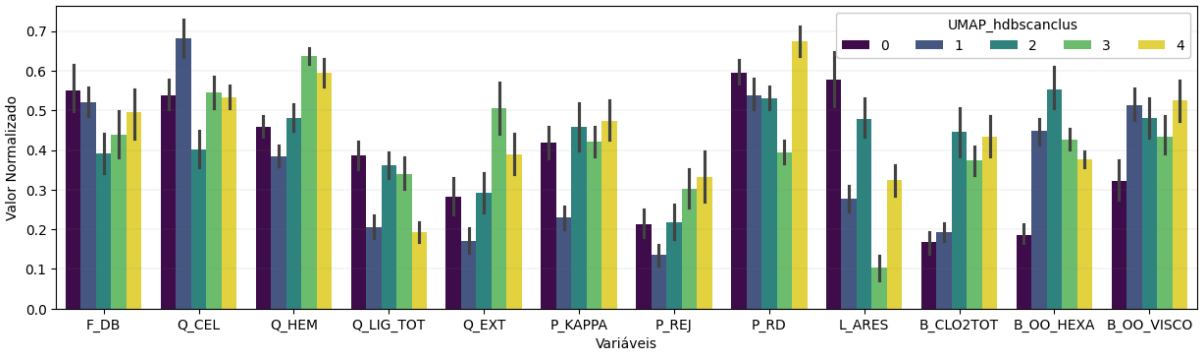
Cluster	**GRA**	**URO_GRA**	**URO**	**RA**	**RB**	**Qtd.**	**Elementos**
0	1	6	0	3	4	7	2062_RB_6, C3836_RB_7, C3841_RA_7, C4536_RB_7, C5489_RB_7, C5722_RA_7, CNB030_RA_7
1	1	14	0	8	7	15	CNB001_RA_7, CNB001_RB_6, CNB004_RA_7, CNB005_RA_7, CNB005_RB_7, CNB006_RA_7, CNB007_RA_7, CNB010_RA_7, CNB010_RB_6, CNB011_RA_7, CNB011_RB_6, CNB016_RB_6, CNB017_RB_6, CNB029_RA_6, CNB029_RB_6
2	1	9	0	6	4	10	122C06_RB_7, 318C09_RB_6, C6494_RA_7, CNB003_RA_8, CNB008_RA_7, CNB009_RB_7, CNB031_RA_7, CNB034_RA_7, CNB039_RB_6, CNB040_RA_7
3	2	9	2	4	9	13	1128_RA_8, 957_RB_6, C4170_RB_6, C7074_RA_8, CNB013_RB_6, CNB018_RA_8, CNB018_RB_7, CNB019_RB_6, CNB020_RA_6, CNB023_RB_7, CNB024_RB_7, CNB032_RB_7, CNB033_RB_6
4	1	10	2	7	6	13	2028_RB_6, 2145_RA_6, 2145_RB_6, 223_RA_6, C4253_RA_6, CNB021_RB_6, CNB022_RA_6, CNB023_RA_6, CNB035_RA_6, CNB035_RB_7, CNB036_RA_6, CNB036_RB_7, CNB037_RB_6

GRA, URO_GRA e URO - Quantidade de elementos em cada espécie
RA e RB - Quantidade de elementos em cada região (RA-Região Alta e RB-Região Baixa

Fonte: Autor (2024).

A análise do perfil das variáveis obtidas pelo agrupamento HDBSCAN com redução de dimensionalidade UMAP, conforme ilustrado na Figura 10, oferece uma visão detalhada da qualidade da madeira e dos processos de polpação e branqueamento em cada cluster. Observa-se que a densidade básica (F_DB) não apresenta diferenças significativas entre os clusters, indicando a necessidade de considerar outras características em conjunto para uma segmentação eficaz. A celulose (Q_CEL) mostra uma distinção clara entre os clusters, especialmente o cluster 1, que se destaca dos demais. A hemicelulose (Q_HEM) e a lignina total (Q_LIG_TOT) apresentam variações nas médias com intervalos de confiança sobrepostos, sugerindo similaridades entre alguns clusters. Os extrativos (Q_EXT) variam significativamente, com o cluster 3 mostrando a maior concentração média e o cluster 1 a menor, indicando uma composição diferenciada. O número kappa (P_KAPPA) e o rejeito (P_REJ) também revelam contrastes notáveis entre os clusters, com o cluster 4 destacando-se em ambos os casos. O rendimento depurado (P_RD), AE residual (L_ARES), consumo total de dióxido de cloro (B_CLO2TOT), a concentração de ácidos hexenurônicos (B_OO_HEXA) e viscosidade (B_OO_VISCO) apresentam diferenças marcantes entre os clusters, com o cluster 4 mostrando a maior média em viscosidade e o cluster 0 a menor.

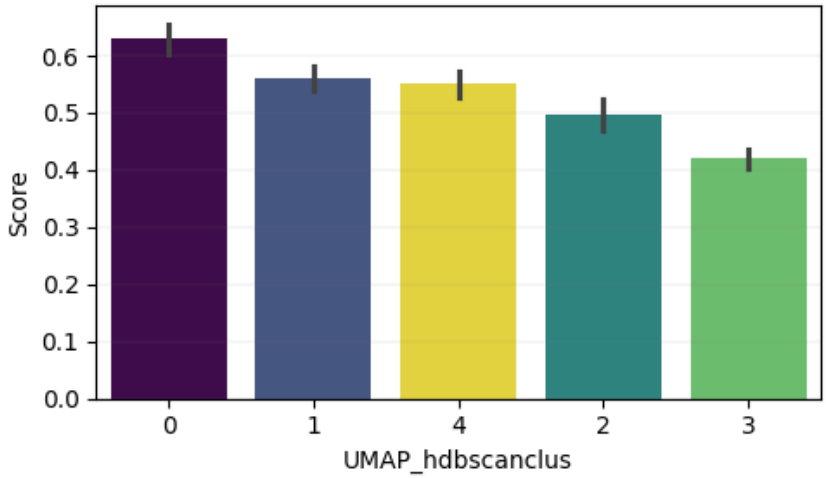
Figura 10 – Perfil normalizado das variáveis por cluster no agrupamento UMAP HDBSCAN



Fonte: Autor (2024).

Para permitir uma avaliação qualitativa dos clusters formados foi criado um índice de pontuação (’score’), incorporando critérios que impactam a qualidade e eficiência do processo de polpação e branqueamento. A variável ’P_RD’ (Rendimento Depurado), recebeu peso de 0,35, onde valores mais altos indicam melhor desempenho. A variável ’B_CLO2TOT’ (Consumo Total de Dióxido de Cloro) recebeu peso de 0,25 e inverte a lógica, favorecendo valores menores. A densidade básica (’F_DB’), recebeu peso de 0,20, considerando melhor os valores mais altos, ’L_ARES’ (AE Residual) e ’B_OO_VISCO’ (Viscosidade), receberam pesos de 0,15 e 0,05, respectivamente, onde valores mais altos indicam melhor resultado. A combinação desses critérios em um índice único fornece uma referência para comparar os clusters e identificar quais representam as melhores e piores condições em termos das métricas para as características avaliadas. A Figura 11 apresenta a classificação após os cálculos, considerando as características citadas e os respectivos pesos. Assim é possível classificar os grupos sob uma perspectiva qualitativa sendo o melhor cluster o mais a esquerda e o pior o mais a direita da figura.

Figura 11 – Classificação dos clusters após cálculo do score

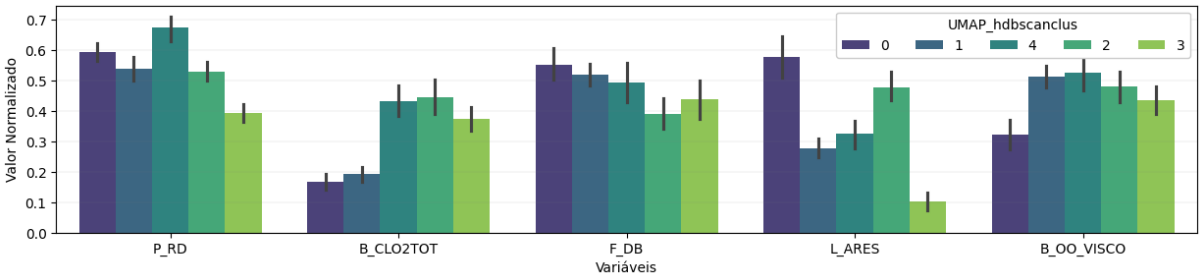


Fonte: Autor (2024).

O gráfico da Figura 11 apresenta a pontuação média (score) para cada cluster gerado pelo agrupamento HDBSCAN, usando a redução de dimensionalidade UMAP. Os scores foram calculados com base nos critérios citados anteriormente. Os resultados mostram que o cluster 0 obteve a pontuação mais alta, sugerindo ser o mais favorável qualitativamente, enquanto o cluster 3, com o score mais baixo, pode representar as condições menos desejáveis. Neste critério avaliado é possível observar uma sobreposição dos clusters 1 e 4 indicando que não haveria diferença estatística entre a adoção de um ou de outro.

A Figura 12 exibe o perfil das variáveis que constituem o score para cada cluster. De menira geral é possível perceber a influência dos pesos em cada variável no cálculo do score. Para algumas variáveis é possível perceber a sobreposição do intervalo de confiança entre os clusters reduzindo a discriminação entre os grupos. O cluster 0 e 1 podem ser considerados de fatos os grupos que apresentam os melhores resultados enquanto o grupo 3 apresentou a pior performance considerando a pontuação do score.

Figura 12 – Perfil das variáveis que compõem o cálculo do score



Fonte: Autor (2024).

Na Tabela 8 destaca-se as cinco características com os maiores índices de Gini para diferentes grupos de variáveis analisadas. O cálculo do índice se deu aplicando os valores médios obtidos para cada variável em cada cluster, o que permite uma visão quantitativa da distribuição interna dos grupos. Esses resultados auxiliam na interpretação de quais características contribuem mais para a separação dos dados e quão eficazes são os clusters formados.

Tabela 8 – Cinco Maiores Índices de Gini por Grupo Características

Grupo Característica	**Característica**	**Lista Médias Grupos**	**Índice Gini**
Todas	Q_S	0.831, 0.291, 0.24, 0.166, 0.084	0.402
Todas	Q_ZN	0.097, 0.493, 0.231, 0.265, 0.197	0.268
Todas	L_ARES	0.578, 0.278, 0.479, 0.103, 0.324	0.261
Todas	Q_LIG_SOL	0.086, 0.476, 0.614, 0.313, 0.525	0.252
Todas	Q_CU	0.226, 0.429, 0.137, 0.186, 0.156	0.231
Química	Q_LIG_SOL	0.086, 0.476, 0.614, 0.313, 0.525	0.252
Química	Q_EXT	0.282, 0.172, 0.292, 0.506, 0.389	0.189
Química	IDADE	0.501, 0.348, 0.612, 0.492, 0.277	0.148
Química	Q_LIG_TOT	0.388, 0.204, 0.362, 0.339, 0.193	0.148
Química	Q_ARA	0.286, 0.307, 0.383, 0.525, 0.502	0.134
Química (inorgânica)	Q_S	0.831, 0.291, 0.24, 0.166, 0.084	0.402
Química (inorgânica)	Q_ZN	0.097, 0.493, 0.231, 0.265, 0.197	0.268
Química (inorgânica)	Q_CU	0.226, 0.429, 0.137, 0.186, 0.156	0.231
Química (inorgânica)	Q_P	0.43, 0.137, 0.284, 0.17, 0.205	0.228
Química (inorgânica)	Q_FE	0.089, 0.196, 0.142, 0.163, 0.055	0.221
Polpação	L_ARES	0.578, 0.278, 0.479, 0.103, 0.324	0.261
Polpação	L_PCALO	0.598, 0.282, 0.757, 0.593, 0.316	0.194
Polpação	L_CALCI	0.347, 0.126, 0.203, 0.349, 0.262	0.183
Polpação	P_REJ	0.214, 0.135, 0.218, 0.303, 0.333	0.161
Polpação	L_SOLID	0.412, 0.706, 0.421, 0.753, 0.471	0.140
Branqueamento	B_D1_CLO2	0.644, 0.17, 0.354, 0.32, 0.335	0.215
Branqueamento	B_CLO2TOT	0.168, 0.192, 0.447, 0.373, 0.433	0.198
Branqueamento	B_OO_KP_SAI	0.211, 0.197, 0.435, 0.339, 0.434	0.173
Branqueamento	B_D0_CLO2	0.211, 0.199, 0.435, 0.34, 0.434	0.172
Branqueamento	B_OO_HEXA	0.186, 0.447, 0.554, 0.427, 0.375	0.162

Fonte: Autor (2024).

A característica Q_S no grupo "Todas"apresenta o maior Índice de Gini (0,402), indicando uma variação significativa e, consequentemente, uma alta capacidade discriminatória entre os clusters. Isso sugere uma distribuição bastante desigual dos dados, que favorece a diferenciação entre os grupos. Em contraste, características como Q_ZN e L_ARES mostram índices mais moderados (0,268 e 0,261, respectivamente), ainda contribuindo para a formação de grupos distintos, mas com menor variação entre os clusters.

No grupo "Química", a característica Q_LIG_SOL, também destacada no grupo geral, reitera sua importância com um Índice de Gini de 0,252. Isso ressalta sua utilidade na diferenciação dos clusters. Outras características, como IDADE e Q_EXT, apresentam índices mais baixos, indicando uma homogeneidade maior dentro dos clusters.

No contexto da "Polpação"e "Branqueamento", características como L_ARES, L_PCALO, e B_D1_CLO2 sobressaem com índices relativamente altos, denotando sua eficácia em separar claramente os tipos de polpa e processos de branqueamento, respectivamente. Tais constatações são importantes para a compreensão das dinâmicas internas dos clusters e podem auxiliar na seleção de variáveis para maximizar a eficiência dos agrupamentos ou estudos futuros.

5 Conclusões

O objetivo deste estudo foi investigar o uso de técnicas de agrupamento não supervisionado para classificar clones de eucalipto baseados em características químicas, físicas e do processo de polpação e branqueamento KRAFT. Os resultados demonstraram que o agrupamento HDBSCAN, combinado com a redução de dimensionalidade por UMAP, foi particularmente eficaz em identificar padrões distintos.

A utilização de índices de pontuação ("score") para classificar os grupos formados apresenta-se como uma estratégia interessante para aplicação dos clones agrupados em cada cluster. Além disso, permite gerar uma avaliação comparativa entre os grupos formados e as características da madeira e do processo KRAFT observados. Este método oferece uma visão clara das variações quantitativas entre os grupos, permitindo decisões estratégicas para a seleção de materiais genéticos ou mesmo na orientação do abastecimento segregado que atenda aos critérios descritos pelo índice.

O uso de um índice de pontuação ("score") para classificar grupos formados através de técnicas de agrupamento não supervisionado representa uma estratégia valiosa na análise de características complexas, como as associadas à produção de celulose de clones de eucalipto. No entanto, essa metodologia não está isenta de limitações, particularmente quanto à seleção de características e à atribuição de pesos, que são críticos para a construção do score. Estes aspectos podem ser considerados arbitrários e sujeitos à subjetividade do especialista, o que introduz um viés potencial na interpretação dos resultados.

A utilização do índice de Gini sugere forte influência de características como, por exemplo, L_ARES (Álcali residual), Q_LIG_SOL (Lignina solúvel), Q_EXT (extrativos) e B_CLO2TOT (Consumo total de dióxido de cloro) na capacidade discriminatória para formação dos clusters. A influência destas características no agrupamento de clones suscita uma oportunidade em realizar novos estudos a fim de reduzir a quantidade de características necessárias para alcançar resultados semelhantes aos obtidos neste trabalho.

Considerando a complexidade e variabilidade intrínseca dos dados, especialmente em um contexto de alta diversidade genética e variabilidade ambiental como é típico em silvicultura de eucalipto, sugere-se ampliar o trabalho de amostragem para cada clone avaliado, com o objetivo de melhorar a robustez e assegurar a generalização dos modelos de agrupamento testados neste estudo.

Referências

ABRANCHES, W. P. d. O. **Efeito do álcali efetivo na produção de polpa kraft branqueada de eucalipto**. Dissertação (Dissertação (Mestrado em Tecnologia de Celulose e Papel)) — Universidade Federal de Viçosa, Viçosa, 2017. 32f.

ALBON, C. **Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning**. First. O'Reilly Media, 2018. Accessed: 2023-04-15. Disponível em: <https://www.oreilly.com/library/view/machine-learning-with/9781491989371/ch04.html>.

ALMEIDA, D. P. d. **Influence of degree of delignification in the production of bleached eucalyptus kraft pulp**. Dissertação (Mestrado) — Universidade Federal de Viçosa, Viçosa, 2010. 89 f. Dissertação (Mestrado em Qualidade da madeira; Tecnologia de celulose e papel).

ALMEIDA, F. S. de. Influência da carga alcalina no processo de polpação lo-solids® para madeiras de eucalipto. 9 2003.

ALVES, I. C. N.; GOMIDE, J. L.; COLODETTE, J. L.; SILVA, H. D. da. Caracterização tecnológica da madeira de lt;igt;eucalyptus benthamiilt;/igt; para produção de celulose kraft. **Ciência Florestal**, Universidade Federal de Santa Maria, v. 21, p. 167–174, 3 2011. ISSN 1980-5098. Disponível em: <https://periodicos.ufsm.br/cienciaflorestal/article/view/2759>.

ANJOS, O.; GARCíA-GONZALO, E.; SANTOS, A. J. A.; SIMõES, R.; MARTíNEZ-TORRES, J.; PEREIRA, H.; GARCíA-NIETO, P. J. **Using Apparent Density of Paper from Hardwood Kraft Pulps to Predict Sheet Properties, based on Unsupervised Classification and Multivariable Regression Techniques**. 2015.

CAIXETA, R. P.; TRUGILHO, P. F.; ROSADO, S. C. da S.; LIMA, J. T. Propriedades e classificação da madeira aplicadas à seleção de genótipos de eucalyptus. **Revista Árvore**, v. 27, p. 43–51, 2 2003. ISSN 0100-6762.

CAMPOS, E. d. S.; FOELKEL, C. **A Evolução Tecnológica do Setor de Celulose e Papel no Brasil**. São Paulo, Brasil: ABTCP - Associação Brasileira Técnica de Celulose e Papel, 2016. ISBN 978-85-61701-02-4.

CARVALHO, D. M. D.; SILVA, M. R. D.; COLODETTE, J. L. Efeito da qualidade da madeira no desempenho da polpaÇÃo kraft wood quality effect on kraft pulping performance. **São Silvestre**, v. 84, p. 677–684, 2014. ISSN 0103-9954.

CASTRO, L. de; FERRARI, D. **Introdução a mineração de dados**. Saraiva Educação S.A., 2017. ISBN 9788547200992. Disponível em: <https://books.google.com.br/books?id=SSlrDwAAQBAJ>.

COLODETTE, J. L.; GOMES, F. J. B. **Branqueamento de Polpa Celulósica**. 1ª. ed. [S.l.]: Editora UFV, 2015. 816 p. ISBN 9788572695329.

COLODETTE, J. L.; MOUNTEER, A. H.; GOMES, F. J. B. Advanced technologies for eucalypt pulp bleaching. In: **Proceedings, ANQUE – International Congress of Chemical Engineering**. Seville, Spain: [s.n.], 2012.

DEPEC-BRADESCO. **Papel e Celulose**. DEPEC - BRADESCO, 2017. Disponível em: <https://www.economiaemdia.com.br/EconomiaEmDia/pdf/infset_papel_e_celulose.pdf>.

EUROPE, F. I. **Wood costs for pulp manufacturers worldwide are on the rise**. 2011. Online. Disponível em: <https://forestindustries.eu/content/wood-costs-pulp-manufacturers-worldwide-are-rise>.

FARDIM, P.; DURáN, N. **Retention of Cellulose, Xylan and Lignin in Kraft Pulping of Eucalyptus Studied by Multivariate Data Analysis: Influences on Physicochemical and Mechanical Properties of Pulp**. 2004. 514-522 p.

FEARON, O.; KUITUNEN, S.; RUUTTUNEN, K.; ALOPAEUS, V.; VUORINEN, T. Detailed modeling of kraft pulping chemistry. delignification. **Industrial and Engineering Chemistry Research**, American Chemical Society, v. 59, p. 12977–12985, 7 2020. ISSN 15205045.

FERNANDES, F. T.; FILHO, A. D. P. C. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. **Revista Brasileira de Saúde Ocupacional**, v. 44, 2019. ISSN 2317-6369.

FIGUEIREDO, J. d. C. **Influência da qualidade da madeira de eucalipto e dos processos de polpação Kraft Convencional e Compact Cooking TM na performance da linha de fibras e propriedades da polpa celulósica**. 67 p. Dissertação (Dissertação) — Universidade Federal de Viçosa, Viçosa, 2019.

FREIRE, S. M. **Bioestatística básica**. Rio de Janeiro: Ed. do Autor, 2021. PDF, [livro eletrônico]. ISBN 978-65-00-35696-0. Disponível em: <https://www.iaeng.org/publication/IMECS2012/IMECS2012_pp471-476.pdf>.

GOMIDE, J. L.; COLODETTE, J. L.; OLIVEIRA, R. C. de; SILVA, C. M. Caracterização tecnológica, para produção de celulose, da nova geração de clones de eucalyptus do brasil. **Revista Árvore**, v. 29, p. 129–137, 2 2005. ISSN 0100-6762.

GOMIDE, J. L.; NETO, H. F.; REGAZZI, A. J. Análise de critérios de qualidade da madeira de eucalipto para produção de celulose kraft. **Revista Árvore**, Sociedade de Investigações Florestais, v. 34, p. 339–344, 2010. ISSN 0100-6762. Disponível em: <https://www.scielo.br/j/rarv/a/6bddGZnPQ6d7b5ScWWrY4Fj/?lang=pt>.

GOUVêA, A. de F. G.; TRUGILO, P. F.; COLODETTE, J. L.; LIMA, J. T.; SILVA, J. R. M. da; GOMIDE, J. L. Avaliação da madeira e da polpação kraft em clones de eucaliptos. **Revista Árvore**, v. 33, p. 1175–1185, 12 2009. ISSN 0100-6762.

Indústria Brasileira de Árvores (Ibá). **Relatório Anual da Indústria Brasileira de Árvores 2023**. 2023. Acesso em: 26 abril 2024. Disponível em: <https://www.iba.org/datafiles/publicacoes/relatorios/relatorio-anual-iba2023-r.pdf>.

JOSWIAK, M.; PENG, Y.; CASTILLO, I.; CHIANG, L. H. Dimensionality reduction for visualizing industrial chemical process data. **Control Engineering Practice**, Elsevier Ltd, v. 93, 12 2019. ISSN 09670661.

LANNA, A. E.; COSTA, M. M.; FONSECA, M. J.; MOUNTEER, A.; COLODETTE, J. L.; GOMIDE, J. L. Maximizing pulp yield potential of cenibra´s wood supply. Belo Horizonte-MG-Brasil, p. 159–167, 2001.

LEITE, F. P. Efeito de fatores de produção vegetal na produtividade e na qualidade da madeira para a produção de polpa celulósica branqueada de eucalipto. 19 f. 2006.

LOVRIć, M.; ĐURIčIć, T.; TRAN, H. T.; HUSSAIN, H.; LACIć, E.; RASMUSSEN, M. A.; KERN, R. Should we embed in chemistry? a comparison of unsupervised transfer learning with pca, umap, and vae on molecular fingerprints. **Pharmaceuticals**, MDPI AG, v. 14, 8 2021. ISSN 14248247.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, v. 35, p. 85–94, 4 2021. ISSN 1806-9592.

MARANESI, G. L. **The influence of industrial process conditions on the quality properties of eucalyptus kraft pulp**. Dissertação (Mestrado) — Universidade Federal de Viçosa, Viçosa, 2010. 87 f. Dissertação (Mestrado em Qualidade da madeira; Tecnologia de celulose e papel).

MARINHO, N. P.; KLOCK, U.; LENGOWSKI, E. C.; MUñIZ, G. I. B. d.; ZAMARIAN, E. H. C. Características da polpa kraft extraída da espécie acácia-negra na produção de papel. **Floresta e Ambiente**, Instituto de Florestas da Universidade Federal Rural do Rio de Janeiro, v. 24, 2017. ISSN 2179-8087. Disponível em: <https://doi.org/10.1590/2179-8087.099214>.

MCKINNEY, W. **Python for Data Analysis**. [S.l.]: O'Reilly Media, Inc., 2012.

MOKFIENSKI, A.; COLODETTE, J. L.; GOMIDE, J. L.; CARVALHO, A. M. M. L. A importância relativa da densidade da madeira e do teor de carboidratos no rendimento de polpa e na qualidade do produto. **Ciência Florestal, Santa Maria**, v. 18, n. 3, 2008. Acessado em 04 julho 2018. Disponível em: <https://periodicos.ufsm.br/cienciaflorestal/article/view/451/0>.

MONTEIRO, C. R. **Classificação estrutural de proteínas por meio de aprendizado não supervisionado**. 79 p. Dissertação (Dissertação) — Universidade Federal de Viçosa, Viçosa, 2019.

MOORI, R. G.; MARCONDES, R. C.; ÁVILA, R. T. A análise de agrupamentos como instrumento de apoio à melhoria da qualidade dos serviços aos clientes. **Revista de Administração Contemporânea**, v. 6, p. 63–84, 4 2002. ISSN 1415-6555.

MORAIS, P. H. D. **Efeito da Idade da Madeira de Eucalipto na sua Química e Polpabilidade, e Branqueabilidade e Propriedades Físicas da Polpa**. Tese (Tese (Mestrado em Agroquímica)) — Universidade Federal de Viçosa, Viçosa, MG, 2008. 64 f.

NETO, H. F. **Qualidade da madeira de eucalipto para produção de celulose kraft**. Tese (Doutorado) — Universidade Federal de Viçosa, Viçosa, MG, 2012. 105 f.

PATINO, C. M.; FERREIRA, J. C. Confidence intervals: a useful statistical tool to estimate effect sizes in the real world. **Jornal Brasileiro de Pneumologia**, v. 41, p. 565–566, 12 2015. ISSN 1806-3713.

PLUMED, F. M. ınez; CONTRERAS-OCHANDO, L.; FERRI, C. esar; ORALLO, J. H. andez; KULL, M.; LACHICHE, N.; QUINTANA, M. ıa Jos Ram ırez; FLACH, P. Crisp-dm twenty years later: From data mining processes to data science trajectories. 2019. Disponível em: <www.sas.com>.

RONCORONI, F.; SANZ-MATIAS, A.; SUNDARARAMAN, S.; PRENDERGAST, D. Unsupervised learning of representative local atomic arrangements in molecular dynamics data. 2 2023. Disponível em: <http://arxiv.org/abs/2302.01465http://dx.doi.org/10.1039/D3CP00525A>.

SANTOS, S. R. **Influência da qualidade da madeira de híbridos de Eucalyptus grandis x Eucalyptus urophylla e do processo Kraft de polpação na qualidade da polpa branqueada**. Dissertação (Dissertação de Mestrado) — Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2005.

SENAI. Serviço Nacional de Aprendizagem Industrial. **Celulose**. São Paulo: SENAI - SP Editora, 2013. (Série Informações Tecnológicas). ISBN 9788565418706.

SINAGA, K. P.; YANG, M. S. Unsupervised k-means clustering algorithm. **IEEE Access**, Institute of Electrical and Electronics Engineers Inc., v. 8, p. 80716–80727, 2020. ISSN 21693536.

STEWART, G.; AL-KHASSAWENEH, M. An implementation of the hdbscan* clustering algorithm. **Applied Sciences (Switzerland)**, MDPI, v. 12, 3 2022. ISSN 20763417.

TIRELLI, A.; CARVALHO, D. O.; OLIVEIRA, L. A.; LIMA, J. P. de; COSTA, N. C.; SANTOS, R. R. dos. Unsupervised machine learning approaches to the q-state potts model. **European Physical Journal B**, Springer Science and Business Media Deutschland GmbH, v. 95, 11 2022. ISSN 14346036.

TRUGILHO, P. F.; BIANCHI, M. L.; GOMIDE, J. L.; SCHUCHARDT, U. ClassificaÇÃo de clones de eucalyptus sp visando À produÇÃo de polpa celulÓsica 1 classification of eucalyptus sp clones for kraft pulp production. p. 895–899, 2004.

VENTORIM, G.; CARASCHI, J. C.; COLODETTE, J. L.; GOMIDE, J. L. A influência dos ácidos hexenurônicos no rendimento e na branqueabilidade da polpa kraft. **Química Nova**, v. 32, p. 373–377, 2009. ISSN 0100-4042.