

UNIVERSIDADE ESTADUAL DE GOIÁS
CÂMPUS CENTRAL SEDE ANÁPOLIS - CET
PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU* EM ENGENHARIA AGRÍCOLA

CÉSAR ROBERTO PEREIRA BASTOS

**ALGORITMO FLORESTA ALEATÓRIA E ESTATÍSTICA MULTIVARIADA
UTILIZADOS PARA CLASSIFICAÇÃO DO ESTÁDIO DE MATURAÇÃO DE
CAGAITAS (*EUGENIA DYSENTERICA*)**

ANÁPOLIS - GO

2025

CÉSAR ROBERTO PEREIRA BASTOS

**ALGORITMO FLORESTA ALEATÓRIA E ESTATÍSTICA MULTIVARIADA
UTILIZADOS PARA CLASSIFICAÇÃO DO ESTÁDIO DE MATURAÇÃO DE
CAGAITAS (*EUGENIA DYSENTERICA*)**

Dissertação apresentada à Universidade Estadual de Goiás - UEG, Câmpus Central Sede Anápolis – CET, como parte das exigências do Programa de Pós-Graduação *Stricto Sensu* em Engenharia Agrícola, para obtenção do título de Mestre em Engenharia Agrícola.

Área de concentração: Engenharia de Sistemas Agroindustriais.

Orientador: Prof. Dr. Francisco Ramos de Melo.

ANÁPOLIS - GO

2025

CÉSAR ROBERTO PEREIRA BASTOS

ALGORITMO FLORESTA ALEATÓRIA E ESTATÍSTICA MULTIVARIADA
UTILIZADOS PARA CLASSIFICAÇÃO DO ESTÁDIO DE MATURAÇÃO DE
CAGAITAS (*EUGENIA DYSENTERICA*)

Dissertação defendida ao Programa de Pós-Graduação
Stricto Sensu em Engenharia Agrícola, Câmpus Anápolis
de Ciências Exatas e Tecnológicas Henrique Santillo,
Universidade Estadual de Goiás, para obtenção de título
de mestre.

Aprovado: 13 de maio de 2025.

Banca examinadora

Documento assinado digitalmente
 FRANCISCO RAMOS DE MELO
Data: 12/09/2025 13:45:18-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Francisco Ramos de Melo
Universidade Estadual de Goiás – UEG

Documento assinado digitalmente
 MARIA JOSELMA DE MORAES
Data: 12/09/2025 16:34:53-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. Maria Joselma de Moraes
Universidade Estadual de Goiás – UEG

Documento assinado digitalmente
 ALISON CARLOS FILGUEIRAS
Data: 18/09/2025 09:27:24-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Alison Carlos Filgueiras
Universidade Estadual de Goiás – UEG

ANÁPOLIS - GO
2025

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UEG
com os dados fornecidos pelo(a) autor(a).

PB327 Pereira Bastos, Cesar Roberto
a Algoritmo Floresta Aleatória e Estatística
Multivariada Utilizados para Classificação do Estádio
de Maturação de Cagaitas (Eugenia Dysenterica) / Cesar
Roberto Pereira Bastos; orientador Francisco Ramos de
Melo. -- Anápolis, 2025.
50 p.

Dissertação (Mestrado - Programa de Pós-Graduação
Mestrado Acadêmico em Engenharia Agrícola) -- Câmpus
Central - Sede: Anápolis - CET - HENRIQUE SANTILLO,
Universidade Estadual de Goiás, 2025.

1. Inteligência artificial. 2. Análise exploratória.
3. Pós-colheita. 4. Cerrado. I. Ramos de Melo,
Francisco, orient. II. Título.

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL (BDTD)

Na qualidade de titular dos direitos de autor, autorizo a Universidade Estadual de Goiás a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UEG), regulamentada pela Resolução, **CsA n.1087/2019** sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

Dados do autor (a)

Nome: Completo: Cesar Roberto Pereira Bastos

E-mail: robertobastos55@gmail.com

Dados do trabalho

Título: Algoritmo Floresta Aleatória e Estatística Multivariada Utilizados para Classificação do Estádio de Maturação de Cagaitas (Eugenia Dysenterica)

Tipo

() Tese (X) Dissertação

Curso/Programa: Programa de Pós-Graduação Stricto Sensu em Engenharia Agrícola

Concorda com a liberação documento ☒ SIM ☐ NÃO¹

Anápolis/GO, 23/06/2025

Local

Data



Documento assinado digitalmente

CESAR ROBERTO PEREIRA BASTOS

Data: 10/09/2025 12:05:05-0300

Verifique em <https://validar.iti.gov.br>

Assinatura do autor



Documento assinado digitalmente

FRANCISCO RAMOS DE MELO

Data: 12/09/2025 14:02:44-0300

Verifique em <https://validar.iti.gov.br>

Assinatura do orientador

¹ Casos de impedimento:

- Período de embargo é de um ano a partir da data de defesa
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da Dissertação/Tese em livro.

RESUMO

O bioma Cerrado, por sua vasta extensão territorial e elevada diversidade biológica, desempenha papel estratégico tanto na segurança alimentar quanto na conservação da biodiversidade no Brasil. Ocupando cerca de 24% do território nacional, esse ecossistema contribui significativamente para a produção agropecuária brasileira. Contudo, a intensificação das intervenções humanas tem comprometido a integridade da biodiversidade regional, tornando essencial o monitoramento e a preservação de suas espécies nativas. A cagaita (*Eugenia dysenterica*), fruto típico e endêmico do Cerrado, destaca-se por seu potencial comercial, mas apresenta desafios relacionados à conservação pós-colheita, devido à textura delicada de sua polpa. Diante disso, o presente estudo teve como objetivo classificar os diferentes estádios de maturação da cagaita com base em atributos físicos, físico-químicos e químicos, visando identificar o ponto ideal de colheita e contribuir para a viabilização de sua comercialização em larga escala. Para isso, foram analisadas variáveis como peso, diâmetro (transversal e longitudinal), área, coloração (luminosidade, °hue e croma), firmeza (casca e polpa), sólidos solúveis, pH, acidez titulável e índice de maturação. Os dados foram obtidos a partir de um delineamento inteiramente casualizado (DIC), com um fator — dias após a antese (DAA) — e cinco tratamentos (28, 31, 34, 37 e 40 DAA), utilizando 40 repetições, com três frutos por unidade experimental. As análises envolveram métodos de estatística multivariada (análise de agrupamentos e análise de componentes principais) e o algoritmo de floresta aleatória, reconhecido por sua capacidade de lidar com classificações complexas. A análise de componentes principais permitiu a identificação das variáveis mais relevantes para a distinção entre os estádios de maturação, enquanto a análise de agrupamentos revelou a formação de três grupos distintos com base nos DAA. O modelo de floresta aleatória apresentou desempenho satisfatório, com métricas equilibradas de precisão, *recall* e *f1-score*, além de um coeficiente de correlação de Matthews (MCC) de 0,62. Esses resultados demonstram o potencial da integração entre estatística multivariada e aprendizado de máquina na caracterização de frutos nativos, oferecendo subsídios para o aprimoramento das práticas de colheita, manejo e processamento na fruticultura do Cerrado.

Palavras-chave: cerrado; pós-colheita; inteligência artificial; análise exploratória.

ABSTRACT

The Cerrado biome, due to its vast territorial extent and high biological diversity, plays a strategic role in both food security and biodiversity conservation in Brazil. Covering approximately 24% of the national territory, this ecosystem contributes significantly to Brazilian agricultural and livestock production. However, increasing human interventions have threatened the region's ecological balance, making the monitoring and preservation of native species essential. *Eugenia dysenterica*, commonly known as cagaita, is an endemic fruit of the Cerrado with commercial potential, yet it poses postharvest conservation challenges due to the delicate texture of its pulp. In this context, the present study aimed to classify the different maturation stages of cagaita based on physical, physicochemical, and chemical attributes, in order to determine the optimal harvest point and support large-scale commercialization. The variables analyzed included weight, diameter (transverse and longitudinal), surface area, color (lightness, hue angle, and chroma), firmness (peel and pulp), soluble solids, pH, titratable acidity, and maturation index. Data were collected using a completely randomized design (CRD) with one factor — days after anthesis (DAA) — comprising five treatments (28, 31, 34, 37, and 40 DAA), with 40 replications and three fruits per experimental unit. Statistical analyses involved multivariate methods (cluster analysis and principal component analysis) and the random forest algorithm, known for its ability to handle complex classification problems. Principal component analysis identified the most relevant variables for differentiating maturation stages, while cluster analysis revealed the formation of three distinct groups based on DAA. The random forest model showed satisfactory performance, with balanced metrics of precision, recall, and F1-score, as well as a Matthews correlation coefficient (MCC) of 0.62. These findings demonstrate the potential of integrating multivariate statistics and machine learning for the characterization of native fruits, providing valuable insights for improving harvest, handling, and processing practices in Cerrado fruit farming.

Keywords: Cerrado; postharvest; artificial intelligence; exploratory analysis.

SUMÁRIO

1	INTRODUÇÃO.....	8
2	REVISÃO BIBLIOGRÁFICA.....	10
2.1	Características e Importância da Cagaita (<i>Eugenia dysenterica</i>).....	10
2.2	Estatística Multivariada na Análise de Dados	12
2.3	Algoritmo Floresta Aleatória e suas Aplicações	14
3	MATERIAL E MÉTODOS.....	17
3.1	Coleta dos dados.....	17
3.2	Área de estudo	18
3.3	Delineamento experimental.....	18
3.4	Estatística multivariada	19
3.4.1	Análise de Componentes Principais – ACP	20
3.4.2	Análise de Agrupamentos – AA.....	21
3.5	Algoritmo floresta aleatória (random forest – RF).....	21
3.5.1	Entradas do modelo de classificação	21
3.5.2	Saídas do modelo de classificação.....	22
3.5.3	Medida de impureza	22
3.5.4	Divisão do conjunto de dados em treinamento e teste.....	22
3.5.5	Avaliação do modelo.....	22
3.5.6	Ferramentas para construção do algoritmo.....	23
4	RESULTADOS E DISCUSSÃO	24
4.1	Análise de Componentes Principais – ACP	24
4.2	Análise de agrupamentos – AA.....	29
4.3	Algoritmo floresta aleatória (random forest – RF).....	33
5	CONSIDERAÇÕES FINAIS	37
	REFERÊNCIAS BIBLIOGRÁFICAS	39
	APÊNDICES	47

1 INTRODUÇÃO

O bioma Cerrado desempenha um papel estratégico na segurança alimentar e na conservação da biodiversidade brasileira, em razão de suas características ecológicas, econômicas e sociais. Ocupando aproximadamente 24% do território nacional, o Cerrado está presente, total ou parcialmente, em 1.389 municípios, distribuídos em 12 estados e no Distrito Federal. Reconhecido como a savana tropical mais biodiversa do mundo, esse bioma contribui significativamente para a produção agrícola nacional, sendo responsável por cerca de 40% das culturas anuais e por parcela expressiva da criação de bovinos, suínos e aves (BOLFE et al., 2020).

A intensa pressão antrópica exercida sobre os recursos naturais do Cerrado tem comprometido a manutenção da biodiversidade local, tornando-se imprescindível o desenvolvimento de estratégias voltadas à conservação e ao uso sustentável de suas espécies endêmicas. Tais estratégias são fundamentais não apenas para a preservação ambiental, mas também para o fortalecimento da economia regional, especialmente por meio da valorização de frutos nativos com potencial comercial (CHO et al., 2021).

Dentre as espécies características desse bioma, destaca-se a cagaita (*Eugenia dysenterica*), planta arbórea da família Myrtaceae, amplamente distribuída nos estados da Bahia, Goiás, Maranhão, Mato Grosso, Mato Grosso do Sul, Minas Gerais, Pará, Piauí, São Paulo, Tocantins e no Distrito Federal. A cagaiteira pode atingir de 10 a 15 metros de altura, apresentando troncos e ramos sinuosos, casca espessa e folhas simples e opostas. Seus frutos são globosos, de coloração amarelo-clara, com polpa fibrosa, epicarpo membranoso e sabor levemente ácido (RODRIGUES-SILVA et al., 2021).

Apesar da reconhecida importância ecológica e econômica da cagaita, observa-se uma lacuna significativa na literatura científica quanto à aplicação integrada de ferramentas multivariadas e de inteligência artificial na classificação de seus estádios de maturação. Essa carência metodológica dificulta o estabelecimento de critérios técnicos para o manejo pós-colheita e a comercialização em larga escala da espécie, comprometendo sua inserção em cadeias produtivas formais.

A qualidade e a conservação pós-colheita desses frutos estão diretamente associadas ao estágio de maturação em que são colhidos. A colheita no momento fisiológico ideal é essencial para garantir frutos com características físico-químicas adequadas ao consumo e à industrialização. No entanto, conforme Carneiro et al. (2015), a textura da polpa da cagaita representa um desafio à sua conservação, dificultando a extensão do período de

comercialização. Isso ocorre devido à elevada atividade metabólica dos frutos, intensificada após a colheita, o que acelera os processos de deterioração.

Neste contexto, este estudo propõe como objetivo central a classificação dos estádios de maturação da cagaita com base em variáveis físicas, físico-químicas e químicas, utilizando técnicas de estatística multivariada (análise de agrupamentos e análise de componentes principais) e o algoritmo de floresta aleatória.

Parte-se da hipótese de que essas ferramentas permitirão identificar padrões multivariados capazes de discriminar os frutos conforme seu grau de maturação. Assim, busca-se responder à seguinte pergunta norteadora: quais variáveis são mais relevantes para a caracterização dos diferentes estádios de maturação da cagaita e como elas podem ser utilizadas para a definição do ponto ideal de colheita?

A literatura científica apresenta diversas aplicações bem-sucedidas dessas técnicas em estudos com frutos nativos. Chaves Neto et al. (2020), por exemplo, utilizaram a análise de agrupamentos para classificar os estádios de maturação da uvaia (*Eugenia pyriformis*), demonstrando a eficácia do método na identificação de agrupamentos hierárquicos baseados em atributos físico-químicos e bioativos. Da mesma forma, Santos et al. (2019) empregaram a análise de componentes principais para avaliar o efeito de métodos de maturação sobre a qualidade de bananas-prata, sendo possível identificar as variáveis mais influentes na diferenciação dos tratamentos. No campo do aprendizado de máquina, Marques (2019) obteve resultados satisfatórios ao utilizar o algoritmo floresta aleatória para a classificação de espécies invasoras em Portugal, evidenciando o potencial da técnica em contextos de classificação.

Diante do exposto, este trabalho teve como objetivo classificar os diferentes estádios de maturação da cagaita por meio da aplicação de estatística multivariada e do algoritmo floresta aleatória, de modo a identificar o ponto ideal de colheita e subsidiar estratégias que favoreçam a comercialização em larga escala dos frutos.

2 REVISÃO BIBLIOGRÁFICA

2.1 Características e Importância da Cagaita (*Eugenia dysenterica*)

A cagaiteira, representada na Figura 1, pertencente à família *Myrtaceae* é uma espécie frutífera nativa do bioma Cerrado, que em razão de suas propriedades laxativas é popularmente conhecida como cagaita. Sendo uma árvore de porte médio, com altura variando entre 4 e 10 metros, tronco cilíndrico e tortuoso com diâmetro entre 20 e 40 centímetros, e casca suberosa, fendida e de aparência bastante característica (MARTINOTTO et al., 2007).

Figura 1 – Árvore da cagaiteira.



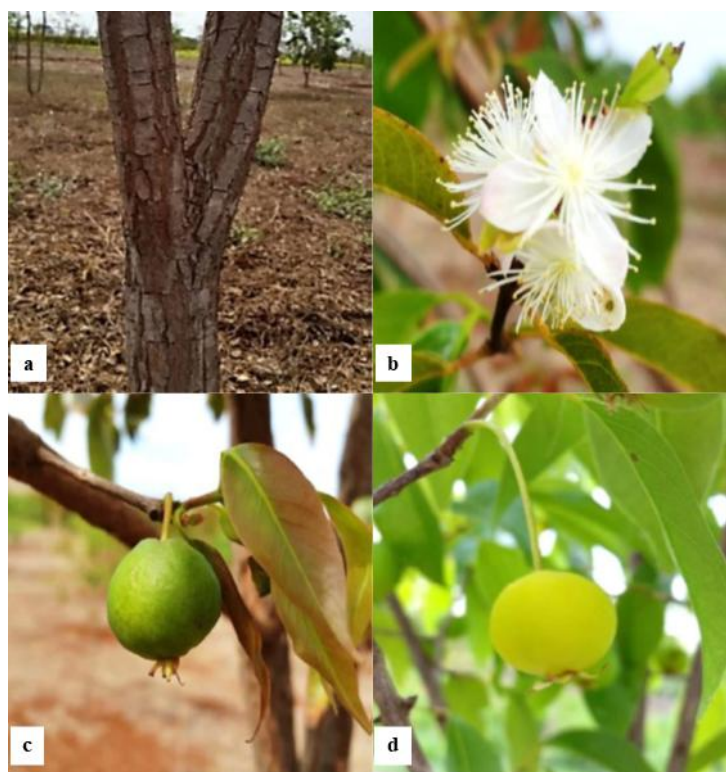
Fonte: Brandizzi (2012) <<https://commons.wikimedia.org/wiki/File:Cagaiteira.JPG>>.

Sua copa é alongada e densa (Figura 1) e adquire tonalidade avermelhada quando há predominância de folhas jovens. As folhas, por sua vez, são reduzidas, membranáceas, simples e opostas-cruzadas (LORENZI, 2002).

De acordo com Embrapa (2025) de uma maneira geral a fase da antese, ou de floração, é considerada a partir do momento em que a maioria das flores estão se abrindo nas inflorescências, passando pela fase de expansão completa, até a fase em que, aparentemente, já ocorreu a liberação do pólen; nesta fase as anteras estão escurecendo e os estames estão com início de murchamento.

A cagaiteira apresenta período da antese entre os meses de agosto e setembro, em geral, a planta apresenta-se totalmente desprovida de folhagem durante esse período (Castro et al., 2013; Reis e Schmiele, 2019), iniciando a frutificação, geralmente, a partir do quarto ou quinto ano de idade (Silva et al., 1992). O intervalo médio entre o início da formação dos frutos e sua maturação é de aproximadamente quatro semanas (Souza et al., 2008). Essas fases fenológicas estão representadas na Figura 2.

Figura 2 – *Eugenia dysenterica*: (a) tronco fissurado, (b) flor, (c) fruto verde com folhas novas e (d) fruto maduro.



Fonte: adaptado de Almeida (2020).

Seus frutos são comestíveis e amplamente apreciados, especialmente para o preparo de sucos, o que justifica seu cultivo frequente em pomares domésticos. Além disso, são avidamente consumidos por diversas espécies de aves, tornando a cagaiteira uma espécie recomendada para reflorestamentos heterogêneos, com foco na recomposição de áreas degradadas de preservação permanente (TERÁN-ORTIZ et al., 2013).

Os frutos são muito frágeis, apresentando casca fina e polpa succulenta, com sabor predominantemente azedo. São amarelos, esféricos e levemente achatados, com massa variando entre 14 e 20 g, comprimento de 3 a 4 cm e diâmetro de 3 a 5 cm, contendo de uma a três sementes por unidade (BUENO et al., 2017).

Rodrigues et al. (2021) identificaram, na polpa da cagaita, compostos como ácidos orgânicos, açúcares, aminoácidos e compostos fenólicos. De forma complementar, Silva et al. (2019) relataram altos teores de compostos fenólicos e níveis intermediários de atividade antioxidante. Esses atributos conferem valor comercial e industrial ao fruto, tornando-o uma promissora matéria-prima para diferentes aplicações.

Do ponto de vista nutricional, a cagaita é considerada uma excelente fonte de vitamina C, com concentração de 24,53 mg por 100 g, valor superior ao observado em frutas convencionalmente cultivadas, como a banana (*Musa* sp.), com 6,4 mg/100 g, e a maçã (*Malus domestica*), com 5,9 mg/100 g (LEMOS-FILHO, 2000).

O aprofundamento dos estudos sobre *Eugenia dysenterica* é fundamental para a compreensão abrangente de suas potenciais aplicações. Entre os anos de 2005 e 2010, houve um avanço significativo nas investigações sobre a composição química do óleo essencial extraído das folhas e dos frutos em diferentes estádios de maturação, sob variadas condições sazonais e ambientais (VIEIRA et al, 2023).

Paralelamente, foram avaliadas as atividades antioxidantes *in vitro*, por meio do método DPPH, de extratos obtidos da polpa, semente e casca dos frutos de *E. dysenterica* (Karadag et al., 2009). Além disso, outros estudos identificaram a presença de metabólitos secundários, como compostos fenólicos e carotenoides, os quais desempenham um papel relevante na proteção das estruturas biológicas, atuando como sequestradores de radicais livres (WANG et al., 2009).

A utilização da antese como marco inicial na avaliação de plantas, especialmente no contexto da produção agrícola, é fundamental, pois essa fase marca o início do processo reprodutivo das plantas, o que é imprescindível para a formação de frutos e sementes, caracterizando a produção final (SIMPSON, 2010).

2.2 Estatística Multivariada na Análise de Dados

A estatística multivariada, também denominada técnicas multivariadas exploratórias, pode ser aplicada em praticamente todos os campos do conhecimento nos quais o pesquisador busca compreender as relações entre múltiplas variáveis de um determinado banco de dados. Essa abordagem é caracterizada pela ausência de intenção de construção de modelos

confirmatórios, ou seja, não se objetiva realizar inferências que extrapolem as observações analisadas, tampouco estimar modelos ou equações voltadas à previsão do comportamento dos dados (FÁVERO, 2015).

Essa característica é fundamental para distinguir as técnicas exploratórias abordadas neste contexto daquelas classificadas como técnicas de dependência, como é o caso dos modelos de regressão simples e múltipla, modelos de regressão logística binária e multinomial, modelos para dados de contagem, modelos de regressão para dados em painel, entre outros (FÁVERO, 2015).

A análise multivariada abrange todas as técnicas estatísticas que realizam, de forma simultânea, a análise de múltiplas medidas obtidas de indivíduos ou objetos sob investigação. Dessa forma, qualquer análise que envolva mais de duas variáveis simultaneamente pode ser classificada, em princípio, como multivariada (HAIR et al., 2009).

No entanto, para que uma análise seja considerada verdadeiramente multivariada, é necessário que todas as variáveis envolvidas sejam aleatórias e inter-relacionadas, de modo que os efeitos individuais de cada uma não possam ser interpretados de forma isolada e significativa (HAIR et al., 2009).

Conquanto a estatística multivariada tenha surgido por volta de 1901, foi apenas nas últimas décadas que seu desenvolvimento e aplicação se tornaram significativamente mais acessíveis, impulsionados pelos avanços tecnológicos. Entre os principais fatores que contribuíram para essa evolução destacam-se os progressos em hardware e software, que permitiram o processamento eficiente de grandes volumes de dados (HAIR et al., 2009; VICINI et al., 2018).

Sem o auxílio de programas computacionais especializados, seria inviável produzir muitos dos gráficos utilizados nessas análises, bem como realizar, com agilidade e clareza, a interpretação dos resultados. A popularização de pacotes estatísticos com interfaces amigáveis introduziu a estatística multivariada à chamada "era do aponte-e-clique", possibilitando que mesmo grandes e complexos conjuntos de dados sejam analisados de forma rápida e acessível (HAIR et al., 2009; VICINI et al., 2018).

A caracterização físico-química e funcional de frutos tem sido amplamente abordada por meio de ferramentas estatísticas multivariadas, que permitem a identificação de padrões complexos e a classificação simultânea de variáveis e amostras. Esse tipo de abordagem tem se mostrado particularmente eficaz na análise de frutos oriundos de populações nativas ou de ocorrência espontânea, os quais apresentam elevada variabilidade entre indivíduos. Métodos como a análise de agrupamentos e a análise de componentes principais, têm sido empregados

com êxito na distinção de grupos com base em características físicas, físico-químicas e químicas, fornecendo subsídios relevantes para a compreensão dos processos de maturação e para a valorização de espécies frutíferas nativas (BASTOS, 2018).

Nesse contexto, Andrade et al. (2017) realizaram um experimento com o objetivo de reduzir a dimensionalidade e identificar, entre as variáveis físicas e químicas dos frutos do meloeiro, aquelas mais relevantes para a caracterização da produção e da qualidade do híbrido *Yellow King*. Por meio da aplicação da análise de componentes principais e da análise fatorial, os autores verificaram que as onze variáveis analisadas, podiam ser reduzidas a dois componentes principais, CP1, referindo-se principalmente a questões da caracterização “sabor de frutos” e CP2, referindo-se a questões da caracterização “física de frutos”. Esses descritores foram considerados eficientes na distinção dos dados estudados, permitindo a formação de grupos distintos e fornecendo subsídios importantes para avaliação dos tratamentos aplicados durante o experimento.

Chaves Neto et al. (2020) realizaram um estudo com o objetivo de caracterizar os atributos físicos e físico-químicos, o teor de compostos bioativos e a atividade antioxidante de frutos de uvaieira (*Eugenia pyriformis* Cambess) provenientes de diferentes plantas de ocorrência espontânea no município de Cuité de Mamanguape, Paraíba, Brasil. Para a análise e interpretação, os autores submeteram os dados à análise multivariada, com agrupamento *two-way* utilizando o método de *Ward*, com base no critério cúbico de agrupamento, com o intuito de classificar simultaneamente as variáveis dependentes e independentes.

Os resultados indicaram que, à medida que os frutos da uvaieira avançam em seu processo de maturação, ocorrem aumentos nos valores de diâmetro, comprimento, massa fresca e teor de sólidos solúveis, acompanhados por uma redução nos níveis de acidez. Observou-se também um incremento na capacidade antioxidante total ao longo da maturação, o qual está associado ao aumento nos teores de carotenoides e flavonoides – compostos que representam os principais responsáveis pelas transformações observadas nos estádios finais de maturação do fruto (CHAVES NETO et al., 2020).

2.3 Algoritmo Floresta Aleatória e suas Aplicações

O *ensemble learning* refere-se a uma abordagem de aprendizado de máquina baseada na combinação de múltiplos modelos preditivos, com o objetivo de construir um modelo final mais robusto e preciso. Um dos métodos mais difundidos dentro dessa categoria é o *Bagging* (*Bootstrap Aggregating*), apresentado por Breiman (1996).

Nesse método, cada modelo individual do conjunto é treinado com um subconjunto de dados de treinamento, obtido por amostragem aleatória com reposição a partir do conjunto original, contendo N observações. Assim, diferentes amostras são utilizadas para gerar diversidade entre os modelos. A combinação dos resultados é realizada por meio de votação majoritária, no caso de tarefas de classificação, ou por meio de média aritmética, quando se trata de regressão (SILVA FILHO et al., 2021).

Floresta Aleatória ou *Random Forest* é um algoritmo *ensemble learning* de aprendizado de máquina supervisionado, amplamente utilizado para tarefas de classificação e regressão, com base em conjuntos de dados previamente coletados. Por exemplo, esse algoritmo pode ser empregado para classificar diferentes espécies animais com base em características como tamanho, peso e aparência, ou ainda para categorizar doenças a partir de sintomas apresentados por um paciente (HARTSHORN, 2016).

Proposto por Breiman (2001), o método combina os princípios do algoritmo *Bagging*, introduzido anteriormente pelo próprio autor, com o *Random Subspace Method*, descrito por Ho (1998). O objetivo principal dessa combinação é reduzir a inter-relação entre os classificadores individuais, treinando-os em diferentes subconjuntos aleatórios de dados e atributos. A base do modelo é a árvore de decisão, que representa uma estrutura hierárquica de decisões análoga a uma tabela de decisão.

O algoritmo constrói diversas árvores de decisão de forma aleatória, originando uma "floresta", na qual cada árvore contribui para a decisão final por meio de um processo de votação (no caso de classificação) ou de média (para regressão), conforme discutido por Sammut e Webb (2011).

A variabilidade necessária para garantir a robustez do modelo é introduzida de duas formas principais: a) cada árvore é treinada com um subconjunto de instâncias da base de dados, selecionado por amostragem aleatória com reposição (método *bootstrap*), e; b) em cada nó interno da árvore, é selecionado aleatoriamente um subconjunto de atributos, limitando o processo de divisão apenas a essas variáveis, o que contribui para a diversificação entre as árvores (RIQUETI et al., 2018).

De-la-Torre et al. (2019) desenvolveram um sistema de classificação automática baseado na captura de imagens de frutos de camapu, com o objetivo de identificar seu estágio de maturação. Para isso, os autores integraram técnicas estatísticas multivariadas com algoritmos de aprendizado de máquina, alcançando uma média da métrica *F-measure* de 71,75. Sendo considerado um bom desempenho e pode ser atribuído tanto à organização das amostras

no espaço de características quanto à eficácia da otimização bayesiana empregada pelos classificadores na definição de hiperplanos de separação apropriados.

Meneses et al. (2024) realizaram um estudo para avaliar os estádios de maturidade fisiológica do café arábica com modelos de aprendizado de máquina (ML), dentre eles o algoritmo floresta aleatória (RF), com o objetivo de determinar a força de desprendimento dos grãos nas cultivares sob condições de irrigação por gotejamento e sequeiro. O modelo RF foi particularmente adequado para cultivares irrigadas por gotejamento, com um coeficiente de correlação de Matthews (MCC) de 0,75. A implementação desses modelos ML para classificação da força de desprendimento tem sido benéfica, melhorando a tomada de decisões em sistemas de colheita mecanizada.

3 MATERIAL E MÉTODOS

3.1 Coleta dos dados

Os dados utilizados neste trabalho foram cedidos por Campos (2021). As metodologias utilizadas na coleta dos dados são apresentadas na Tabela 1.

Tabela 1 – Metodologias utilizadas na análise das cagaitas para obtenção das características físicas, físico-químicas e químicas.

Variáveis Analisadas	Metodologias
Massa Fresca	Obtidas por meio de uma balança de alta precisão fabricada pela marca Shimadzu, modelo BL3200H, valores em gramas (g).
Diâmetro (transversal e longitudinal)	Diâmetro longitudinal foi avaliado na direção horizontal do fruto e o diâmetro transversal (DT) foi medido desde a área de fixação do pedúnculo até sua área oposta, seguindo uma orientação vertical, com paquímetro digital Stainless Hardened de 150 mm, resultados em milímetros (mm).
Área	Área = $4 \cdot \pi \cdot r^2$, em que: $\pi = 3,14$ e r = raio.
Coloração (Luminosidade, °Hue e Croma)	Luminosidade obtida com colorímetro da marca CR-400 (Konica Minolta); $^{\circ}\text{Hue} = 180 + \arctang\left(\frac{b^*}{a^*}\right)$; $\text{Croma} = \sqrt{a^{*2} + b^{*2}}$; Em que: a^* = coordenada cromática obtida no colorímetro CR-400; b^* = coordenada cromática obtida no colorímetro CR-400.
Firmeza (casca e polpa)	Obtidas com aparelho Texture Analyser CT3 da marca Brookfield, com uma velocidade de penetração de 7,0 mm s ⁻¹ a uma profundidade de 2 mm, resultados em Newton (N).
Sólidos solúveis	Cada fruto foi submetido à maceração utilizando um almofariz e pistilo. A leitura dos sólidos solúveis no refratômetro mini digital Reichert, em °Brix, utilizando uma pipeta, conforme recomendado pelo AOAC (2012).
Potencial hidrogeniônico	Determinado utilizando o pHmetro portátil modelo K39 - 0014P da marca Kasvi, de acordo com AOAC (2012).
Acidez titulável	Foram pesados 5 g de cada fruto macerado e adicionados 95 ml de água destilada, para homogeneização, depois foi utilizado NaOH 0,1 mol L ⁻¹ para determinar a porcentagem de ácido cítrico. O ponto de viragem foi identificado visualmente pela alteração de cor de cada amostra, conforme recomendado pelo AOAC (2012).
Índice de maturação	Foi calculado por intermédio da correlação entre o teor de sólidos solúveis e a acidez titulável.

Fonte: Adaptado de Campos (2021).

3.2 Área de estudo

As cagaitas foram colhidas na área experimental da Escola de Agronomia, Engenharia de Alimentos e Engenharia Florestal, da Universidade Federal de Goiás, localizada no município de Goiânia/GO, situado em latitude 16° 36' 16.920" Sul, longitude 49° 15' 44.280" Oeste e altitude de 749 metros. Conforme os tipos de clima da classificação Koppen-Geiger, para Goiás, o clima característico é Aw – Clima tropical de altitude, com chuva de verão e mês mais frio com temperatura média superior a 18°C (NASCIMENTO et al., 2016).

Subsequentemente, os frutos foram transportados ao Laboratório de Pós-Colheita do curso de Engenharia Agrícola da Universidade Estadual de Goiás, Campus Central Henrique Santillo, situado no município de Anápolis/GO, onde foram realizadas as análises para a obtenção das características físicas, físico-químicas e químicas das cagaitas.

3.3 Delineamento experimental

O delineamento experimental empregado por Campos (2021) foi o inteiramente casualizado (DIC), com um fator, denominado dias após antese (abertura da flor), utilizando 40 repetições por tratamento, sendo 3 frutos diferentes por repetição. Os frutos foram coletados a partir do 28 DAA, com o intervalo de três dias, complementando 5 tratamentos (28, 31, 34, 37 e 40 DAA) e foram estudados no mesmo dia da coleta.

Foram efetuadas as análises físicas, físico-químicas e químicas, sendo elas: massa fresca (MF), diâmetro (transversal (DT) e longitudinal (DL)), área (A), coloração (luminosidade (L), °HUE e CROMA), firmeza (casca (FC) e polpa (FP)), sólidos solúveis (SS), potencial hidrogeniônico (PH), acidez titulável (AT) e índice de maturação (IM) (CAMPOS, 2021).

Diferentemente do estudo conduzido por Campos (2021), que empregou o método de agrupamento de *Ward* associado à distância de *Canberra* como medida de dissimilaridade, a presente pesquisa adotou o algoritmo de agrupamento *Complete Linkage* (método do vizinho mais distante), utilizando a distância Euclidiana como índice de dissimilaridade. Tal escolha se deu com base nos testes realizados no *software PAST* (Hammer et al., 2001), nos quais essa combinação apresentou o melhor coeficiente de correlação cofenética (CCC).

No que se refere à Análise de Componentes Principais (ACP), este trabalho também se distingue do de Campos (2021) pela adoção dos gráficos de carregamentos para a seleção das variáveis mais representativas em cada componente principal, bem como pelo uso de critérios baseados nos vetores do diagrama de dispersão para o refinamento das variáveis utilizadas no

algoritmo de aprendizado de máquina Floresta Aleatória, o qual difere dos algoritmos empregados por Campos (2021).

3.4 Estatística multivariada

Para correta aplicação das técnicas multivariadas exploratórias é necessário que exista uma estrutura de inter-relação entre as variáveis, caso não seja observada ligação entre estas, a técnica indicada para análise é a univariada (VICINI et al., 2018).

A fim de facilitar a análise, realizou-se a redução dos dados por meio do cálculo da média aritmética simples nos resultados dos 3 frutos analisados em cada uma das 13 variáveis presentes neste estudo, resultando em um conjunto de dados com 2.600 observações.

Tendo em vista que os dados, utilizados nos cálculos, não possuem unidades de medida iguais, os mesmos foram normalizados por Min-Max, conforme Equação (1) adaptada de HAN; PEI; TONG (2022), para terem médias zero e desvios-padrão unitários, e assim evitar erros nos resultados (MANLY e ALBERTO, 2019).

$$v'_i = \frac{v_i - \min_A}{\max_A - \min_A} \quad (1)$$

Em que:

v_i = valor original

$\min_A$ = valor mínimo da variável

$\max_A$ = valor máximo da variável

A forma como os dados são normalizados é um fator relevante para o bom resultado da técnica utilizada para classificação, mostrando-se mais eficazes os classificadores aplicados aos dados normalizados por Min-Max do que quando aplicados a dados normalizados pela normalização *Standard* (ANTÓNIO, 2019).

Com a finalidade de testar a medida de adequação dos dados para se ter uma confirmação de que as variáveis estão bem ligadas foi realizado o teste *Kaiser-Meyer-Olkin Measure of Adequacy* (KMO), neste teste um resultado igual ou superior a 0,7, indica que a análise realizada pode ser considerada própria, revelando que as variáveis estão inter-relacionadas (FORNARI, 2019).

Para identificar se as variáveis seguem uma distribuição normal multivariada foi executado o teste *Bartlett Test of Sphericity* (BTS), o valor de α utilizado será de 5%, a hipótese nula poderá ser rejeitada caso o valor encontrado seja inferior ao valor de α (HAIR et al., 2009).

Para realização das análises de estatística multivariada, utilizou-se o *software Past* versão 5.1 para *Windows* (HAMMER et al., 2001).

3.4.1 Análise de Componentes Principais – ACP

A aplicação da técnica de ACP torna-se viável ao identificar que as variáveis estão inter-relacionadas, positiva ou negativamente. Nesta situação é concebível que as 13 variáveis originais analisadas neste estudo, podem ser adequadamente representadas por duas ou três componentes principais (MANLY e ALBERTO, 2019).

A ACP foi executada a fim de compreender as interações entre as variáveis, bem como quais são as mais relevantes para a classificação do estágio de maturação das cagaitas. Essa técnica reduz o espaço dimensional inicial, que é composto pelo número de variáveis envolvidas, representando-as por combinações lineares, as quais são chamadas de “componentes”. Esses componentes têm a função de representar a maior parte dos dados, para que seja o mais fiel possível a eles (ROCHA et al., 2020).

Para obter os componentes principais, mediante os dados contidos nas variáveis envolvidas no estudo, que representam as características físicas, físico-químicas e químicas da cagaita com o estágio de maturação, foram calculados a matriz de correlação, os autovalores e os autovetores e, por fim, as combinações lineares que constituíram as componentes principais (VICINI et al., 2018).

Com o intuito de facilitar a visualização dos componentes principais mais relevantes para o estudo, estes foram apresentados em forma de tabela. Para a definição do número de fatores a serem listados, adotou-se o critério proposto por Kaiser (1958), o qual estabelece que apenas os componentes principais com autovalores superiores a 1 devem ser considerados significativos. Assim, os componentes com autovalores inferiores a esse limite foram considerados insignificantes e, portanto, desconsiderados na interpretação dos resultados.

A decisão sobre o número de componentes a serem mantidos, foi tomada baseada no critério de percentagem de variância, o qual se caracteriza pela obtenção de um percentual cumulativo especificado da variância total extraída por componentes sucessivos, no caso deste trabalho até 65%, pois buscasse reduzir ao máximo o número de variáveis a serem utilizadas na classificação com o algoritmo floresta aleatória, aliado à menor perda possível de informação (VICINI et al., 2018).

3.4.2 Análise de Agrupamentos – AA

Por ser a AA uma técnica exploratória de análise multivariada de dados, esta permitiu classificar os tratamentos do estudo em classes homogêneas de acordo com as similaridades existentes entre eles (LOESCH e HOELTGEBAUM, 2012).

Essa análise foi aplicada utilizando-se o método aglomerativo hierárquico. Os grupos (também conhecidos como classes ou *clusters*) foram construídos devido a um determinado nível de proximidade entre os tratamentos comparados.

O algoritmo de agrupamento dos dados utilizado foi o *Complete linkage* (método do vizinho mais distante) com a distância euclidiana, um coeficiente que quantifica o quão parecidos dois objetos são por meio da dissimilaridade entre eles, pois quanto menor for o valor observado mais próximos os objetos estão uns dos outros (REGAZZI, 2001 apud VICINI, 2006). As técnicas hierárquicas produzem um gráfico chamado dendograma, o qual será utilizado para fazer as inferências sobre o estudo (MANLY e ALBERTO, 2019). A eficiência do dendograma será determinada pelo coeficiente de correlação cofenética, que verificará a proximidade das duas matrizes (matriz de similaridade e matriz cofenética), sendo que um resultado à partir de 0,8 poderá ser considerado um bom ajuste (BUSSAB et al., 1990).

3.5 Algoritmo floresta aleatória (random forest – RF)

No presente trabalho, o foco é o aprendizado supervisionado, que utiliza o conjunto de dados para que o algoritmo possa aprender e identificar os fatores significantes para o estágio de maturação da cagaita, construindo um modelo de classificação que permitiu a identificação das cagaitas com os dias após antese (BITENCOURT et al., 2021).

Antes de utilizar os dados no algoritmo RF para realizar a classificação, foram seguidas as seguintes etapas de tratamento dos dados (RIBEIRO e ZÁRATE, 2016; BITENCOURT et al., 2021):

- Remoção ou substituição de quaisquer dados ausentes, inconsistentes ou irrelevantes que possam prejudicar o desempenho do modelo, e;
- Seleção das características relevantes e descarte das que não contribuem significativamente para a classificação.

3.5.1 Entradas do modelo de classificação

As entradas do modelo RF, consistiram nos dados das variáveis que representam as características físicas, físico-químicas e químicas das cagaitas, selecionadas após a aplicação

da ACP, bem como na classe de estágio de maturação associada a cada fruto, neste trabalho descrita pelos DAA, selecionados após aplicação da AA. Esses dados são essenciais para treinar e testar o modelo e fazer agrupamentos precisos (RAIMUNDO, 2021).

3.5.2 Saídas do modelo de classificação

As saídas do modelo foram as classificações previstas do estágio de maturação das cagaitas com base nas características fornecidas como entrada. O modelo atribuiu uma classe de estágio de maturação a cada fruto, indicando em qual estágio de maturação ele se encontra (LACERDA, 2023).

3.5.3 Medida de impureza

Como medida de impureza usada no processo de construção das árvores de decisão dentro do RF, neste trabalho foi utilizado o índice de Gini, por ele ser um dos mais utilizados e também por ser o critério padrão em problemas de classificação e na linguagem de programação *Python*. Ele é usado para determinar a qualidade da divisão dos dados em cada nó da árvore, com o objetivo de criar divisões que maximizem a homogeneidade das classes (HARTSHORN, 2016; SANTANA, 2020; MARTELETTO, 2022).

3.5.4 Divisão do conjunto de dados em treinamento e teste

Os dados foram divididos em dois conjuntos: conjunto de treinamento com 70% dos dados e o conjunto de teste com 30%. O conjunto de treinamento foi usado para treinar o modelo RF, enquanto o conjunto de teste foi usado para avaliar o desempenho do modelo após o treinamento (VASCONCELOS et al., 2019; BITENCOURT et al., 2021; OLIVEIRA et al., 2021).

3.5.5 Avaliação do modelo

A análise avaliativa do modelo proposto pelo RF, após utilização dos dados de treinamento e teste, utilizou a matriz de confusão, que registra a quantidade de Verdadeiros Positivos (VP, quando os frutos são classificados corretamente pelo modelo, exibidos na diagonal principal), Verdadeiros Negativos (VN, quando os frutos são classificados verdadeiramente reprovados para uma determinada classe), Falsos Positivos (FP, casos em que o fruto é classificado em uma determinada classe, mas na verdade, pertence a outra) e Falsos Negativos (FN, é quando o modelo classifica o fruto como não pertencente a uma determinada classe, quando na realidade ele pertence à mesma). Partindo dessas quatro medidas, serão

calculados os parâmetros estatísticos: precisão, *recall*, *f1-score* e coeficiente de correlação de Matthews (MCC) (MATTHEWS, 1975; BITENCOURT et al., 2021; CAMPOS, 2021).

3.5.6 Ferramentas para construção do algoritmo

Para construção do algoritmo foram utilizadas algumas bibliotecas e ferramentas em linguagem de programação *Python*, descritas abaixo (HARTSHORN, 2016):

- *Pandas*: biblioteca de análise de dados que fornece estruturas de dados flexíveis e eficientes, para manipular e analisar dados;
- *Scikit-learn*: uma das bibliotecas mais populares para aprendizado de máquina em *Python*;
- *Matplotlib* e *Seaborn*: bibliotecas para visualização de dados em *Python*, e;
- *Google Colab*: utilizado para construção de algoritmos.

4 RESULTADOS E DISCUSSÃO

4.1 Análise de Componentes Principais – ACP

Os resultados obtidos por meio da ACP aplicados às características físicas, físico-químicas e químicas em diferentes dias após antese são apresentados na Tabela 1. O teste KMO que confirma a existência previa de inter-relação entre as variáveis, mostrou o valor de 0,77 como resultado, indicando que as variáveis possuem boa inter-relação.

Já o *p-value* do teste BTS foi de $1,82 \times 10^{-8}$. Uma vez que o valor de α utilizado neste estudo foi de 5%, a hipótese nula pode ser rejeitada, tendo em vista que o valor encontrado foi menor que o valor de α , significando que a ACP será capaz de fornecer uma redução útil da dimensionalidade.

Cada variável em estudo (13 variáveis) gerou um componente principal – CP. No entanto, na Tabela 2 são apresentados apenas 4, tendo em vista que os demais apresentaram autovalores abaixo de 1 (Kaiser, 1958), o que significa que esses componentes (5 ao 13) fornecem pouca informação para explicar o estágio de maturação baseado nas variáveis em diferentes dias após antese.

Também são apresentadas as variâncias explicadas e as variâncias totais. Seguindo o critério proposto por Vicini et al. (2018), considera-se aceitável, em estudos com maior controle experimental, como este, a utilização das componentes principais que, em conjunto, sejam capazes de sintetizar uma variância acumulada em torno de 65%. Os dois primeiros componentes são os que explicam a maior variação total acumulada no conjunto dos dados. Onde CP1 explica 44,49% e CP2 21,89%, representando um total de 66,38% da variabilidade dos dados.

Tabela 2 - Componentes principais, autovalores, variância explicada (%) e variância total (%).

Componente Principal	Autovalor	Variância Explicada (%)	Variância Total (%)
CP1	5,78	44,49	44,49
CP2	2,85	21,89	66,38
CP3	1,25	9,62	76,00
CP4	1,12	8,63	84,63

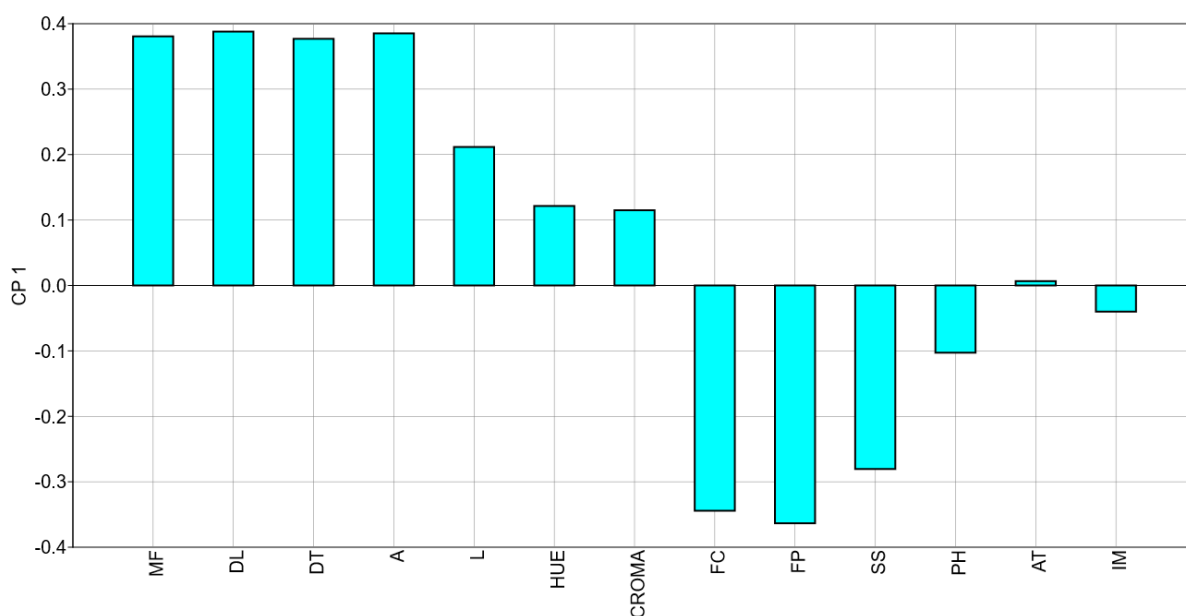
Fonte: Próprio autor (2025).

Nas Figuras 3 e 4 são expostas as inter-relações das variáveis físicas, físico-químicas e químicas das cagaitas, com os dois primeiros componentes principais (CP1 e CP2), nos quais é

possível verificar que a maioria das variáveis apresentaram as melhores relações mútuas, dados os altos valores de variância explicada.

Na CP1, seguindo o critério definido por HAIR et al. (2009), que considera cargas fatoriais entre $\pm 0,3$ e $\pm 0,4$ como atendendo o nível mínimo para interpretação de estrutura de correspondência, o ponto de corte foi estabelecido em 0,3 para seleção de variáveis com forte interdependência positiva e -0,3 para forte interdependência negativa, há interconexão em 6 das 13 variáveis em estudo, em que se correlaciona positivamente MF, DL, DT e A, e negativamente FC e FP, isso quer dizer que quanto maiores e mais pesados os frutos se encontram, menos consistentes estarão.

Figura 3 – Gráfico de carregamentos da inter-relação entre as características físicas, físico-químicas e químicas das cagaitas em diferentes estádios de maturação e o componente principal 1. Em que: MF = Massa Fresca, DL = Diâmetro Longitudinal, DT = Diâmetro Transversal, A = Área, L = Luminosidade, FC = Firmeza da Casca, FP = Firmeza da Polpa, SS = Sólidos Solúveis, PH = Potencial hidrogeniônico, AT = Acidez Titulável; e IM = Índice de Maturação.



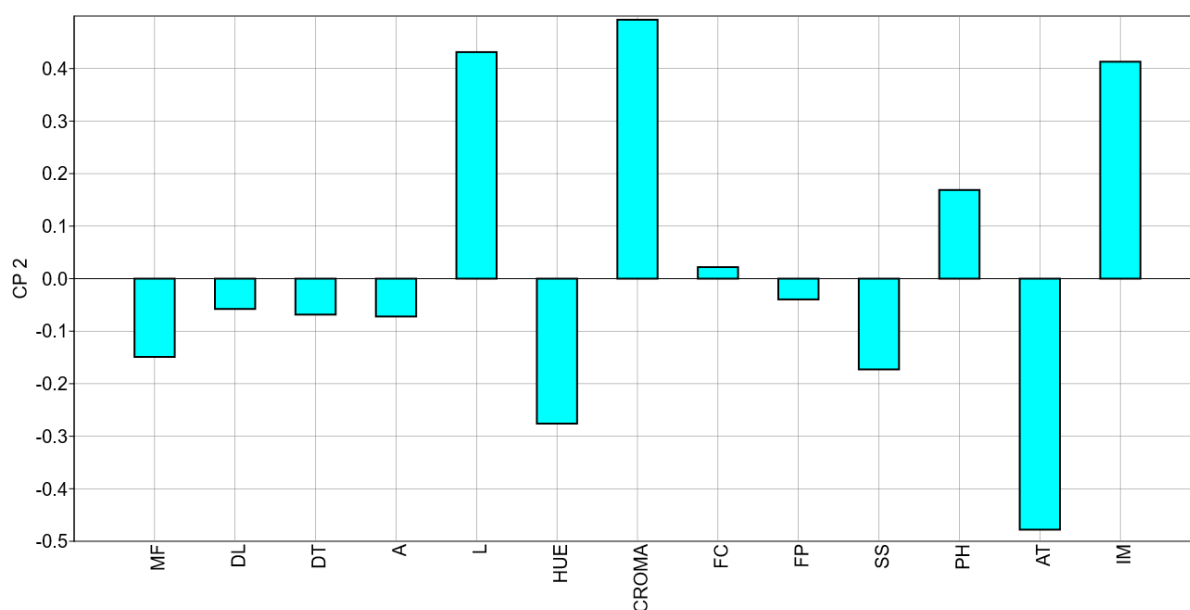
Fonte: Próprio autor (2025).

Na CP2, seguindo o mesmo critério de ponto de corte, mencionado anteriormente, para seleção das variáveis, ocorre inter-relação em 4 das variáveis em estudo, sendo elas positivamente L, CROMA e IM, e negativamente AT. A redução da acidez da cagaita está relacionada com a respiração do fruto, isso quer dizer que quanto mais maduros os frutos se encontram, menor é a acidez dos mesmos (TERÁNORTIZ et al., 2013), corroborando com este trabalho, em que no decorrer dos dias analisados, os frutos apresentaram redução da acidez.

O índice de maturação, de acordo com Lattuada et al. (2020), é um dos principais indicativos para avaliar se os frutos estão maduros e também, um dos principais parâmetros para se avaliar o flavor. Quanto maior essa relação se encontra, mais maduros os frutos estão. Foi possível observar que nesse trabalho ocorreu esse aumento, indicando que os frutos se apresentaram mais maduros ao longo dos diferentes estádios de maturação.

Quanto a coloração, os frutos ao longo dos DAA apresentaram aumento no brilho e as cores mais vividas. A coloração da casca dos frutos pode apresentar alterações ao longo da maturação, sendo que este fato se deve a degradação da clorofila ou revelação de pigmentos, como carotenoides e antocianinas (TUCKER,1993).

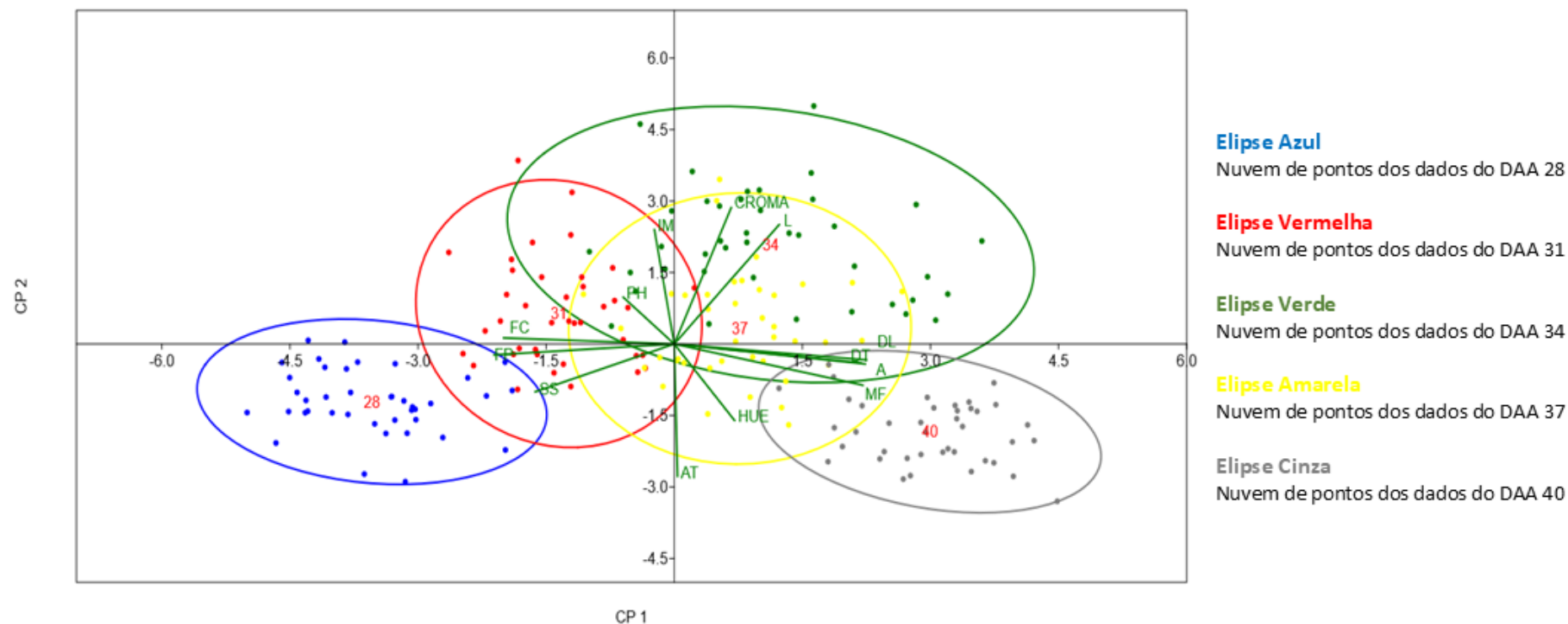
Figura 4 – Gráfico de carregamentos da inter-relação entre as características físicas, físico-químicas e químicas das cagaitas em diferentes estádios de maturação e o componente principal 2. Em que: MF = Massa Fresca, DL = Diâmetro Longitudinal, DT = Diâmetro Transversal, A = Área, L = Luminosidade, FC = Firmeza da Casca, FP = Firmeza da Polpa, SS = Sólidos Solúveis, PH = Potencial hidrogeniônico, AT = Acidez Titulável; e IM = Índice de Maturação.



Fonte: Próprio autor (2025).

Uma vez conhecidas as variáveis de maior importância para os CP1 (MF, DL, DT, A, FC e FP) e CP2 (L, CROMA, IM e AT), as mesmas serão analisadas a seguir com base na Figura 5, no intuito de realizar mais uma etapa de “filtragem” das variáveis, com o objetivo de obter resultados satisfatórios de classificação no RF com gasto computacional mínimo.

Figura 5 – Diagrama de dispersão da inter-relação entre os dias após antese, as características físicas, físico-químicas e químicas das cagaitas e os componentes principais 1 e 2. Em que: DAA = Dias Após Antese, L = Luminosidade, DL = Diâmetro Longitudinal, DT = Diâmetro Transversal, A = Área, MF = Massa Fresca, AT = Acidez Titulável, SS = Sólidos Solúveis, FP = Firmeza da Polpa, FC = Firmeza da Casca, PH = Potencial hidrogeniônico e IM = Índice de Maturação.



Fonte: Próprio autor (2025).

O comprimento do vetor (seta) está diretamente relacionado à sua importância dentro do componente em que está inserido, de modo que vetores mais longos indicam maior relevância. As variáveis situadas ao longo do eixo horizontal possuem maior influência sobre o CP1, enquanto aquelas localizadas no eixo vertical exercem maior impacto sobre o CP2. Já as variáveis que formam um ângulo de aproximadamente 45° em relação a esses eixos apresentam contribuição significativa para ambos os componentes simultaneamente.

Para determinar quais dentre as variáveis selecionadas seguirão para a etapa de classificação com o RF o primeiro critério foi o tamanho do vetor e o segundo, que funcionou como critério de desempate, foi a proximidade do vetor com o eixo do seu respectivo componente. Levando em consideração também a manutenção do mesmo número de inter-relações positivas e negativas, ou seja, ao selecionar um vetor localizado na região positiva do plano cartesiano, um vetor correspondente na região negativa também será incluído, garantindo uma distribuição simétrica das variáveis em relação aos componentes analisados.

O eixo horizontal (CP1), representa um gradiente de mudança estrutural e físico-química do fruto, com características físicas (MF, DL, DT, A) em um extremo e firmeza (FC, FP) no outro.

Para o CP1, observamos que FC e FP, que foram as únicas classificadas anteriormente como variáveis com forte inter-relação negativa, estão bem próximas ao eixo horizontal, dessa forma selecionaremos duas variáveis na região positiva do plano cartesiano, na qual podemos observar que DL apresenta o maior vetor e está perfeitamente alinhada ao eixo horizontal, sendo a primeira escolhida.

No caso da escolha da segunda variável com forte inter-relação positiva, DT, A e MF, aparecem aparentemente empatadas no critério de tamanho do vetor, sendo selecionada DT, por meio do critério de desempate que analisa a proximidade do vetor com o eixo do CP1.

O eixo vertical (CP2), está associado a mudanças químicas e coloração, diferenciando frutos com maior acidez (AT) dos que apresentam maior intensidade de cor (CROMA, L).

Para o CP2, observa-se que a variável AT é a única previamente classificada com uma forte inter-relação negativa. Além disso, ela se apresenta nitidamente alinhada com o eixo vertical. Dessa forma, para manter o equilíbrio na seleção das variáveis, escolhe-se uma variável localizada na região positiva do plano cartesiano. Nessa região, CROMA e L apresentam os vetores de maior magnitude, sendo CROMA a variável selecionada devido à sua maior proximidade em relação ao eixo do CP2.

A Figura 5 também exibe diferentes grupos coloridos com elipses, sugerindo uma possível segmentação dos dados.

Grupo azul (DAA 28), possui indivíduos com valores baixos no CP1, indicando frutos menos desenvolvidos fisicamente e com maior firmeza (FC, FP). No CP2 seus valores aparecem próximos de zero ou negativos o que sugere que esses frutos podem ter menor intensidade de cor e acidez elevada. Representando dessa forma frutos em estágio inicial de maturação.

O grupo vermelho (DAA 31), demonstra frutos que ainda apresentam características físicas iniciais, mas já começam a perder firmeza. O CP2 positivo indica mudança na coloração e menor acidez, sugerindo um estágio de transição para maturação intermediária.

Diferentemente dos grupos citados anteriormente, os grupos verde (DAA 34) e amarelo (DAA 37), estão intimamente ligados, uma vez que suas elipses ocupam boa parte uma da outra, com a maior parte dos seus pontos próximos e ocupando a parte superior direita do gráfico. Esses grupos trazem frutos localizados entre o ponto médio das características e totalmente desenvolvidos fisicamente, com maiores valores de MF, DL, DT e A, ou seja, um estágio intermediário de maturação com equilíbrio entre firmeza, cor e acidez. São frutos maduros e prontos para consumo ou processamento e tendem a se deslocar mais ainda para a direita do gráfico à medida que amadurecem ainda mais (SILVA, 2016).

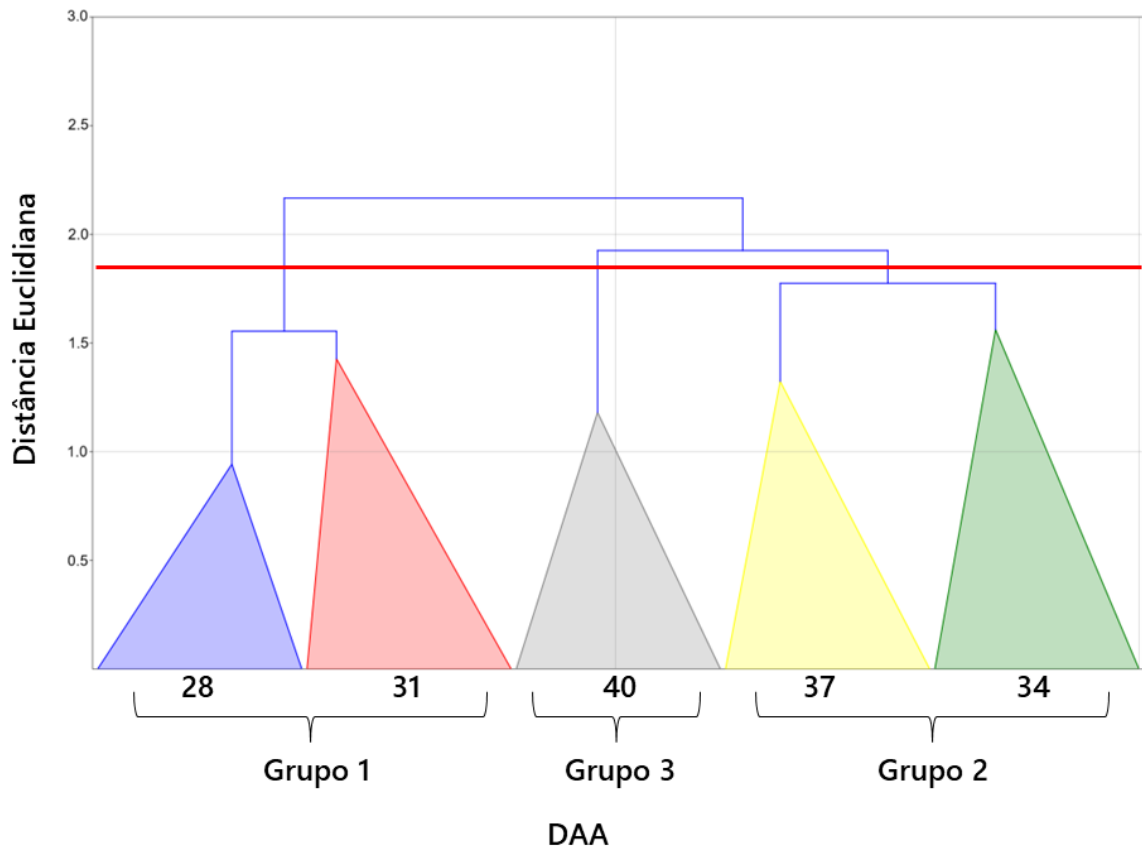
Por sua vez, no grupo cinza os frutos apresentam tamanho no último estágio de desenvolvimento (CP1 positivo), levando em consideração os DAAs aqui analisados, mas características químicas e de coloração menos expressivas (CP2 negativo). Isso pode indicar uma variação específica no amadurecimento, como frutos que não desenvolvem alta intensidade de cor ou frutos em processo de senescência.

Espera-se validar o agrupamento previamente identificado entre os DAAs 34 e 37, bem como investigar a possível existência de outros agrupamentos na análise subsequente.

4.2 Análise de agrupamentos – AA

O dendrograma da análise de agrupamentos (Figura 6), gerado com base nas características físicas, físico-químicas e químicas dos frutos da cagaiteira em diferentes dias após antese, buscou agrupar os diferentes DAA. Para revelar a existência de classes semelhantes entre os DAA, foi adotado o algoritmo *Complete Linkage* com a distância Euclidiana como índice de dissimilaridade. O coeficiente de correlação cofenética (CCC) calculado foi de $r = 0,7$.

Figura 6 – Dendrograma da análise de agrupamentos das cagaitas em diferentes dias após antese (DAA).



Fonte: Próprio autor (2025).

De acordo com Rohlf (1970) caso haja pouca correlação, entre 0,6 e 0,7, pode-se investigar com maior cautela a suposição de que se tem um sistema de *clusters* bem ajustados. Nesses casos é útil averiguar se a falta de ajuste diz respeito ao dendrograma como um todo ou se, há distorção apenas em certas áreas. Esses procedimentos foram realizados e não se observou nada de anormal nos dados e análises do presente trabalho.

Outros algoritmos e distâncias dissimilaridade foram testados (dados não mostrados), sendo a combinação entre *Complete Linkage* e distância Euclidiana aquela que apresentou o dendrograma com menor distorção nas hastes de união dos *clusters*, quando utilizada a função “*Collapse to groups*” no software Past 5.1 (Hammer et al., 2001), permitindo assim uma melhor leitura da inter-relação dos agrupamentos formados.

De acordo com Vaz Patto et al. (2004), os coeficientes de correlação cofenética $\geq 0,56$ indicam que o dendrograma reproduz satisfatoriamente a informação contida na matriz de correlação e na consequente formação dos *clusters*.

Assim como relatado por Rohlf e Fisher (1968), ao utilizar matrizes de distância como dados de entrada do agrupamento é possível observar que a magnitude do CCC está relacionada com a quantidade de observações avaliadas (5 DAA x 40 repetições = 200 observações) e diminui ao passo que o número de observações aumenta.

Houmanat et al. (2021) avaliando características fenotípicas qualitativas e quantitativas e marcadores moleculares em 44 acessos de cártamo obtiveram como resultado um CCC de 0,67.

Almici (2022) analisando a variabilidade genética de *Carthamus tinctorius* L. via diferentes matrizes de dados moleculares, morfológicos e agronômicos, encontrou o valor de 0,57 CCC, considerando-o significativo.

Essas evidências encontradas na literatura, permitem concluir que valor de 0,7 encontrado para o CCC está dentro do recomendando, mostrando assim a confiabilidade do dendrograma.

A linha de corte foi estabelecida entre as alturas 1,5 e 2,0 da ordenada, uma vez que o nível de coesão das classes pode ser arbitrariamente fixado pelo pesquisador interessado em defini-las de acordo com os objetivos do seu estudo (SOKAL e ROHLF, 1962).

Analisando o dendrograma, observa-se que foram revelados três grupos, sendo o Grupo 1 composto pelos DAAs 28 e 31, Grupo 2 composto pelos DAAs 34 e 37 e o Grupo 3 pelo DAA 40. Pode-se confirmar, também, que os DAAs 28 e 31 são os que possuem maior semelhança no dendrograma, por terem sido o primeiro grupo a ser formado, ao contrário do DAA 40, que foi o último grupo a ser formado, mantendo-se distinto dos demais. A dissimilidade maior entre os grupos 1 e 3 no dendrograma confirma que os frutos colhidos na maturidade fisiológica (28-31 DAA) são drasticamente diferentes daqueles que se aproximam do estágio final de maturação (40 DAA).

A formação de grupos distintos mostra que a maturação da cagaita não ocorre de forma abrupta, mas sim gradual, passando por um estágio intermediário bem definido (Grupo 2). Isso sugere que:

- Existe um momento ideal de colheita (próximo dos 34-37 DAA), quando os frutos ainda mantêm estrutura firme, mas já apresentam melhor palatabilidade (menor acidez e coloração mais intensa e vibrante);
- Frutos colhidos no DAA 28 ou 31 podem ser excessivamente ácidos e firmes, o que pode impactar seu consumo *in natura* ou sua aceitação para processamento;

- Frutos colhidos no 40 DAA já estão em estágio avançado de maturação, podem apresentar textura amolecida, porém não apresentam características de senescência ou perda de qualidade.

Nesse contexto, as informações relatadas por Camilo et al. (2013) contribuem para a compreensão do processo de maturação da cagaita ao indicarem que a emissão de botões e flores ocorre majoritariamente em setembro, com a frutificação concentrando-se entre setembro e outubro. Segundo os autores, o desenvolvimento e a maturação dos frutos ocorrem de forma relativamente rápida e estão associados ao período chuvoso, com o pico de produção registrado em outubro. Esses dados sustentam a ideia de que a maturação é um processo contínuo, reforçando a existência de um estágio intermediário bem definido e a importância de identificar o momento ideal de colheita com base nas alterações características físicas, físico-químicas e químicas, observadas ao longo dos diferentes dias após a antese.

Almeida (2020) analisando o cultivo consorciado de cagaiteiras com plantas de cobertura, relatou que os frutos coletados com 34 dias após a antese, apresentavam epicarpo visualmente amarelo.

Reis e Schmiele (2019) relataram que a frutificação da cagaita ocorre cerca de um mês após o florescimento, que acontece entre agosto e setembro.

A análise de agrupamentos indica que o ponto ideal de colheita para consumo *in natura* dos frutos pode estar entre 34 e 37 DAA, quando os frutos atingem o equilíbrio entre doçura e acidez, mantendo ainda uma boa firmeza.

Frutos mais imaturos (28-31 DAA) podem ser aproveitados para sucos ou polpas mais ácidas e frutos mais maduros (40 DAA) podem ser utilizados para produtos fermentados ou doces, onde a degradação da firmeza não compromete a qualidade.

Souza et al. (2008) destacaram que o desenvolvimento e a maturação dos frutos ocorrem entre 30 e 40 dias após a antese das flores, resultado que está em consonância com os resultados obtidos no presente estudo.

Oliveira et al. (2011) ao detalhar a produção de vinho a partir de cagaita, relataram que os frutos utilizados no processo eram maduros e decíduos, ou seja, naturalmente desprendidos da planta, por isso, coletados manualmente do solo. Os frutos decíduos, atingem a plena maturação antes da queda (40 DAA), o que favorece sua utilização na fabricação de vinho de frutas.

Carneiro et al. (2015) ao colherem frutos de *E. dysenterica* ainda ligados à planta-mãe, para estudar o efeito da temperatura e do uso de embalagem na conservação pós-colheita, observaram estágio de maturação verde-maduro.

Silva (2017), ao elaborar uma geleia utilizando um mix de polpa de cagaita (*Eugenia dysenterica*) e mangaba (*Hancornia speciosa*), juntamente com a avaliação dos parâmetros de qualidade, destacou que o rendimento da cagaita na produção de polpa está diretamente relacionado à qualidade dos frutos. Quando utilizados frutos em boas condições, é possível obter um rendimento de até 72% de polpa.

Assumpção et al. (2013) observaram que o néctar misto de polpa de mangaba e cagaita se destacou como um produto atrativo devido às suas características sensoriais, além de apresentar grande potencial tecnológico e nutricional. As duas formulações analisadas obtiveram boa aceitação pelos consumidores, sugerindo que o produto possui viabilidade para inserção no mercado com perspectivas promissoras.

A transição entre os grupos 2 e 3 da Figura 6, pode indicar que, dependendo das condições de armazenamento, os frutos podem amadurecer rapidamente e perder qualidade. Sendo necessária a implementação de estratégias como refrigeração para retardar esse processo e aumentar a vida útil.

Quanto à conservação dos frutos pós-colheita, Sobral et al. (2023), realizaram um trabalho envolvendo cagaita, ciriguela e mangaba em condições do semiárido mineiro, reforçando a necessidade de estudos sobre a conservação desses frutos, uma vez que seu tempo de vida útil é extremamente curto. Diante desse cenário, os autores apontaram que a adoção de técnicas de pós-colheita, como refrigeração e atmosfera modificada, pode contribuir para a extensão da vida útil desses frutos, além de agregar valor comercial ao produto.

As evidências apresentadas na literatura corroboram as inferências obtidas por meio da avaliação visual do dendrograma.

4.3 Algoritmo floresta aleatória (*random forest* – RF)

Com o objetivo de classificar as cagaitas colhidas em diferentes dias após a antese com base nos grupos identificados na análise de agrupamentos, utilizou-se o conjunto de dados obtidos na análise de componentes principais, composto pelas variáveis Diâmetro Longitudinal – DL, Diâmetro Transversal – DT, Firmeza da Casca – FC, Firmeza da Poupá – FP, CROMA e Acidez Titulável – AT.

Esse conjunto de dados foi pré-processado para lidar com valores ausentes na variável AT, os quais foram substituídos pela mediana. Além disso, aplicou-se a função *BorderlineSMOTE* para balanceamento dos grupos no conjunto de treinamento e utilizou-se a função *StandardScaler* para padronização dos dados. Após esse pré-processamento, o conjunto

foi empregado no treinamento e teste do algoritmo Floresta Aleatória (Random Forest – RF), com 70% da base destinada ao treinamento e 30% ao teste.

As RFs apresentam alta acurácia preditiva, o que contribui para a obtenção de resultados mais precisos e para uma identificação mais exata das causas analisadas. Para interpretar os resultados e compreender melhor o processo de aprendizado do modelo, foram extraídas as regras de associação geradas durante sua construção.

Para otimizar os hiperparâmetros usados na criação das RFs, sendo os principais deles, *n_estimators* e *random_state*, cujo primeiro indica quantas árvores de decisão serão criadas dentro do RF e o segundo desempenha um papel fundamental na reprodutibilidade dos resultados e na estabilidade do modelo, realizou-se um experimento avaliativo com duas técnicas de busca por hiperparâmetros ideais, Grid Search e Random Search. Infelizmente após a aplicação das duas técnicas não observou-se melhora significativa nos resultados, por isso, o estudo prosseguiu com os valores padrão de *n_estimators* = 100 e *random_state* = 42, visando um baixo custo computacional.

Após a implementação e execução do modelo em *Python*, os resultados foram analisados por meio da precisão, *recall*, *f1-score*, matriz de confusão e coeficiente de correlação de Matthews (MCC).

A análise dos resultados indica que o modelo atingiu uma acurácia global de 76%, evidenciando um desempenho satisfatório na classificação dos estádios de maturação dos frutos. A Tabela 2 apresenta as métricas de precisão, *recall* e *f1-score* para cada grupo predito.

Tabela 2 – Métricas de desempenho do modelo de classificação.

Grupo	Precisão	<i>Recall</i>	<i>F1-score</i>
1	0,81	0,82	0,81
2	0,71	0,69	0,70
3	0,73	0,75	0,74

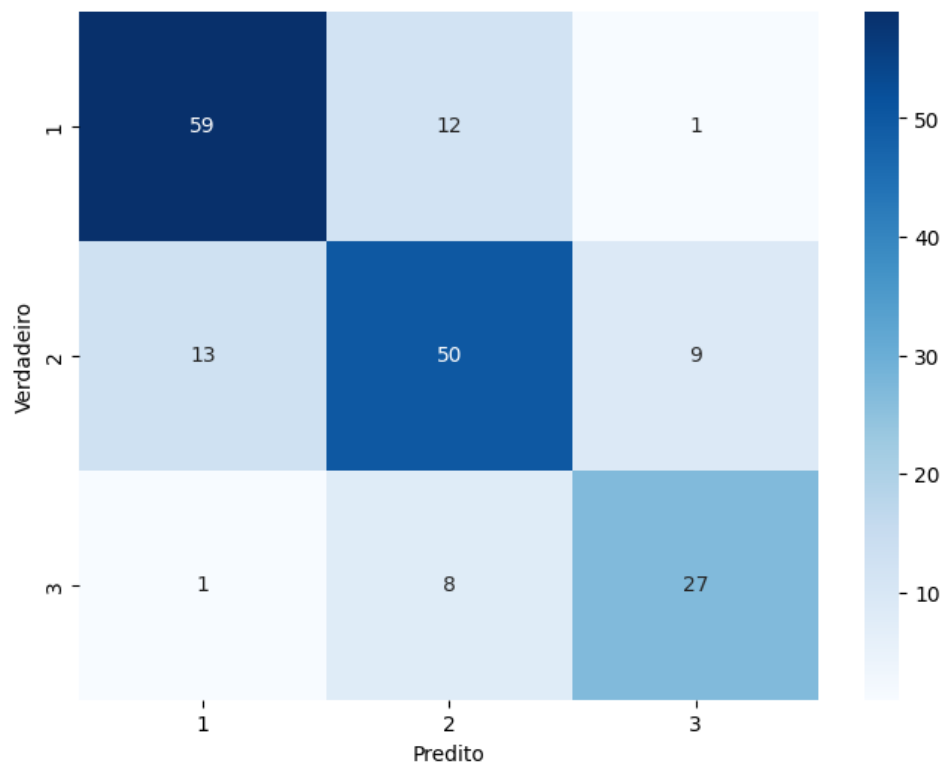
Fonte: Próprio autor (2025).

Os resultados demonstram que o grupo 1 obteve a maior taxa de acerto, com precisão de 81% e *recall* de 82%, indicando que a maioria dos frutos nesse estágio foi corretamente classificada. O grupo 2 apresentou uma leve redução no desempenho, com precisão de 71% e *recall* de 69%, sugerindo que houve certa confusão entre os estádios adjacentes. Por fim, o grupo 3 alcançou precisão de 73% e *recall* de 75%, o que sugere que o modelo foi capaz de diferenciar de maneira satisfatória os frutos mais maduros.

A média ponderada das métricas apresentou valores equilibrados, com um *f1-score* médio de 0,76, reforçando a adequação do modelo para a tarefa proposta. No entanto, a matriz de confusão evidencia algumas instâncias de classificação incorreta, particularmente entre os grupos 1 e 2, o que pode ser atribuído à similaridade entre esses estádios em determinadas características físico-químicas.

A matriz de confusão obtida para o teste do modelo treinado é apresentada na Figura 7.

Figura 7 – Matrizes de confusão com resultados do teste para o modelo floresta aleatória.



Fonte: Próprio autor (2025).

A matriz de confusão gerada, apresentada na Figura 7, destaca a performance do modelo em termos de classificação correta e incorreta entre os três grupos que simbolizam os estádios de maturação avaliados anteriormente. Observa-se que:

- **Grupo 1** foi corretamente classificado em 59 das 72 instâncias, revelando que o modelo possui boa precisão nesse estágio específico;
- **Grupo 2** apresentou uma classificação correta em 50 de 72 casos, indicando um equilíbrio entre precisão e erros de classificação, e;
- **Grupo 3** alcançou uma classificação correta em 27 de 36 instâncias, mostrando-se eficiente no reconhecimento deste estágio específico de maturação.

O resultado do coeficiente de correlação de Matthews (MCC) foi de 0,62, indicando que o modelo de classificação possui um bom desempenho, levando em consideração que os resultados desse índice variam entre -1 e +1 (VILELA JUNIOR et al., 2022). Quanto mais próximo de +1 melhor é o resultado, igual a 0 é nada mais que uma previsão aleatória e -1 indica desacordo total entre previsão e observação. Ele está conseguindo capturar padrões significativos nos dados, não está fazendo previsões aleatórias e tem uma correlação positiva moderada com os valores reais. Pode haver erros na predição, mas ele está conseguindo separar bem os grupos.

Cucchi (2020) utilizou algoritmos inteligentes, incluindo o Random Forest (RF), para estimar os teores de pigmentos fotossintéticos presentes nas membranas dos tilacoides de folhas de cinco espécies da família *Myrtaceae*, demonstrando que tais algoritmos são eficientes para fins de classificação com base em análise espectroscópica. A eficácia do modelo foi avaliada por meio do coeficiente de correlação de Matthews (MCC), com valores superiores a 0,5, indicando desempenho eficiente.

Da mesma forma, Meireles et al. (2022), ao avaliar um modelo preditivo baseado em RF, em um estudo fitoquímico das raízes de *Euphorbia phosphorea* Mart. (*Euphorbiaceae*), reportaram um valor de MCC igual a 0,56, o qual foi considerado satisfatório para os fins propostos.

Por fim, Figueiredo et al. (2024) investigaram o uso de modelos preditivos baseados em RF aplicados à análise de raízes de espécies do gênero *Maytenus*, amplamente utilizadas na medicina popular, especialmente no tratamento de distúrbios do trato gastrointestinal. Os autores avaliaram a eficácia dos modelos frente à ação sobre quatro microrganismos distintos, observando valores de MCC entre 0,50 e 0,64. Esses resultados permitiram concluir que os modelos apresentaram bons níveis de precisão, desempenho, robustez e capacidade preditiva.

Esses resultados indicam que o RF, foi eficaz na previsão do estágio de maturação dos frutos da cagaita, podendo ser utilizado como uma ferramenta auxiliar para a caracterização e acompanhamento do desenvolvimento dos frutos ao longo do tempo.

5 CONSIDERAÇÕES FINAIS

Diante dos resultados obtidos, verifica-se que a utilização integrada da estatística multivariada e do algoritmo de aprendizado de máquina Floresta Aleatória (*Random Forest* – RF) mostrou-se eficaz na classificação dos diferentes estádios de maturação dos frutos de cagaita. A Análise de Componentes Principais (ACP) permitiu a identificação das variáveis mais relevantes para a classificação, com destaque para DL, DT, FC, FP, CROMA e AT, que explicaram parte significativa da variabilidade presente nos dados.

A análise de agrupamentos evidenciou a formação de três grupos distintos, em que o Grupo 1 corresponde aos dois primeiros estádios de maturação que foram determinados por meio dos DAAs 28 e 31. O Grupo 2 representou os DAAs 34 e 37 e o Grupo 3 foi melhor representado pelo DAA 40. Esses agrupamentos foram validados por um coeficiente de correlação cofenética satisfatório (0,7), o que confere robustez à estrutura de agrupamento formada.

Quanto ao algoritmo floresta aleatória, foram usados 70% dos dados para treinamento e 30% para teste. No relatório das métricas o modelo demonstrou uma precisão geral satisfatória para todos os estádios agrupados. Os valores de *recall* indicam a capacidade do modelo em identificar corretamente as instâncias de cada estágio, com um *recall* médio de 0,75. O *f1-score*, que pondera igualmente precisão e *recall*, mostrou um desempenho coeso do modelo, com um valor médio de 0,75.

O coeficiente de correlação de Matthews (MCC), que faz uma avaliação geral do modelo RF, foi de 0,62, evidenciando um desempenho aceitável para aplicações práticas. No entanto, recomenda-se a investigação de técnicas adicionais de pré-processamento de dados ou a combinação com outros algoritmos, objetivando aprimorar ainda mais a acurácia dos modelos classificatórios aplicados à fruticultura nativa.

Os resultados indicam que o modelo de Floresta Aleatória implementado é relativamente eficaz na tarefa de classificação dos estádios de maturação dos frutos. Apesar de uma performance robusta nos Grupos 1 (DAAs 28 e 31) e 3 (DAA 40), o modelo apresentou dificuldades moderadas na classificação do Grupo 2 (DAAs 34 e 37), o que pode estar relacionado a uma sobreposição nas características dos dados para esses grupos.

As análises revelaram três grupos distintos conforme os dias após a antese (DAA), indicando maturação gradual da cagaita. O Grupo 1 (28 e 31 DAA) apresentou frutos mais firmes e ácidos, característicos da maturidade fisiológica. O Grupo 2 (34 e 37 DAA) indicou o ponto ótimo de colheita, com melhor equilíbrio entre firmeza, acidez e coloração. Já o Grupo 3 (40 DAA) corresponde a frutos mais maduros, com textura amolecida, mas ainda com boa

qualidade. Esses resultados sugerem que o uso dos frutos deve ser direcionado conforme seu estágio de maturação para consumo in natura, processamento que gere sucos ou polpas ou ainda serem utilizados para produtos fermentados ou doces.

Como sugestão para investigações futuras, recomenda-se a avaliação do desempenho do algoritmo de floresta aleatória em conjunto com outros algoritmos de aprendizado de máquina, por meio da formação de comitês de classificadores (*ensemble methods*), com o objetivo de aprimorar a precisão e a robustez na tarefa de classificação dos estádios de maturação dos frutos. Além disso, propõe-se o desenvolvimento de modelos automáticos de classificação baseados em processamento digital de imagens, os quais podem ampliar a aplicabilidade prática da metodologia, especialmente em contextos de produção e comercialização em larga escala.

REFERÊNCIAS BIBLIOGRÁFICAS

- ALMEIDA, M. J. de. **Cultivo consorciado de cagaiteira (*Eugenia dysenterica* DC): Influência na qualidade tecnológica dos frutos**. 2020. 53 f. Dissertação (Mestrado em Agroquímica), Instituto Federal de Educação, Ciência e Tecnologia Goiano, Rio Verde, 2020. Disponível em: <<https://repositorio.ifgoiano.edu.br/handle/prefix/1236>>. Acesso em: 10 mar. 2025.
- ALMICI, M. da S. **Acesso da variabilidade genética de *Carthamus tinctorius* L. via diferentes matrizes de dados moleculares, morfológicos e agronômicos**. 2022. 70 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas), Universidade Estadual de Mato Grosso, Cáceres, 2022. Disponível em: <https://portal.unemat.br/media/files/DISERTAO_OK_FICHA_CATALOGRAFICA.pdf>. Acesso em: 10 mar. 2025.
- ANDRADE, A. R. S. de; CRUZ, A. F. da S.; BEZERRA, T. M. Ê.; LIMA, J. R. de S. Aplicação da estatística multivariada na identificação das características biométricas de produção e qualidade de frutos do melão. **Brazilian Journal of Applied Technology for Agricultural Science**, Guarapuava-PR, v.10, n.1, p.25-40, 2017. Disponível em: <<https://revistas.unicentro.br/index.php/repaa/article/view/4707/3401>>. Acesso em: 05 abr. 2025.
- ANTÔNIO, C. T. M. **Classificação de navios baseados em assinaturas acústicas**. 2019. 104 f. Dissertação (Mestrado em Ciências Militares Navais) - Escola Naval, Lisboa, 2019. Disponível em: <<http://hdl.handle.net/10400.26/30352>>. Acesso em: 03 mar. 2024.
- ASSOCIATION OF OFFICIAL ANALYTICAL CHEMISTS. Official methods of analysis. 19th Edition, Washington: AOAC, 2012.
- ASSUMPÇÃO, C. F.; BACHIEGA, P.; SANTANA, A. T. M. C.; MORZELLE, M. C.; BOAS, B. M. V.; SOUZA, E. C. de. Néctar misto de mangaba (*Hancoria speciosa* Gomes) e cagaíta (*Eugenia dysenterica*): Perfil sensorial e características físico-químicas. **Revista Brasileira de Produtos Agroindustriais**, v. 15, n. 3, p. 219-224, 2013. Disponível em: <https://www.researchgate.net/profile/Carolina-Assumpcao-4/publication/277937634_NECTAR_MISTO_DE_MANGABA_Hancoria_speciosa_Gomes_E_CAGAITA_Eugenia_dysenterica_PERFIL_SENSORIAL_E_CARACTERISTICAS_FISICO-QUIMICAS/links/56c486d308ae7fd4625a2836/NECTAR-MISTO-DE-MANGABA-Hancoria-speciosa-Gomes-E-CAGAITA-Eugenia-dysenterica-PERFIL-SENSORIAL-E-CARACTERISTICAS-FISICO-QUIMICAS.pdf>. Acesso em: 10 mar. 2025.
- BASTOS, C. R. P.; DE FARIA, V. A.; PEREIRA, P. V. D. PRODUÇÃO DE GRÃOS NO ESTADO DO TOCANTINS: UMA ANÁLISE ATRAVÉS DE TÉCNICAS ESTATÍSTICAS. **Revista Integralização Universitária**, n. 19, p. 07-19, 2018. Disponível em: <<https://doi.org/10.31501/1982-9280.2018V12N19p.%2007-19>>. Acesso em: 08 jun. 2025.
- BITENCOURT, W. A.; SILVA, D. M.; XAVIER, G. do C. Pode a inteligência artificial apoiar ações contra evasão escolar universitária?. **Revista Ensaio: Avaliação e Políticas Públicas em Educação**, v. 30, p. 669-694, 2021. Disponível em: <<https://doi.org/10.1590/S0104-403620220003002854>>. Acesso em: 02 mar. 2024.

BOLFE, E. L.; SANO, E. E.; CAMPOS, S. K. **Dinâmica agrícola no cerrado: análises e projeções**. Embrapa, Brasília, p. 308, 2020. Disponível em: <<http://www.alice.cnptia.embrapa.br/alice/handle/doc/1121716>>. Acesso em: 18 mar. 2024.

Breiman, L. Bagging predictors. **Machine Learning**, v. 24, p. 123–140, 1996. Disponível em: <<https://doi.org/10.1007/BF00058655>>. Acesso em: 05 abr. 2025

Breiman, L. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001. Disponível em: <<https://doi.org/10.1023/A:1010933404324>>. Acesso em: 05 abr. 2025

BUENO, G. H.; GUEDES, M. N. S.; SOUZA, A. G. de; MADEIRA, A. P. C.; GARCIA, E. M.; TAROCO, H. A.; MELO, J. O. F. Caracterização física e físico-química de frutos de *Eugenia dysenterica* DC. originados em região de clima tropical de altitude. **Brazilian Journal of Biometrics**, [S. l.], v. 35, n. 3, p. 515–522, 2017. Disponível em: <http://200.131.250.9/index.php/BBJ/article/view/101>. Acesso em: 05 abr. 2025.

BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 93p.

CAMILO, Y. M. V.; SOUZA, E. R. B.; VERA, R.; NAVES, R. V. Fenologia, produção e precocidade de plantas de *Eugenia dysenterica* visando melhoramento genético. **Revista de Ciências Agrárias**, v. 36, n. 2, p. 192-198, 2013. Disponível em: <<https://doi.org/10.19084/rca.16297>>. Acesso em: 10 mar. 2025.

CAMPOS, G. A. M. **Classificação do estágio de maturação de cagaitas utilizando estatística multivariada e redes neurais artificiais**. 2021. 58 f. Dissertação (Mestrado em Engenharia Agrícolas), Câmpus Central – Universidade Estadual de Goiás, Anápolis, 2021.

CARNEIRO, J. DE O.; SOUZA, M. A. DE A. DE S.; RODRIGUES, Y. J. DE M.; MAPELI, A. M. Efeito da temperatura e do uso de embalagem na conservação pós-colheita de frutos de cagaita (*Eugenia dysenterica* DC.). **Revista Brasileira de Fruticultura**, Jaboticabal, v. 37, p. 568-577, 2015. Disponível em: <<https://doi.org/10.1590/0100-2945-157/14>>. Acesso em: 18 mar. 2024.

CASTRO, J. V.; RAMADAN, R.; SOUSA, W. D. S.; GONÇALVES, R. A.; LEANDRO, W. M. Crescimento de mudas de Cagaita, produzida em Sistema Agroecológico, sob dose de Biofertilizante a Base do Kombucha. **Cadernos de Agroecologia**, Recife, v. 8, n. 2, 2013. Disponível em: <<https://revista.aba-agroecologia.org.br/cad/article/view/14999/9591>>. Acesso em: 05 abr. 2025.

CHAVES NETO, J. R.; SILVA, S. de M.; DANTAS, R. L. Atributos de qualidade, compostos bioativos e atividade antioxidante de frutos de uvaieira durante a maturação. **Agrarian**, Dourados, v. 13, n. 49, p. 269–308, 2020. Disponível em: <<https://ojs.ufgd.edu.br/index.php/agrarian/article/view/8684>>. Acesso em: 22 mar. 2024.

CHO, D. F.; SCHWAIDA, S. F.; CICERELLI, R. E.; ALMEIDA, T.; RAMOS, A. P. M.; SANO, E. E. Desempenho do Algoritmo de Classificação de Imagens Random Forest para Mapeamento do Uso e Cobertura do Solo no Cerrado Brasileiro. **Anuário do Instituto de Geociências**, Rio de Janeiro, v. 44, 2021. Disponível em: <<https://revistas.ufrj.br/index.php/aigeo/article/view/37979>>. Acesso em: 18 mar. 2024.

CUCCHI, P. A. **Identificação de marcadores em espécies de *Myrtaceae* por meio de espectroscopia, análise de imagens e quimiometria**. 2020. 118 f. Dissertação (Programa de Pós-Graduação em Química - Mestrado) - Universidade Estadual do Centro-Oeste, Guarapuava-PR. Disponível em: <<http://tede.unicentro.br:8080/jspui/handle/jspui/1600>>. Acesso em: 06 abr. 2025.

DE-LA-TORRE, M.; ZATARAIN, O.; AVILA-GEORGE, H.; MUÑOZ, M.; OBLITAS, J.; LOZADA, R.; MEJÍA, J.; CASTRO, W. Multivariate Analysis and Machine Learning for Ripeness Classification of Cape Gooseberry Fruits. **Processes**, v. 7, 2019. Disponível em: <<https://doi.org/10.3390/pr7120928>>. Acesso em: 06 abr. 2025.

EMBRAPA. Biologia Reprodutiva e Fenologia. Disponível em: <<https://www.cnpf.embrapa.br/pesquisa/efb/temp/biolo.htm>>. Acesso em: 08 jun. 2025.

FÁVERO, L. P. **Análise de Dados: Técnicas Multivariadas Exploratórias com SPSS e STATA**. Rio de Janeiro: GEN Atlas, 2015. *E-book*. p.IX. ISBN 9788595155589. Disponível em: <https://app.minhabiblioteca.com.br/reader/books/9788595155589/>. Acesso em: 05 abr. 2025.

FIGUEIREDO, P. T. R. de; DUARTE, M. C.; CORDEIRO, L. V.; MENEZES, R. P. B. de; LIMA, E. de O.; SCOTTI, M. T.; TAVARES, J. F.; COSTA, V. C. de O.; SILVA, M. S. da. Alcaloides sesquiterpênicos isolados de *Maytenus distichophylla* (Celastraceae). **Química Nova**, v. 47, n. 4, p. 1-6, 2024. Disponível em: <<http://dx.doi.org/10.21577/0100-4042.20230123>>. Acesso em: 06 abr. 2025.

FORNARI, J. L. **Caracterização físico-química e fúngica das silagens de milho utilizadas na alimentação de bovinos leiteiros na microrregião Oeste do Paraná**. 2019. 90 f. Dissertação (Mestrado em Zootecnia) - Universidade Estadual do Oeste do Paraná, Marechal Cândido Rondon, 2019. Disponível em: <<https://tede.unioeste.br/handle/tede/4417>>. Acesso em: 26 fev. 2024.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009. *E-book*. Disponível em: <https://app.minhabiblioteca.com.br/reader/books/9788577805341/>. Acesso em: 05 abr. 2025.

HAMMER, Ø.; HARPER, D. A. T.; RYAN, P. D. PAST: Paleontological Statistics Software Package for Education and Data Analysis – versão 5.1 (February 2025). Oslo: Natural History Museum – University of Oslo, 2001. Disponível em: <<https://www.nhm.uio.no/english/research/resources/past/>>. Acesso em: 12 abr. 2025.

HAN, Jiawei; PEI, Jian; TONG, Hanghang. **Data mining: concepts and techniques**. Morgan kaufmann, 2022.

HARTSHORN, S. **Machine learning with random forests and decision trees: a visual guide for beginners**. 1. ed. Michigan: Fairlynerdy, 2016. E-book (82 p.). Disponível em: <<http://www.fairlynerdy.com/our-books/>>. Acesso em: 17 mar. 2024.

HO, T. K. The random subspace method for constructing decision forests. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 20, n. 8, p. 832-844, 1998.

Disponível em: <<https://doi.org/10.1109/34.709601>>. Acesso em: 06 abr. 2025.

HOUMANAT, K.; DOUAIK, A.; CHARAFI, J.; HSSAINI, L.; EL FECHTALI, M.; NABLOUSSI, A. Appropriate statistical methods for analysis of safflower genetic diversity using agglomerative hierarchical cluster analysis through combination of phenotypic traits and molecular markers. **Crop Science**, v.61, n.6, p. 4164-4180, 2021. Disponível em: <<https://doi.org/10.1002/csc2.2059>>. Acesso em: 10 mar. 2025.

KAISER, H. F. The varimax criterion for analytic rotation in fator analysis. **Psychometrika**, New York, v. 23, n. 3, p. 187- 200, 1958.

KARADAG, A.; OZCELIK, B.; SANER, S.; Review of methods to determine antioxidant capacities. **Food Analytical Methods**, v. 2, p. 41-60, 2009. Disponível em: < <https://doi.org/10.1007/s12161-008-9067-7>>. Acesso em: 05 abr. 2025.

LACERDA, C. R. B. **Classificação supervisionada: árvores de decisão e florestas aleatórias**. 2023. Dissertação (Mestrado em Matemática) – Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Coimbra, 2023. Disponível em: <https://estudogeral.uc.pt/retrieve/265409/Mestrado_01_09_2023.pdf>. Acesso em: 03 mar. 2024.

LATTUADA, D. S.; BARROS, N.; HAGEMANN, A. SOUZA, V. D de. Caracterização físico-química e desenvolvimento pós-colheita de jabuticabas (*Plinia peruviana* e *P. cauliflora*). **Série Botânica**, Porto Alegre, 2020.

LEMOS-FILHO, J. P. de. Fotoinibição em três espécies do cerrado (*Annona crassifolia*, *Eugenia dysenterica* e *Campomanesia adamantium*) na estação seca e na chuvosa. **Brazilian Journal of Botany**, v. 23, n. 1, p.45-50, 2000. Disponível em: < <https://doi.org/10.1590/S0100-84042000000100005>>. Acesso em: 05 abr. 2025.

LOESCH, C.; HOELTGEBAUM, M. **Métodos estatísticos multivariados**. São Paulo: Saraiva, 2012.

LORENZI, H. **Árvores brasileiras: manual de identificação e cultivo de plantas arbóreas do Brasil**. vol. 2. Instituto Plantarum, Nova Odessa, SP, 2002.

MANLY, B. F. J.; ALBERTO, J. A. N. **Métodos estatísticos multivariados: uma introdução**. 4. ed. Porto Alegre: Bookman, 2019. E-book. ISBN 9788582604991. Disponível em: <<https://app.minhabiblioteca.com.br/#/books/9788582604991/>>. Acesso em: 21 fev. 2024.

MARQUES, B. G. **Análise de espécies invasoras e respetiva proposta de intervenção – concelho de Gavião**. 2019. 91 f. Dissertação (Mestrado em Recursos Florestais), IPC - Instituto Politécnico de Coimbra, ESAC - Escola Superior Agrária de Coimbra, Coimbra, 2019. Disponível em: <<http://hdl.handle.net/10400.26/34899>>. Acesso em: 24 mar. 2024.

MARTELETTO, S. R. **Técnicas de seleção de atributos através de Random Forests: um estudo de caso para detecção de tendências em séries temporais financeiras**. 2022. 86f. Dissertação (Mestrado em Modelagem de Sistemas Complexos) – Escola de Artes, Ciências e

Humanidades, Universidade de São Paulo, São Paulo, 2022. Disponível em: <<https://doi.org/10.11606/D.100.2022.tde-08032023-134543>>. Acesso em: 03 mar. 2024.

MARTINOTTO, C.; PAIVA, R.; SOUZA, H. D.; MASETTO, T. E.; VARGAS, D. P.; SANTOS, B. R. Indução de calos em explantes foliares de cagaiteira com 2,4-D e BAP em meio líquido. **Revista Brasileira de Horticultura Ornamental**, v. 13, p. 1131-1134, 2007. Disponível em: <<https://doi.org/10.14295/oh.v13i0.1646>>. Acesso em: 05 abr. 2025.

MATTHEWS, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. **Biochimica et Biophysica Acta (BBA) - Protein Structure**, v. 405, n. 2, p. 442-451, 1975. Disponível em: <[https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)>. Acesso em: 06 abr. 2025.

MEIRELES, R. A. R.; ABREU, L. S.; SILVA, J. P. R. e; CAVALCANTI, A. B. S.; MENEZES, R. P. B. de; RODRIGUES, G. C. S.; RODRIGUES JUNIOR, V. S.; MELO, J. I. M. de; KATO, M. J.; COSTA, V. C. de O.; SCOTTI, M. T.; TAVARES, J. F. Diterpenos Ent-Abietanos de Euphorbia Phosphorea (Euphorbiaceae). **Química Nova**, v. 45, n. 9, p. 1092-1097, 2022. Disponível em: <<http://dx.doi.org/10.21577/0100-4042.20170932>>. Acesso em: 06 abr. 2025.

MENESES, M. D.; CARREIRA, V. dos S.; MOREIRA, B. R. de A.; VALE, W. G. do; ROLIM, G. de S.; SILVA, R. P. da. Machine learning models for classifying coffee fruits detachment force. **Agronomy Journal**, v. 116, p. 2362–2369, 2024. Disponível em: <<https://doi.org/10.1002/agj2.21633>>. Acesso em: 06 abr. 2025.

NASCIMENTO, D. T. F.; OLIVEIRA, I. Jose de; LUIZ, Gislaine Cristina. Panorama dos sistemas de classificação climática e as diferentes tipologias climáticas referentes ao estado de Goiás e ao Distrito Federal/Brasil. **Elisee: Revista de geografia da UEG**, v. 5, n. 2, p. 59-86, 2016. Disponível em: <<https://www.revista.ueg.br/index.php/elisee/article/view/5769>>. Acesso: 17 mar. 2024.

OLIVEIRA, M. E. A.; GALVÃO, L. R.; OLIVAL, A. de A. Floresta aleatória na seleção de perfis de score de pagamento: aplicação em crédito rural. In: CIENTÍFICA DAS CIÊNCIAS SIMPÓSIO DE INICIAÇÃO AGRÁRIAS (SICCA), 9., p. 55, 2021, Alta Floresta. **Anais...** Universidade do Estado de Mato Grosso, 2021. Disponível em: <https://www.researchgate.net/profile/Patricia-Crivelari-Costa/publication/358337026_IXCIENTIFICA_DAS_CIENCIAS_SIMPOSIO_DE_INICIACAO_AGRARIAS_SICCA_EditoresOrganizadores/links/61fc3bfa007fb5044734d047/IXCIENTIFICA-DAS-CIENCIAS-SIMPOSIO-DE-INICIACAO-AGRARIAS-SICCA-Editores-Organizadores.pdf#page=54>. Acesso em: 02 mar. 2024.

OLIVEIRA, M. E. S.; PANTOJA, L.; DUARTE, W. F.; COLLELA, C. F.; VALARELLI, L. T.; SCHWAN, R. F.; DIAS, D. R. Fruit wine produced from cagaita (*Eugenia dysenterica* DC) by both free and immobilised yeast cell fermentation. **Food Research International**. V. 44, Issue 7, 2011. Disponível em: <https://doi.org/10.1016/j.foodres.2011.02.028>. Acesso em: 10 mar. 2025.

RAIMUNDO, F. J. dos S. N. **Melhorar a Sustentabilidade da Irrigação usando Machine Learning**. 2021. 96f. Dissertação (Mestrado em Engenharia de Telecomunicações e Informática) – Departamento de Ciências e Tecnologias da Informação, Instituto Universitário

de Lisboa, Lisboa, 2021. Disponível em: <<http://hdl.handle.net/10071/23850>>. Acesso em: 03 mar. 2024.

REIS, A. F.; SCHMIELE, M. Características e potencialidades dos frutos do Cerrado na indústria de alimentos. **Brazilian Journal of Food Technology**, v. 22, 2019. Disponível em: <https://doi.org/10.1590/1981-6723.15017>. Acesso em: 10 mar. 2025.

RIBEIRO, C. E.; ZÁRATE, L. E. Data Preparation for Longitudinal Data Mining: a case study on human ageing. **Journal of Information and Data Management**, v. 7, n. 2, p. 116-129, 2016. Disponível em: <<https://periodicos.ufmg.br/index.php/jidm/article/view/339>>. Acesso em: 02 mar. 2024.

RIQUETI, G. A.; RIBEIRO, C. E.; ZÁRATE, L. E. Classificando perfis de longevidade de bases de dados longitudinais usando Floresta Aleatória. Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 6. 2018, São Paulo. **Anais...** Porto Alegre: Sociedade Brasileira de Computação, 2018, p. 33-40. Disponível em: <<https://doi.org/10.5753/kdmile.2018.27382>>. Acesso em: 06 abr. 2025.

ROCHA, G. M.; GARCIA, J. M.; SILVA, A. L. DA; LONGO, R. M. Atributos físicos do solo de florestas urbanas como indicadores de degradação. **Ciência Florestal**, Santa Maria, v. 30, p. 767-778, 2020. Disponível em: <<https://www.scielo.br/j/cflo/a/vZgkRpTDGzzNqNLMZzLkPKc/?format=pdf&lang=pt>>. Acesso em: 21 fev. 2024.

RODRIGUES, D. B.; MENDONÇA, H. de O. P.; NOGUEIRA, L. A.; PAULA, A. C. C. F. F. de; MELO, A. C. de; REINA, L. D. C. B.; SILVA, M. R.; ARAÚJO, R. L. B. de; AUGUSTI, R.; MELO, J. O. F. Characterization of volatile compounds and bioactive compounds of pulp and jelly of cagaita by solid phase microextraction in the headspace mode and mass spectrometry by *paper spray*. **Research, Society and Development**, 2021.

RODRIGUES-SILVA, P. L.; AMORIM, G. C.; ANDRADE, I. E. P. C.; CUNHA, V. A.; FIGUEIREDO, L. H. M.; GROSSI-DE-SA, M. F. Monitoramento Tecnológico da Planta Cagaita (*Eugenia dysenterica*) e Aplicações Biotecnológicas Potenciais. **Cadernos de Prospecção**, [S. l.], v. 14, n. 4, p. 1248–1264, 2021. Disponível em: <<https://periodicos.ufba.br/index.php/nit/article/view/38459>>. Acesso em: 19 mar. 2024.

ROHLF, F. J. Adaptative hierarquical clustering schemes. **Systematic Zoology**, v. 19, n. 1, p. 58-82, 1970. Disponível em: <<https://doi.org/10.1093/sysbio/19.1.58>>. Acesso em: 10 mar. 2025.

ROHLF, F. J.; FISHER, D. R. Tests for hierarchical structure in random data sets. **Systematic Biology**, v.17, n.4, p. 407-412, 1968. Disponível em: <<https://doi.org/10.1093/sysbio/17.4.407>>. Acesso em: 10 mar. 2025.

SAMMUT, C.; WEBB, G. I. **Encyclopedia of Machine Learning**. Springer Reference. Springer, New York, 2011.

SANTANA, F. B. de. **Floresta aleatória para desenvolvimento de modelos multivariados de classificação e regressão em química analítica**. 2020. 1 recurso online (139 p.) Tese (Doutorado em Ciências) - Universidade Estadual de Campinas, Instituto de Química,

Campinas, SP. Disponível em: <<https://hdl.handle.net/20.500.12733/1639152>>. Acesso em: 3 mar. 2024.

SANTOS, W. W. V.; SILVA, K. R. de O.; BARBOSA, R. C.; OLIVEIRA, J. B. de; SILVA, J. S. A. da; MEDEIROS, E. V. de. Efeito de diferentes métodos de maturação sobre a qualidade da banana prata. **Diversitas Journal**, Santana do Ipanema, v. 4, n. 3, p. 1092–1104, 2019. Disponível em: <https://diversitas.emnuvens.com.br/diversitas_journal/article/view/838>. Acesso em: 22 mar. 2024.

SILVA FILHO, F. R. da; VENANCIO, M. V. de L.; SOUSA, I. A. A. de. Aprendizado por Agrupamento. **Revista Sistemas e Mídias Digitais**, v. 6, n. 1, 2021. Disponível em: <<https://revistasmd.virtual.ufc.br/arquivos/volume-6/numero-1/rsmd-v6-n1-2.pdf>>. Acesso em: 06 abr. 2025.

SILVA, F. S. **Elaboração de geleia com mix de polpa de cagaita (*Eugenia dysenterica*) e mangaba (*Hancornia speciosa*) e avaliação dos parâmetros de qualidade**. 2017. 74f. Dissertação (Mestrado em Ciência e Tecnologia de Alimentos), Universidade Federal do Tocantins, Palmas, 2017. Disponível em: <<http://hdl.handle.net/11612/524>>. Acesso em: 10 mar. 2025.

SILVA, J. A.; SILVA, D. B.; JUNQUEIRA, N. T. V.; ANDRADE, L. R. M. de. **Coleta de sementes, produção de mudas e plantio de espécies frutíferas nativas dos cerrados: informações exploratórias**. Planaltina: EMBRAPA/CPAC, 1992. 23 p. (Documentos, 44).

SILVA, M. M. M da. **Estudo do desenvolvimento fisiológico da cagaita (*Eugenia dysenterica*)**. 2016. 120 f. Dissertação (Mestrado em Agronomia). Escola de Agronomia - Universidade Federal de Goiás, Goiânia, 2016. Disponível em: <<http://repositorio.bc.ufg.br/tede/handle/tede/6789>>. Acesso em: 16 abr. 2025.

SILVA, M. R.; FREITAS, L. G.; SOUZA, A. G.; ARAÚJO, R. L. B.; LACERDA, I. C. A.; PEREIRA, H. V.; AUGUSTI, R.; MELO, J. O. F. Antioxidant activity and metabolomic analysis of cagaitas (*Eugenia dysenterica*) using paper spray mass spectrometry. **Journal of the Brazilian Chemical Society**, v. 30, n. 5, p. 1034-1044, 2019. Disponível em: <<https://doi.org/10.21577/0103-5053.20190002>>. Acesso em: 05 abr. 2025.

SIMPSON, M. G. Species and Conservation in Plant Systematics. **Species and Conservation in Plant Systematics**, v. 2, p. 649-668, 2010. Disponível em: <<https://doi.org/10.1016/B978-0-12-374380-0.50019-1>>. Acesso em: 08 jun. 2025.

SOBRAL, R. R. S.; MIZOBUTSI, G. P.; MIZOBUTSI, E. H.; SANTOS, R. C. dos; AGUIAR, F. S.; CASTRICINI, A.; SOARES, C. F.; ALMEIDA, L. B. Caracterização pós-colheita de cagaita, ciriguela, mangaba em condições do semiárido mineiro. **Contribuciones a Las Ciencias Sociales**, v.16, n.10, p. 22261-22280, 2023. Disponível em: <<https://doi.org/10.55905/revconv.16n.10-210>>. Acesso em: 10 mar. 2025.

SOKAL, R. R.; ROHLF, F. J. The comparison of dendrograms by objective methods. **Taxon**, v. 11, n. 2, p. 33-40, 1962. Disponível em: <<https://doi.org/10.2307/1217208>>. Acesso em: 10 mar. 2025.

SOUZA, E. R. B. de; NAVES, R. V.; BORGES, J. D.; VERA, R.; FERNANDES, E. P.; SILVA, L. B. e; TRINDADE, M. da G. Fenologia de cagaiteira (*Eugenia dysenterica* DC.) no Estado de Goiás. **Revista Brasileira de Fruticultura**, v. 30, p. 1009-1014, 2008. Disponível em: < <https://doi.org/10.1590/S0100-29452008000400028>>. Acesso em: 10 mar. 2025.

TERÁN-ORTIZ, G. P.; SILVA, M. G. da; LIMA, H. C. de; PACIULLI, S. de O. D.; SILVA, V. A. Avaliação pós-colheita de cagaita (*Eugenia dysenterica*). Congresso Brasileiro de Processamento de Frutas e Hortaliças, 3. 2013, Ilhéus. **Anais...** Repositório Acesso Livre à Informação Científica da Embrapa (Alice). Disponível em: < <http://www.alice.cnptia.embrapa.br/alice/handle/doc/1036208>>. Acesso em: 05 abr. 2025.

TUCKER, G.A. Introduction. In: SEYMOR, G.B.; TAYLOR, J.E.; TUCKER, G.A. **Biochemistry of fruit ripening**. New York: Chapman & Hall, 1993. p. 2-51.

VASCONCELOS, G. T.; SOUZA, K. X. S. de; CAMARGO NETO, J.; OLIVEIRA, S. R. de M. Utilizando processamento em cascata e agrupamento em imagens para otimizar modelos de classificação de solos. CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 12. 2019, Indaiatuba. **Anais...** Ponta Grossa: SBIAGRO, 2019. Disponível em: <[https://ainfo.cnptia.embrapa.br/digital/bitstream/item/216893/1/PC-Utilizando-processamento-SBIAGRO-2019](https://ainfo.cnptia.embrapa.br/digital/bitstream/item/216893/1/PC-Utilizando-processamento-SBIAGRO-2019.pdf)>.pdf. Acesso em: 02 mar. 2024.

VAZ PATTO, M. C.; SATOVIC, Z.; PÊGO, S.; FEVEREIRO, P. Assessing the genetic diversity of Portuguese maize germplasm using microsatellite markers. **Euphytica**, v.137, p.63-72, 2004. Disponível em: <<https://link.springer.com/article/10.1023/B:EUPH.0000040503.48448.97>>. Acesso em: 10 mar. 2025.

VICINI, L. **Geração de subsídios para a tomada de decisão na cadeia produtiva da bovinocultura do Brasil**. Santa Maria: EDUFSM, 2006.

VICINI, L.; SOUZA, A. M.; MORALES, F. E. C.; SOUZA, F. M. **Técnicas Multivariadas Exploratórias: teorias e aplicações no software Statistica**. Santa Maria: EDUFSM, 2018. E-book. Disponível em: <<https://plataforma.bvirtual.com.br>>. Acesso em: 19 fev. 2024.

VIEIRA, F. M. DA S.; MARTINS, C. F.; FIUZA, T. S.; PAULA, J. R. de. O progresso científico no estudo de *Eugenia dysenterica*: uma análise cientométrica. **Revista Virtual de Química**, v. 15, n. 6, p. 1059-1069, 2023. Disponível em: <<http://dx.doi.org/10.21577/1984-6835.20230012>>. Acesso em: 05 abr. 2025.

VILELA JUNIOR, G. de B.; LIMA, B. N.; RIBEIRO, H. F.; RODRIGUES, M. F.; PEREIRA, A. de A.; OLIVEIRA, J. R. L de; SILIO, L. F.; PASSOS, R. P. Importância do Índice Fowlkes-Mallows (FMI), do Coeficiente de Correlação de Matthews (MCC) e do Índice Youden (IY) nos classificadores de inteligência artificial na área da saúde. **Revista CPAQV**, v. 14, n. 3, 2022. Disponível em: <<https://revista.cpaqv.org/index.php/CPAQV/article/view/996/712>>. Acesso em: 06 abr. 2025.

WANG, A. N.; YI, X. W.; YU, H. F.; DONG, B.; QIAO, S. Y. Free radical scavenging activity of *Lactobacillus fermentum* in vitro and its antioxidative effect on growing-finishing pigs. **Journal of Applied Microbiology**, v. 107, n. 4, p. 1140-1148, 2009. Disponível em: < <https://doi.org/10.1111/j.1365-2672.2009.04294.x>>. Acesso em: 05 abr. 2025.

APÊNDICES

Apêndice A – Algoritmo Floresta Aleatória.

#Carregar bibliotecas

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import classification_report, confusion_matrix, matthews_corrcoef

from imblearn.over_sampling import BorderlineSMOTE

Carregar a planilha enviada para análise

file_path = '/content/drive/MyDrive/Pré projeto/Dados_Dissertacao_Gabriella/Dados_Lab_CP_6var.xlsx'

dados = pd.read_excel(file_path)

Exibir as primeiras linhas e informações gerais para entender a estrutura do arquivo

print(dados.head())

print(dados.info())

Tratar valores ausentes na coluna 'AT' substituindo pela mediana

if 'AT' in dados.columns:

 dados['AT'] = dados['AT'].fillna(dados['AT'].median())

Separar as variáveis preditoras (X) e a variável-alvo (y)

X = dados.drop(columns=['DAA']) # Todas as colunas exceto a classe

y = dados['DAA']

Dividir os dados em treino (70%) e teste (30%)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)

```
# Aplicar balanceamento das classes no conjunto de treino
```

```
smote = BorderlineSMOTE(random_state=42)
```

```
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

```
# Padronizar os dados
```

```
scaler = StandardScaler()
```

```
X_train_scaled = scaler.fit_transform(X_train_resampled)
```

```
X_test_scaled = scaler.transform(X_test)
```

```
# Criar e treinar o modelo Random Forest
```

```
modelo_rf = RandomForestClassifier(n_estimators=100, random_state=42)
```

```
modelo_rf.fit(X_train_scaled, y_train_resampled)
```

```
# Fazer previsões nos dados de teste
```

```
y_pred = modelo_rf.predict(X_test_scaled)
```

```
# Gerar relatório de classificação
```

```
report = classification_report(y_test, y_pred, target_names=[str(cls) for cls in
sorted(y.unique())])
```

```
# Personalizar a exibição
```

```
print("Resultados do modelo:\n")
```

```
print("=" * 30)
```

```
print(report)
```

```
print("=" * 30)
```

```
# Plotar matriz de confusão
```

```
cm = confusion_matrix(y_test, y_pred)
```

```
plt.figure(figsize=(8, 6))
```

```
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=sorted(y.unique()),
yticklabels=sorted(y.unique()))
```

```
plt.xlabel('Predito')
```

```
plt.ylabel('Real')
```

```
plt.title('Matriz de Confusão')
```

```
plt.show()
```

```
# Calcular e exibir o Coeficiente de Correlação de Matthews (MCC)
```

```
mcc = matthews_corrcoef(y_test, y_pred)
```

```
print(f"Coeficiente de Correlação de Matthews (MCC): {mcc:.4f}")
```