



Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul



PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DOS MATERIAIS

Sexagem de Aves Utilizando Espectroscopia no Infravermelho por Transformada de Fourier e *Machine Learning*

SILVANO DIAS PEREIRA NAVES

Orientador: Prof. Dr. Bruno Spolon Marangoni

Coorientador: Prof. Dr. Cícero Rafael Cena da Silva

Tese apresentada à Universidade Federal de Mato Grosso do Sul - UFMS – Instituto de Física - INFI, para obtenção do título de Doutor em Ciência dos Materiais.

Campo Grande – MS
Maio / 2026

INSTITUTO DE FÍSICA- INFI
Cidade Universitária | Unidade 5 |
Fone 67 3345 7485
79070-900 | Campo Grande | MS

RESUMO

A identificação precoce do sexo em aves desempenha um papel crucial tanto na reprodução, criação e comercialização dessas espécies quanto na produção de conhecimento científico. A técnica molecular baseada em DNA tem sido a principal abordagem para determinação do gênero, destacando-se por sua precisão e ausência de riscos para a ave, no entanto, enfrenta desafios relacionados à alta demanda de tempo e custos, além da necessidade de laboratórios especializados. O presente trabalho investiga a aplicação da espectroscopia no Infravermelho por Transformada de Fourier (FTIR) associada a algoritmos de aprendizado de máquina como uma técnica não invasiva e precisa para a determinação precoce do sexo de aves, através de medidas feitas em uma região da pena conhecida com o vexilo. A pesquisa foca em quatro espécies: *Oryzoborus maximiliani* (Bicudo), *Nymphicus hollandicus* (Calopsita), *Oryzoborus angolensis* (Curió) e *Psittacula krameri* (Ring Neck). A metodologia envolveu a coleta de espectros FTIR de penas de pássaros jovens (com sexagem confirmada por análise de DNA) utilizando o acessório de Refletância Total Atenuada (ATR). Os espectros obtidos no intervalo de 3800 a 800 cm^{-1} foram tratados por SNV (*Standard Normal Variate*) e submetidos à Análise de Componentes Principais (PCA) para reduzir dimensionalidade e identificar variáveis significativas. A combinação de FTIR associado ao classificador de aprendizado de máquina *Random Forest*, mostrou resultados promissores. Foram obtidas acurácias significativas na identificação do sexo das aves, com destaque para Bicudo e Curió (94,4%), Calopsita (77,8%) e Ring Neck (72,2%). Dessa forma, o estudo mostra grande potencial de aplicação para sexagem de aves e indica a necessidade de estudos adicionais para permitir sua aplicação generalizada em determinadas espécies.

Palavras-chave: Sexagem de Aves, FTIR, Aprendizado de Máquina, Identificação Precoce de Gênero, Método Não-Invasivo.

ABSTRACT

Early identification of sex in birds plays a crucial role in reproduction, breeding, commercialization, and scientific knowledge production. DNA-based molecular techniques have been the primary approach for sex determination, distinguished by their accuracy and risk-free nature for the bird, but they face challenges related to time consumption, costs, and the need for specialized laboratories. This study investigates the application of Fourier Transform Infrared Spectroscopy (FTIR) combined with machine learning algorithms as a non-invasive and precise technique for early sex determination in birds, through measurements made in a feather region known as the vexil. The research focuses on four species: *Oryzoborus maximiliani* (*Bicudo*), *Nymphicus hollandicus* (*Cockatiel*), *Oryzoborus angolensis* (*Curió*), and *Psittacula krameri* (Ring Neck). The methodology involved collecting feather FTIR spectra from young birds (sex confirmed by DNA analysis) using the Attenuated Total Reflectance (ATR) accessory. Spectra obtained in the range of 3800 to 800 cm^{-1} were processed using Standard Normal Variate (SNV) and subjected to Principal Component Analysis (PCA) to reduce dimensionality and identify significant variables. The combination of FTIR with the Random Forest classifier demonstrated promising results. Significant accuracies were obtained in the identification of bird sex, with emphasis on *Bicudo* and *Curió* (94.4%), *Cockatiel* (77,8%) and Ring Neck (72,2%). In this way, the study shows great potential for application in avian sexing and indicates the need for additional studies to allow its generalized application in certain species.

Keywords: Bird Sexing, FTIR, Machine Learning, Early Gender Identification, Non-Invasive Method.

LISTA DE FIGURAS

Figura 1: Nomes das partes principais de uma pena.....	16
Figura 2: Modos Vibracionais simétricos e assimétricos.....	19
Figura 3: Interferômetro de Michelson.	20
Figura 4: Refletância Total Atenuada (ATR).	21
Figura 5: Validação Cruzada do tipo LOOCV, onde o número de dobras é igual a quantidade de amostras.	28
Figura 6: Matriz de Confusão para as Classes "Positivo" e "Negativo".	28
Figura 7: Média dos espectros FTIR de cada espécie.....	38
Figura 8: Média e desvio padrão dos espectros FTIR separados por sexo para o pássaro Bicudo.	39
Figura 9: Scores e Loadings para o pássaro Bicudo (3800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.	40
Figura 10: Acurácias em função da quantidade de PC's e árvores para o pássaro Bicudo (3800 a 800 cm^{-1}).	40
Figura 11: Acurácias em função da quantidade de PC's e árvores para o pássaro Bicudo (3000 a 2800 cm^{-1}).	41
Figura 12: Acurácias em função da quantidade de PC's e árvores para o pássaro Bicudo (1800 a 800 cm^{-1}).	41
Figura 13: Scores e Loadings para o pássaro Bicudo (3000 a 2800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.....	42
Figura 14: Matrizes de confusão de validação LOOCV (a) e teste (b) para o pássaro Bicudo (3000 a 2800 cm^{-1}).	43

Figura 15: Média e desvio padrão dos espectros FTIR separados por sexo para o pássaro Calopsita.....	44
Figura 16: Scores e Loadings para o pássaro Calopsita (3800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.....	45
Figura 17: Acurácias em função da quantidade de PC's e árvores para o pássaro Calopsita (3800 a 800 cm^{-1}).	45
Figura 18: Acurácias em função da quantidade de PC's e árvores para o pássaro Calopsita (3000 a 2800 cm^{-1}).	46
Figura 19: Acurácias em função da quantidade de PC's e árvores para o pássaro Calopsita (1800 a 800 cm^{-1}).	46
Figura 20: Scores e Loadings para o pássaro Calopsita (3000 a 2800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.....	47
Figura 21: Matrizes de confusão de validação LOOCV (a) e teste (b) para o pássaro Calopsita (3000 a 2800 cm^{-1}).	47
Figura 22: Média e desvio padrão dos espectros FTIR separados por sexo para o pássaro Curió.	49
Figura 23: Scores e Loadings para o pássaro Curió (3800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.	50
Figura 24: Acurácias em função da quantidade de PC's e árvores para o pássaro Curió (3800 a 800 cm^{-1}).	50
Figura 25: Acurácias em função da quantidade de PC's e árvores para o pássaro Curió (3000 a 2800 cm^{-1}).	51
Figura 26: Acurácias em função da quantidade de PC's e árvores para o pássaro Curió (1800 a 800 cm^{-1}).	51

Figura 27: Scores e Loadings para o pássaro Curió (1800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.	52
Figura 28: Matrizes de confusão de validação LOOCV (a) e teste (b) para o pássaro Curió (1800 a 800 cm^{-1}).	52
Figura 29: Média e desvio padrão dos espectros FTIR separados por sexo para o pássaro Ring Neck.	54
Figura 30: Scores e Loadings para o pássaro Ring Neck (3800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.....	55
Figura 31: Acurácias em função da quantidade de PC's e árvores para o pássaro Ring Neck (3800 a 800 cm^{-1}).	55
Figura 32: Acurácias em função da quantidade de PC's e árvores para o pássaro Ring Neck (3000 a 2800 cm^{-1}).	56
Figura 33: Acurácias em função da quantidade de PC's e árvores para o pássaro Ring Neck (1800 a 800 cm^{-1}).	56
Figura 34: Scores e Loadings para o pássaro Ring Neck (3000 a 2800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.....	57
Figura 35: Matrizes de confusão de validação LOOCV (a) e teste (b) para o pássaro Ring Neck (3000 a 2800 cm^{-1}).	57

LISTA DE TABELAS

Tabela 1: Atribuição de picos e seus modos vibracionais em espectros FTIR de vexilos de penas.	38
Tabela 2: Resumo dos resultados.	58

LISTA DE ABREVIATURAS E SIGLAS

- ATR – Refletância Total Atenuada (*Attenuated Total Reflectance*)
- DNA – Ácido Desoxirribonucleico (*Deoxyribonucleic Acid*)
- FIR – Infravermelho distante (*Far Infrared*)
- FN – Falso Negativo (*False Negative*)
- FP – Falso Positivo (*False Positive*)
- FTIR – Infravermelho por Transformada de Fourier (*Fourier Transform Infrared Spectroscopy*)
- HCA – Análise Hierárquica de Cluster (*Hierarchical Cluster Analysis*)
- IA – Inteligência Artificial (*Artificial Intelligence*)
- LAMP – Amplificação Isotérmica Mediada por Alça (*Loop-Mediated Isothermal Amplification*)
- LDA – Análise Discriminante Linear (*Linear Discriminant Analysis*)
- LOOCV – Validação Cruzada *Leave-One-Out* (*Leave-One-Out Cross-Validation*)
- MIR – Infravermelho Médio (*Middle Infrared*)
- MS – Mato Grosso do Sul
- NIR – Espectroscopia no Infravermelho Próximo (*Near-Infrared Spectroscopy*)
- PCA – Análise de Componentes Principais (*Principal Component Analysis*)
- PC – Componente Principal (*Principal Component*)
- PC's – Componentes Principais (*Principal Components*)
- PCR – Reação em Cadeia da Polimerase (*Polymerase Chain Reaction*)
- RIA – Radioimunoensaio (*Radioimmunoassay*)
- ROC – Característica de Operação do Receptor (*Receiver Operating Characteristic*)
- SeZn – Seleneto de Zinco (*Zinc Selenide*)
- SVM – Máquina de Vetores de Suporte (*Support Vector Machine*)
- SNV – Variação Normal Padrão (*Standard Normal Variate*)
- TN – Verdadeiro Negativo (*True Negative*)
- TP – Verdadeiro Positivo (*True Positive*)
- UV – Ultravioleta (*Ultraviolet*)

SUMÁRIO

1	INTRODUÇÃO.....	9
2	OBJETIVOS.....	12
2.1	Objetivo Geral.....	12
2.2	Objetivos Específicos.....	12
3	REVISÃO BIBLIOGRÁFICA.....	13
3.1	Sexagem de Aves.....	13
3.2	Espectroscopia no Infravermelho por Transformada de Fourier (FTIR).....	17
3.3	Análise de Componentes Principais (PCA).....	22
3.4	Aprendizado de Máquina.....	25
3.5	Florestas Aleatórias (<i>Random Forests</i>).....	30
4	METODOLOGIA.....	34
4.1	Coleta e Preparação das Amostras.....	34
4.2	Procedimento para realização das medidas.....	34
4.3	Medidas por FTIR.....	34
4.4	Pré processamento e Análise de Componentes Principais.....	35
4.5	Aprendizado de Máquina (<i>Machine Learning</i>).....	36
5	RESULTADOS E DISCUSSÃO.....	38
5.1	<i>Oryzoborus maximiliani</i> (Bicudo).....	39
5.2	<i>Nymphicus hollandicus</i> (Calopsita).....	44
5.3	<i>Oryzoborus angolensis</i> (Curió).....	49
5.4	<i>Psittacula krameri</i> (Ring Neck).....	54
6	CONCLUSÃO.....	59
7	SUGESTÕES DE TRABALHOS FUTUROS.....	61
	REFERÊNCIAS BIBLIOGRÁFICAS.....	62

1 INTRODUÇÃO

A identificação precoce do sexo em aves é amplamente reconhecida por seus benefícios diretos para a reprodução, criação e comercialização dessas espécies. Contudo, sob a perspectiva da produção de conhecimento, essa prática também desempenha um papel fundamental em áreas cruciais, como a preservação de espécies ameaçadas, o estudo de doenças relacionadas ao sexo, o melhoramento genético e outras iniciativas que demandam soluções inovadoras para atender às crescentes necessidades do mercado (1,2).

A técnica de referência utilizada por autores para fornecer a determinação de gênero aviário é a técnica molecular de análise do DNA (ácido desoxirribonucleico). Embora seja considerada uma técnica precisa, confiável e que não apresenta risco de danos a ave, se demonstra, onerosa, demorada e muitas vezes demanda a necessidade de realização em laboratórios especializados(3,4,5). Além desta técnica, outros métodos de identificação da sexagem de aves podem ser realizados por análises como: endoscopia, laparoscopia, citogenética, análise morfológica (formas, cores, tamanhos), análise de comportamento e o dimorfismo sexual. No entanto, estes, por muitas vezes apresentam restrições, tais como incertezas ou necessidade de que atinjam a maturidade sexual, dentre outros fatores limitantes (6,7).

Com o avanço das tecnologias e o surgimento de técnicas automatizadas, vieram novas formas de se identificar precocemente a sexagem de algumas espécies de aves, que passaram a não se restringir a métodos demorados, de custos elevados, limitadores e por vezes incertos e invasivos. Estudos realizados nesta última década apontam métodos modernos, rápidos, eficazes e menos onerosos que se utilizaram da espectroscopia de Infravermelho por Transformada de Fourier (FTIR) na sexagem de algumas espécies de aves (8,9,10).

A espectroscopia FTIR foi utilizada para analisar penas de contorno em crescimento de filhotes de peru, onde foram observadas variações espectrais na faixa de 1000 a 1250 cm^{-1} . A classificação foi realizada por um método não supervisionado PCA (*Principal Component Analysis*). A avaliação dos primeiros e terceiros componentes principais permitiu a classificação dos filhotes em machos e fêmeas com uma precisão superior a 95% (9).

Em outro trabalho, foi demonstrado que apenas uma pequena quantidade de polpa de penas é suficiente para determinar o sexo de pombos. Espectros de FTIR submetidas à PCA e ao classificador LDA (*Linear Discriminant Analysis*) obtiveram informações relacionadas ao sexo

descritas pelas componentes principais 2 e 4. A componente principal 2 (PC2) refletiu diferentes quantidades de proteínas, enquanto o PC4 revelou variações nas bandas de amida I e amida II (1544 cm^{-1} e 1682 cm^{-1}), além da região de vibrações de fosfato de ácidos nucleicos (1042 cm^{-1} e 1222 cm^{-1}). As amostras de polpa de penas dos pombos machos foram classificadas com 100% de acurácia (10).

Os resultados indicam que a utilização dessa técnica apresenta potencial para identificação precoce na sexagem com precisão e rapidez. No entanto, nem todos os métodos ou técnicas podem ser aplicados a todas as espécies de aves, e muitas vezes, sozinhos, não são suficientes para determinar com eficácia o sexo de algumas espécies. Assim, há a necessidade de desenvolvimento de novos métodos que superem as limitações das técnicas atuais e já consolidadas.

Considerando os avanços tecnológicos que estão em constante evolução e se renovam em um ritmo acelerado, e atendendo ao advento da Inteligência Artificial (IA), que surge como um sistema especialista, inteligente, baseado em conhecimento e aprendizado contínuo (11,12), as empresas buscam incessantemente o aprimoramento para automação, eficiência e uma melhor compreensão de seus dados, além de tomadas de decisões cada vez mais complexas.

Dessa forma, destacando a relevância da temática e a importância de poder identificar precocemente a sexagem de aves de maneira a se adaptar a fenômenos e práticas mais dinâmicas, além de oferecer uma alternativa aos métodos já estabelecidos, este trabalho tem como objetivo investigar a viabilidade da espectroscopia FTIR associada ao aprendizado de máquina para a sexagem de aves. Essa técnica, apesar de ser amplamente utilizada em diversas áreas da ciência para identificação e análise de materiais, não é uma das técnicas mais comumente utilizadas na sexagem de aves. Utilizou-se dela por ser uma técnica analítica que permite a identificação de moléculas com base em suas assinaturas vibracionais espectrais em determinados comprimentos de onda do espectro e foi realizada no intervalo espectral do infravermelho médio (MIR) com o uso do acessório de Refletância Total Atenuada (ATR) no cristal de seleneto de zinco (SeZn).

Estudos anteriores já utilizaram a técnica FTIR na análise da sexagem de aves tais como perus, frangos e alguns tipos de pássaros (9,10), obtendo resultados a partir dela por si só ou da conjunção de outras técnicas. Uma das principais diferenciações deste estudo é a análise a partir da região da pena conhecida como vexilo. Este fator representa uma originalidade em nosso

trabalho, pois, diferentemente da maior parte dos estudos anteriores, especialmente internacionais, que se baseiam na análise do cálcio, por conter células do folículo da pena e do epitélio fornecendo informações genéticas.

Para alcançar o objetivo geral deste estudo, foram definidos os seguintes objetivos específicos: obter os espectros FTIR das penas de ambos os sexos de aves das espécies: *Oryzoborus maximiliani* (Bicudo), *Nymphicus hollandicus* (Calopsita), *Oryzoborus angolensis* (Curió) e *Psittacula krameri* (Ring Neck); realizar a análise não supervisionada, utilizando PCA, para avaliar as principais contribuições para a separação dos sexos; e aplicar o aprendizado de máquina para analisar os dados resultantes do PCA, automatizando a classificação.

A escolha por este estudo foi motivada pela escassez de pesquisas na literatura relacionadas à sexagem precoce de aves das espécies citadas, além do crescente interesse da indústria avícola na melhoria genética para diversos fins. Destaca-se, ainda, a importância na preservação da espécie, bem como o crescimento do turismo ornitológico e o crescente envolvimento da indústria em concursos de aves no Brasil e em nível internacional, como os renomados eventos no sul da Ásia, na Espanha, entre outros (13).

Os espectros obtidos foram normalizados utilizando o método SNV (*Standard Normal Variate*) e submetidos à PCA. O potencial de agrupamento das amostras foi avaliado utilizando diferentes intervalos. Por fim, os *scores* obtidos pelo PCA foram utilizados para o treinamento do classificador *Random Forest*, testando diferentes parâmetros através do método de validação LOOCV (*Leave-One-Out Cross-Validation*). Foram observados e respeitados os cuidados com a procedência, armazenamento, acondicionamento e manuseio das amostras. Destaca-se que, inicialmente, o sexo de todas as amostras foi identificado pela técnica de referência (análise de DNA) para possibilitar um confronto rigoroso dos resultados ao final do estudo proposto.

Neste trabalho, apresentaremos no Capítulo 3 a revisão bibliográfica. No Capítulo 4 a metodologia do estudo, descrevendo sobre as amostras; os procedimentos para realização das medidas; a caracterização dos dados espectrais e, na sequência, a análise multivariada realizada. No Capítulo 5, serão expostos os resultados e discussões e, no Capítulo 6, a conclusão do estudo, bem como as contribuições para a academia, para a indústria avícola e para a pesquisa científica.

2 OBJETIVOS

2.1 Objetivo Geral

Investigar a aplicação da espectroscopia óptica no Infravermelho por Transformada de Fourier (FTIR), associada a algoritmos de aprendizado de máquina, como método não invasivo e preciso para a determinação precoce do sexo de aves.

2.2 Objetivos Específicos

- Obter os espectros FTIR do vaxilo das penas de ambos os sexos das aves das espécies: *Oryzoborus maximiliani* (Bicudo), *Nymphicus hollandicus* (Calopsita), *Oryzoborus angolensis* (Curió) e *Psittacula krameri* (Ring Neck);
- Avaliar qualitativamente os espectros por Análise de Componentes Principais (PCA), reduzindo a dimensionalidade dos dados e identificando o intervalo espectral mais adequado para a formação de grupos com base nos sexos masculino e feminino. Comparar as bandas espectrais que mais contribuem para a separação dos grupos em cada espécie de ave;
- Avaliar quantitativamente o desempenho do uso da espectroscopia FTIR associada ao aprendizado de máquina na sexagem de aves, utilizando os *scores* obtidos por PCA no intervalo espectral mais promissor;
- Avaliar a acurácia da aplicação do Random Forest para treinamento e classificação da sexagem.

3 REVISÃO BIBLIOGRÁFICA

3.1 Sexagem de Aves

O termo sexagem se refere ao ato de determinar se um indivíduo de determinada espécie é fêmea ou macho (14). A determinação da sexagem de indivíduos da mesma espécie pode ser realizada de diferentes formas, desde uma simples inspeção visual quando as diferenças são mais evidentes até técnicas avançadas em espécies onde a sexagem é de difícil determinação. Diversas metodologias podem ser empregadas para a sexagem de indivíduos da mesma espécie, estas técnicas podem incluir estratégias baseadas na morfologia do animal, como por exemplo medições corporais e laparoscopia (7,15). Além disso, a morfologia do indivíduo pode ser investigada por exames de imagem, como a tomografia por ressonância magnética nuclear (16) e a ultrassonografia em tempo real (17).

Recentemente, a sexagem de animais, especialmente a sexagem de aves, vem evoluindo com o emprego de técnicas avançadas, tanto para interesse industrial, como as aves de corte (com a determinação da sexagem *in ovo*), como para animais domésticos. A inovação permite a evolução no manejo, na tomada de decisões, na conservação e na pesquisa (18,19). Em espécies de aves sem dimorfismo sexual evidente e criadas em cativeiro, estratégias de protocolos de sexagem seguros e eficazes são fundamentais para a adequada reprodução e o comércio legal de aves (20). Além disso, esse procedimento é essencial para a pesquisa e centros de conservação, permitindo a reprodução em cativeiro, garantindo a formação de casais heterossexuais, para posterior soltura ou reintrodução na natureza (21).

Para atender a essa demanda, nos últimos anos diferentes técnicas de sexagem de aves foram desenvolvidas, testadas e implementadas, beneficiando diretamente a conservação e as pesquisas ornitológicas. Essas técnicas incluem a já citada laparoscopia, além da citogenética, tomografia por ressonância magnética, radioimunoensaio (RIA) e LAMP (Loop-Mediated Isothermal Amplification) (22,23,24,25).

No entanto, essas metodologias acabam apresentando desvantagens, como o elevado risco de dano físico ao animal, o alto custo dos equipamentos, bem como a sua manutenção, a necessidade de equipe técnica especializada e treinada, além das dificuldades na aplicação desses métodos em larga escala, mais especificamente em campo, e sua dificuldade de

assertividade em indivíduos filhotes ou imaturos (23,24). Assim, as técnicas mais utilizadas atualmente baseiam-se na extração e avaliação de material genético, na hibridização de regiões específicas e em amplificações de fragmentos de material genético via PCR (*Polymerase Chain Reaction*, ou Reação em Cadeia da Polimerase) (26,10). O método PCR destaca-se dentre os demais pela sua maior eficácia, rapidez e precisão, além de não representar riscos biológicos para os indivíduos avaliados, por se tratar de técnica minimamente invasiva (26).

No entanto, apesar de sua eficiência, a técnica PCR exige infraestrutura laboratorial, equipamentos apropriados e profissionais qualificados, o que acaba por consequência, fazendo que criadores e conservacionistas dependam de empresas especializadas, que geralmente estão localizadas em grandes centros urbanos o que atrasa e encarece as análises. Frequentemente, as amostras de penas precisam ser enviadas por correios para a realização da sexagem nos grandes laboratórios, o que representa um grande obstáculo para atividades conduzidas no campo e no interior do país. Assim, é fundamental o desenvolvimento de tecnologias mais acessíveis e práticas, que facilitem o trabalho de pesquisadores e produtores e possam promover maior eficiência na sexagem de animais.

Recentemente, técnicas não invasivas baseadas em espectroscopia óptica e aprendizado de máquina (*Machine Learning*) surgiram como alternativas eficazes para sexagem de aves através de suas penas. He et al. (27) utilizaram espectroscopia no Ultra Violeta (UV) e aprendizado profundo para a sexagem de aves canoras baseado no reflexo das suas penas. O estudo demonstrou que padrões de refletância espectral diferem entre machos e fêmeas, permitindo uma distinção precisa. O estudo indica que diferentes espécies de aves podem conter informação espectral de sua sexagem em suas penas, principalmente em idade adulta.

Mehdizadeh et al. (28) aplicaram espectroscopia no infravermelho próximo (NIR *Near-Infrared Spectroscopy*) e visão computacional para a sexagem de codornas com base nas características da sua plumagem. O estudo utilizou o classificador Máquina de Vetores de Suporte (SVM - *Support Vector Machine*), alcançando 100% de precisão na classificação dos machos e 92% na identificação das fêmeas. Esses resultados indicam a eficiência na sexagem de aves através de métodos não invasivos, utilizando apenas as propriedades ópticas das penas, que permitem a construção de um protocolo prático e eficiente. A alta acurácia no estudo desde os

primeiros dias de vida das aves indicam a possibilidade de sexagem em aves filhotes através de suas penas e espectroscopia vibracional associada com aprendizado de máquina.

Harz et al. (29) relataram a aplicação da espectroscopia Raman com ressonância em UV como um método minimamente invasivo para a determinação do sexo das aves, por meio da avaliação de amostras da polpa extraída das penas. A sexagem foi realizada por SVM, atingindo acurácia de 95%. Além disso, a análise de componentes principais (PCA - *Principal Component Analysis*) também foi aplicada para determinar as regiões espectrais importantes para a classificação dos espectros Raman de amostras da polpa de penas de frangos.

Steiner et al. (10) utilizaram espectroscopia no infravermelho por transformada de Fourier (FTIR - *Fourier Transform Infrared Spectroscopy*) para sexagem de pombos domésticos através da polpa das suas penas. Utilizando imagens espectroscópicas de FTIR, filtradas por PCA e a classificadas com análise discriminante linear (LDA - *Linear Discriminant Analysis*), as amostras de polpa de pena dos machos foram classificadas com 100% de precisão. Enquanto sete das oito amostras de fêmeas foram classificadas corretamente, demonstrando a eficiência das técnicas espectroscópicas na sexagem de aves a partir de suas penas.

Além disso, Steiner et al. (9) utilizaram a microespectroscopia FTIR para analisar a composição molecular das células germinativas e realizar a sexagem de filhotes de peru, a partir da polpa das suas penas. A classificação espectral foi realizada por um método não supervisionado baseado em PCA. A avaliação da primeira e da terceira componentes principais permitiram a sexagem com uma precisão superior a 95%. Reforçando a capacidade da técnica espectroscópica na sexagem de aves filhotes.

Steiner et al. (8) demonstraram que a espectroscopia FTIR é uma ferramenta adequada para a determinação rápida e objetiva do sexo das aves. O método baseia-se nas diferenças no tamanho dos cromossomos. A partir de espectros obtidos de células germinativas extraídas da polpa da pena de filhotes e do disco germinativo de ovos fertilizados (não incubados) foi possível identificar alterações significativas entre células de aves machos e fêmeas principalmente na região das vibrações de fosfato, em torno de 1080 e 1120 cm^{-1} , reforçando a sensibilidade da técnica para determinação da sexagem de aves ainda filhotes.

Com o objetivo de desenvolver uma técnica eficiente e ainda mais prática, neste trabalho as penas de aves filhotes foram analisadas por espectroscopia FTIR na região conhecida como vexilo (Figura 1), ao invés da polpa das penas, com o objetivo de identificar variações na composição e nas estruturas das penas que permitam a sexagem de aves de forma rápida e eficiente, utilizando o classificador florestas aleatórias (*Random Forests*) removendo variáveis correlacionadas pelo uso do PCA.

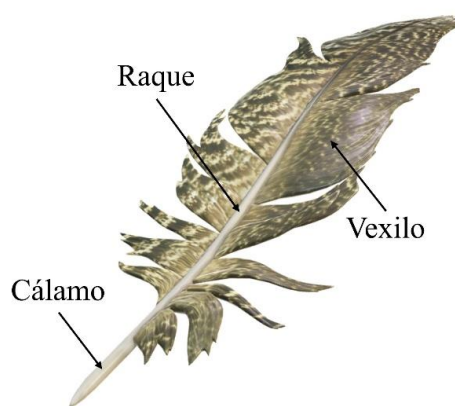


Figura 1: Nomes das partes principais de uma pena.
Fonte: O autor, 2025.

3.2 Espectroscopia no Infravermelho por Transformada de Fourier (FTIR)

A espectroscopia é técnica de caracterização de materiais e uma ciência que estuda a interação entre a radiação eletromagnética e a matéria. Este campo científico permite compreender como átomos e moléculas se comportam em contato com a radiação eletromagnética em diferentes comprimentos onda, avaliando se as moléculas absorvem, emitem ou dispersam energia (30,31). Através da análise das transições energéticas das moléculas em estudo é possível obter representações gráficas dessas interações, que fornecem informações específicas sobre transições energéticas e consequentemente características físicas e químicas dos materiais (32,31).

As transições energéticas dos sistemas estudados estão diretamente relacionadas com radiação eletromagnética incidente. No caso das moléculas, ou seja, na espectroscopia molecular, as transições podem ser classificadas em: transições eletrônicas, vibracionais e rotacionais. As transições eletrônicas correspondem à movimentação de elétrons para estados de energia superiores e predominam nas regiões do ultravioleta e parte do espectro visível, estão relacionadas com os orbitais moleculares. Já as transições vibracionais estão associadas às oscilações das ligações químicas e se manifestam nas regiões do infravermelho e parte do espectro visível. As transições rotacionais dizem respeito à rotação das moléculas ao redor de seus próprios eixos, sendo observadas na faixa do micro-ondas (33,31). Cada técnica espectroscópica emprega diferentes níveis de energia e depende de dispositivos específicos para emitir a luz em diferentes comprimentos de onda e detectar as radiações transmitidas ou absorvidas.

A radiação no infravermelho foi identificada pela primeira vez no início do século XIX por William Herschel, utilizando termômetros de mercúrio para medir a temperatura de diferentes cores da luz visível refratadas por um prisma. Neste experimento, observou-se que a temperatura aumentava além da região da cor vermelha, onde não havia luz perceptível ao olho humano, por isso o nome infravermelho (abaixo do vermelho). Os resultados revelaram a existência de uma radiação invisível, impulsionando os estudos sobre a interação da radiação infravermelha com a matéria. (34,35).

O espectro eletromagnético é classificado de acordo com as diferentes regiões de comprimentos de onda. O infravermelho abrange comprimentos de onda entre 0,78 e 1000 μm , sendo subdividida em três seções principais: o infravermelho próximo (NIR, *Near Infrared*), que vai de 0,78 a 2,5 μm (12800 - 4000 cm^{-1}); o infravermelho médio (MIR, *Middle Infrared*), de 2,5 a 50 μm (4000 - 200 cm^{-1}); e o infravermelho distante (FIR, *Far Infrared*), que compreende a faixa entre 50 e 1000 μm (200 - 10 cm^{-1}) (33,36). No presente estudo, foram utilizadas técnicas de caracterização no infravermelho médio, amplamente empregadas na caracterização de materiais.

No infravermelho, a absorção ou emissão de radiação das moléculas ocorre devido às variações periódicas no momento de dipolo elétrico, que representam a separação entre cargas positivas e negativas da eletrosfera. A radiação absorvida ou emitida possui frequência equivalente à frequência da oscilação do dipolo (35,31). Consequentemente, moléculas diatômicas homonucleares, como H_2 e Cl_2 não apresentam absorção no infravermelho devido à ausência de variação no momento de dipolo. Em contrapartida, moléculas diatômicas heteronucleares, sofrem variações de dipolo ao vibrar, tornando-se ativas nessa região do espectro.

A forma com que as moléculas vibram depende das maneiras pelas quais seus átomos podem se movimentar, definindo seus graus de liberdade. Para uma molécula com N átomos, há 3N modos diferentes de oscilação, incluindo deslocamentos de translação, rotações e vibrações. As translações são deslocamentos da molécula inteira, as rotações são movimentos de rotação em torno de eixos específicos e as vibrações são as oscilações dos átomos em torno de suas ligações químicas em posições de equilíbrio. Moléculas não lineares possuem 3N - 6 graus de liberdade vibracionais, enquanto moléculas lineares apresentam 3N - 5, pois não ocorre rotação ao longo do eixo molecular (36).

Os tipos de vibração molecular são chamados de modos vibracionais e podem ser classificados em dois tipos: estiramento (*stretching*) e deformação angular (*bending*). O estiramento ocorre quando os átomos se movem (aproximando ou afastando) ao longo do eixo da ligação química. A deformação angular são alterações nos ângulos das ligações químicas, podendo apresentar movimentos de flexão. Os modos vibracionais podem ser simétricos, quando

os átomos se deslocam de forma semelhante, ou assimétricos, quando os deslocamentos são invertidos (36).

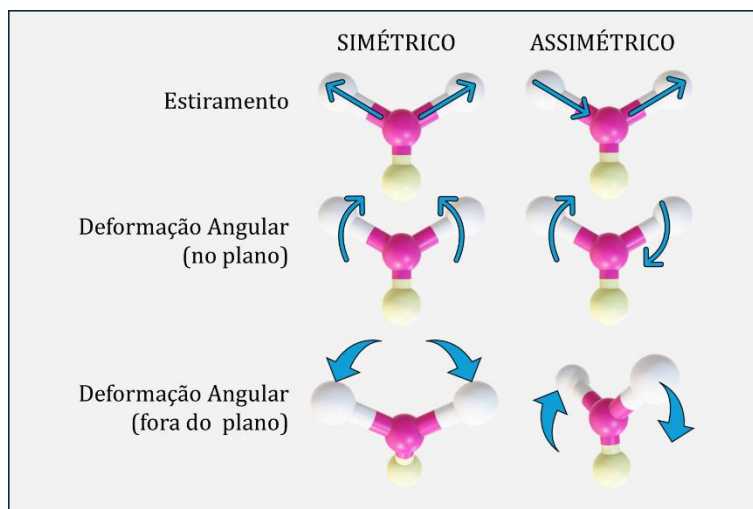


Figura 2: Modos Vibracionais simétricos e assimétricos.
Fonte: O autor, 2025.

Na espectroscopia de MIR e Raman, predominam as transições fundamentais, que envolvem a passagem do estado fundamental para o primeiro estado excitado, na espectroscopia NIR, ocorrem sobretons e combinações de transições vibracionais, resultando em estados energéticos mais elevados (36).

Na espectroscopia MIR a técnica mais eficiente para aquisição de espectros é a utilização do interferômetro de Michelson associado ao tratamento de dados utilizando a operação matemática de transformada de Fourier. A técnica FTIR teve suas bases estabelecidas no final do século XIX, onde o físico Albert Abraham Michelson criou o interferômetro (em 1891) para realizar medidas de comprimentos muito pequenos com precisão. No ano seguinte, Lord Rayleigh demonstrou que o interferograma gerado pela técnica estava matematicamente relacionado ao espectro da radiação por meio da transformada de Fourier. No entanto, somente após 1960, com o avanço da computação, a técnica se tornou mais acessível (37).

Os espectrômetros FTIR modernos utilizam o interferômetro de Michelson, composto por uma fonte de radiação, um divisor de feixe (*beamsplitter*) e dois espelhos, um fixo e outro móvel. A radiação incidente é separada, refletida e recombinada, formando um interferograma que, ao ser processado por meio da transformada de Fourier, gera o espectro da amostra (38). Essa abordagem apresenta diversas vantagens em relação aos espectrômetros

dispersivos tradicionais, como maior rapidez na aquisição dos dados, melhor relação sinal-ruído, maior precisão na determinação dos comprimentos de onda, maior sensibilidade e um intervalo espectral mais amplo (37).

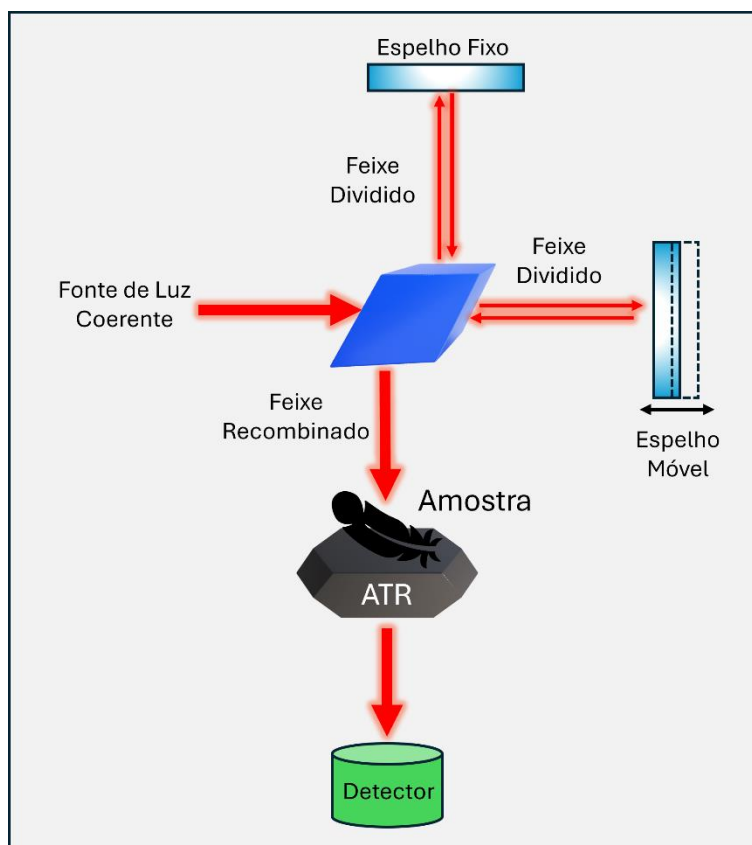


Figura 3: Interferômetro de Michelson.
Fonte: O autor, 2025.

Na espectroscopia FTIR, a técnica de reflectância total atenuada (ATR - *Attenuated Total Reflectance*) é amplamente utilizada para analisar substâncias sólidas e líquidas. Nesse método, um feixe de radiação infravermelha incide sobre um cristal de alto índice de refração, onde a amostra é depositada. A interação entre a amostra e a radiação ocorre por meio de uma onda evanescente, resultante das reflexões internas do cristal, que penetra superficialmente na amostra, permitindo a obtenção do espectro de absorção de maneira ampla e uniforme. Essa abordagem facilita a preparação das amostras e possibilita análises não destrutivas, permitindo a análise direta de sólidos, líquidos, pós e filmes finos, com maior reprodutibilidade e precisão (39).

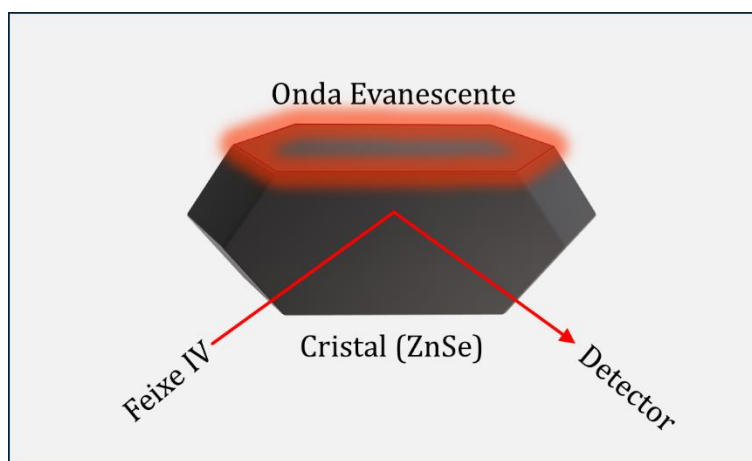


Figura 4: Refletância Total Atenuada (ATR).
Fonte: O autor, 2025.

No espectrômetro FTIR, é possível realizar múltiplas varreduras ou número de *scans*, somando os interferogramas obtidos em um processo conhecido como co-adição. Esse procedimento melhora a qualidade do espectro obtido (40,41). Outro aspecto essencial na análise FTIR é a resolução utilizada, que indica a capacidade do sistema de distinguir bandas próximas dentro do espectro. Uma maior resolução oferece a vantagem de um espectro mais detalhado, no entanto, exige um deslocamento mais extenso do espelho móvel, prolongando o tempo de aquisição o que pode resultar em uma aumento da captação de ruído adicional (42,43).

Assim, quando se configura no equipamento uma maior resolução espectral, o deslocamento do espelho móvel é maior, o que eleva o comprimento do interferograma e consequentemente aumenta o número de pontos de amostragem. O comprimento de deslocamento do espelho necessário para a resolução configurada, em geral em cm^{-1} , pode variar de acordo com a configuração (disposição de elementos) do espectrômetro e as características da radiação infravermelha utilizada (fonte). Durante a leitura, o *software* do espectrômetro aplica a transformada de Fourier, interpolando os dados do interferograma para ajustar a resolução do espectro de acordo com o incremento/passo de número de onda configurado na coleta (*i.e.*, $0,5 \text{ cm}^{-1}$).

3.3 Análise de Componentes Principais (PCA)

A análise multivariada é um campo interdisciplinar que integra estatística multivariada, modelagem matemática, com aplicações na ciência da computação e na química analítica (44,45). Essa abordagem permite examinar múltiplas variáveis simultaneamente, o que permite uma compreensão mais profunda e objetiva das relações entre elementos de múltiplas variáveis, proporcionando uma visão mais abrangente sobre um conjunto de dados.

Como resultado da aplicação da análise multivariada em dados de origem química, surge o termo quimiometria em 1971 para designar a aplicação de modelos matemáticos, estatística no contexto da química analítica (46). Tempos após seu surgimento, a disciplina se consolidou com o avanço da computação e conseqüentemente seu acesso, permitindo a implementação dessas técnicas em equipamentos laboratoriais (47,48).

A análise multivariada é diversificada e pode ser aplicada em diversas áreas. Além disso, suas técnicas são empregadas em diferentes etapas, desde a análise exploratória de dados, calibração multivariada, processamento de sinais analíticos, planejamento e otimização experimental, modelagem e monitoramento de processos, até mesmo em aplicações práticas do dia-a-dia utilizando inteligência artificial (46).

Técnicas exploratórias de análise de dados são essenciais para identificar padrões e tendências em conjuntos de dados com muitas variáveis. Elas auxiliam na simplificação das informações, tornando-as mais acessíveis e interpretáveis, além de detectar dados discrepantes (*outliers*) (46). Métodos como a Análise de Componentes Principais (PCA, *Principal Component Analysis*) e a Análise Hierárquica de Cluster (HCA, *Hierarchical Cluster Analysis*) são os exemplos mais comuns, amplamente utilizados para transformar grandes volumes de dados em representações mais intuitivas, reduzindo um conjunto de dados evidenciando suas informações mais relevantes (49,50).

A principal função do PCA é reduzir a dimensionalidade de um conjunto de dados, a partir da geração de combinações lineares das variáveis originais, criando variáveis, denominadas componentes principais (PC's – *Principal Components*). Cada componente principal é obtida de modo a capturar a maior variância possível dos dados originais, permitindo representar grandes conjuntos de dados de maneira simplificada sem comprometer informações

essenciais, o que permite a visualização gráfica das relações existentes entre indivíduos/amostras (45).

Matematicamente, os dados são organizados em uma matriz $\mathbf{X}$, composta por “n” amostras (linhas) e “m” variáveis (colunas). No processo de PCA, essa matriz é decomposta em duas outras matrizes: a matriz $\mathbf{T}$, que contém as projeções dos dados no espaço dos componentes principais (*scores*), e a matriz $\mathbf{P}$, que representa os pesos ou as contribuições de cada variável original para a variância de cada componente principal (*loadings*) (45).

A decomposição de $\mathbf{X}$ gera as PC's, que são novas variáveis, derivadas de combinações lineares das variáveis originais correlacionadas. O vetor $\mathbf{T}$ expressa a projeção das variáveis originais nessas novas componentes, permitindo identificar visualmente padrões e relações entre as amostras. Já o vetor $\mathbf{P}$ indica a influência de cada variável original na composição das componentes principais, geometricamente, representa os cossenos dos ângulos entre as componentes e os eixos das variáveis originais (45).

Geralmente, um pequeno número de componentes principais é suficiente para descrever a variância de um conjunto de dados espectroscópicos, facilitando a identificação de similaridades e diferenças entre as amostras. A interpretação dos resultados pode ser realizada por meio de gráficos de *scores*, que ilustram as relações entre as amostras e gráficos de *loadings*, que destacam quais variáveis influenciam as similaridades ou diferenças observadas.

No entanto, para um conjunto de dados obtido por espectroscopia molecular, antes de realizar o PCA, o conjunto de dados precisam ser uniformizado, minimizando eventuais artefatos introduzidos durante a leitura que podem afetar a qualidade das análises. Para isso, técnicas de pré-tratamento são frequentemente utilizadas, minimizando efeitos de deslocamento de linha de base e normalizando suas intensidades (51).

A dispersão da luz em meios sólidos é um fenômeno comum em técnicas analíticas, principalmente devido as partículas possuírem dimensões muito próximas aos comprimentos de onda utilizados. O método de pré-processamento SNV (*Standard Normal Variate*) é um dos métodos mais utilizados. O SNV corrige deslocamentos na linha de base (*offset*) e normaliza as intensidades de transmitância, subtraindo cada intensidade pela média do respectivo espectro, dividindo o valor obtido pelo desvio padrão desse mesmo espectro, conforme a Equação 1:

$$x_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \quad (1)$$

Onde x_{ij} representa o valor da intensidade de transmitância “j” no espectro “i”, μ_i é a média das intensidades do espectro “i” e σ_i é o seu respectivo desvio padrão. Dessa forma, os espectros são centralizados em torno da média (resultando em média zero para todos os espectros) e normalizados pelo desvio padrão (tornando-o igual a 1 para todos os espectros) (52).

3.4 Aprendizado de Máquina

Recentemente a inteligência artificial e mais especificamente o aprendizado de máquina ou *machine learning*, tem sido cada vez mais relevante e empregado em diversas áreas, refletindo um crescente interesse no uso de técnicas de análise de dados como ferramenta para identificar padrões e tornar processos mais eficientes. O aprendizado de máquina pode ser definido como um conjunto de técnicas matemáticas e estatísticas que possibilitam que recursos computacionais possam extrair padrões e realizar previsões a partir de um conjunto de dados, construindo modelos, que podem ser compreendidos como funções matemáticas que são capazes de tomar decisões com base em informações previamente processadas (53).

As estratégias para o uso de *machine learning* podem ser classificadas de forma geral em métodos supervisionados e não supervisionados. Os métodos não supervisionados são aqueles em que o algoritmo busca identificar padrões, estruturas ou agrupamentos nos dados sem a necessidade de rótulos ou classificações pré-definidas. Diferentemente do aprendizado supervisionado, no qual os modelos são treinados com exemplos rotulados, no aprendizado não supervisionado o algoritmo analisa os dados em busca de regularidades e correlações, sem distinguir previamente categorias ou classes. Um exemplo clássico dessa abordagem é o algoritmo K-Means, utilizado para tarefas de clusterização (54).

Já os métodos supervisionados são aqueles em que o conjunto de dados possui identificações ou rótulos associadas que tem relação direta com cada indivíduo ou amostra, permitindo que o modelo aprenda a relação entre as variáveis de entrada e a associem ela a uma saída desejada, ou seja, façam uma previsão. Os métodos supervisionados podem ainda ser divididos em duas categorias principais, modelos de classificação, que atribuem os dados a categorias predefinidas e modelos de regressão, que fazem previsões numéricas contínuas a partir dos padrões identificados nos dados de treinamento. (54).

Assim, os métodos supervisionados podem prever propriedades quantitativas (regressão) como a concentração de um analito ou qualitativas (classificação) como definir como “positivo” ou “negativo” uma amostra suspeita de alguma doença, sendo que neste trabalho utilizamos um método supervisionado chamado *Random Forests* para identificar a sexagem de aves por meio da classificação binária (macho ou fêmea).

O processo de criação de modelos supervisionados começa com a divisão dos dados em dois conjuntos: o conjunto de treinamento, usado para construir o modelo de aprendizado de máquina, e o conjunto de teste, que permite avaliar a capacidade do modelo de generalizar para novos dados e fazer previsões precisas. Não é recomendável utilizar todos os dados tanto para treinamento quanto para teste, pois isso pode levar à criação de modelos mal ajustados (55,56,57).

O ajuste inadequado pode ocorrer de duas formas. A primeira é quando o modelo não captura informações suficientes para se tornar robusto, ou seja, ele é incapaz de extrair características relevantes dos dados e não consegue diferenciar corretamente nem mesmo as classes de treino. Isso é conhecido como subajuste ou *underfitting*. A segunda forma ocorre quando o modelo é muito complexo e acaba armazenando as características dos dados de treinamento no modelo, apresentando um bom desempenho apenas com os dados de treino. No entanto, quando confrontado com dados novos, seu desempenho é insatisfatório. Esse fenômeno é chamado de sobreajuste ou *overfitting*. (58).

A divisão dos dados é feita de modo a garantir uma distribuição homogênea das classes nos conjuntos de treinamento e teste, ou seja, os dados são estratificados. Além disso, é fundamental que esses conjuntos representem adequadamente a variabilidade sistemática do problema, assegurando que, além das classes, características secundárias conhecidas também sejam distribuídas de forma equilibrada. Isso evita que fatores externos influenciem indevidamente o desempenho do modelo, garantindo uma avaliação mais precisa e realista da sua capacidade de generalização (55,56,57).

Essa divisão de grupos pode ser realizada utilizando algoritmos como os métodos Kennard-Stone, Duplex e redes neurais de Kohonen (59). Neste estudo, utilizou-se uma abordagem mais simples, porém eficiente, a amostragem aleatória estratificada por classe para a separação dos conjuntos de treinamento e teste.

O treinamento de um modelo envolve a otimização de uma função de perda, ajustando automaticamente seus parâmetros internos para minimizar os erros de previsão. Esse processo busca encontrar os melhores valores dos parâmetros do modelo para que ele generalize bem em novos dados (57).

Após o treinamento do modelo, sua capacidade de generalização é avaliada utilizando o conjunto de teste, separado no início do processo. Se o desempenho não for satisfatório, ajustes nos hiperparâmetros podem ser feitos para otimizar a performance do modelo. Esse processo, conhecido como otimização de hiperparâmetros, depende do tipo de classificador utilizado, pode incluir técnicas como ajuste da taxa de aprendizado, número de componentes principais (PCA) ou o uso de métodos de regularização, como Lasso e Ridge, para evitar *overfitting* (56).

Assim, os parâmetros de um modelo são ajustados automaticamente a partir dos dados de treinamento, já os hiperparâmetros, são definidos manualmente antes do treinamento e podem ser otimizados por técnicas como a busca em grade (*Grid Search*). A busca em grade é um método sistemático que testa várias combinações possíveis de um conjunto pré-definido de valores de hiperparâmetros, selecionando ao final o modelo ótimo, ou seja, aquele que resulta no melhor desempenho com base em alguma métrica de validação (60,57).

Durante o treinamento, um sorteio aleatório define quais amostras serão usadas como teste para validar seus parâmetros, processo conhecido como validação. Para um conjunto de amostras limitado uma estratégia amplamente utilizada é a validação cruzada, contribuindo para redução do *overfitting*. O método mais geral de validação cruzada é o *k-fold*, onde são divididos aleatoriamente em “k” subconjuntos (ou pastas) que são usados para validação enquanto os demais servem para o treinamento (55).

Uma alternativa ao *k-fold* é o método LOOCV (*Leave-One-Out Cross Validation*), no qual, ao invés de dividir os dados em múltiplos subconjuntos, apenas uma única amostra é separada para validação, enquanto as demais são usadas para treinamento. Esse processo se repete para cada amostra, garantindo que todas sejam utilizadas tanto no treinamento quanto na validação. O LOOCV é especialmente útil para conjuntos pequenos de amostras, pois maximiza os dados usados no treinamento, embora seja computacionalmente mais exigente. Além disso, ele reduz a variabilidade nos subconjuntos de treinamento e teste, o que pode ser vantajoso em processos de ajuste de hiperparâmetros, como o *Grid Search* (55).

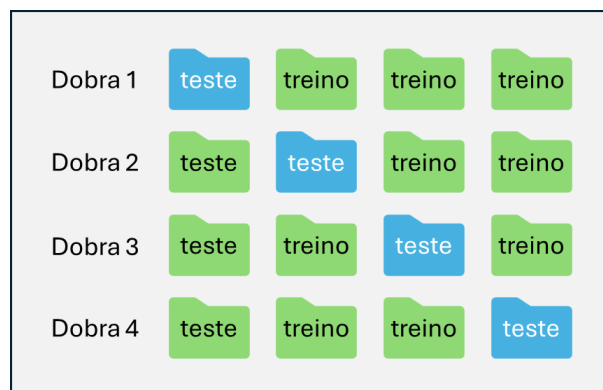


Figura 5: Validação Cruzada do tipo LOOCV, onde o número de dobras é igual a quantidade de amostras.
Fonte: O autor, 2025.

A partir das previsões do modelo, seu desempenho pode ser avaliado por meio de métricas conhecidas como figuras de mérito. Para isso, utiliza-se uma tabela chamada matriz de confusão, que permite visualizar o desempenho do modelo por classe. Para modelos de classificação, uma das métricas mais comuns é a acurácia, que representa a proporção de previsões corretas em relação ao total de previsões. Em problemas de classificação binária, também é possível calcular as taxas de verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN), além de métricas como *F1-score* e área sob a curva ROC (*Receiver Operating Characteristic*), que avaliam o desempenho considerando o equilíbrio entre as classes (61,62).

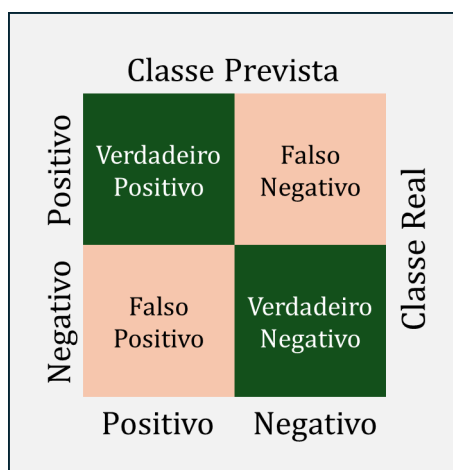


Figura 6: Matriz de Confusão para as Classes "Positivo" e "Negativo".
Fonte: O autor, 2025.

Os métodos supervisionados de classificação podem ser lineares ou não lineares. Um método é linear quando a relação entre as variáveis independentes e a variável dependente pode ser descrita por uma combinação linear das variáveis. Já os métodos não lineares não fazem essa suposição. Entre os métodos lineares, destaca-se o LDA (*Linear Discriminant Analysis*). Já a Floresta Aleatória (*Random Forest*), proposta por Breiman (63), é um método não linear.

Os modelos de classificação podem ainda ser classificados como paramétricos e não paramétricos. Os métodos paramétricos assumem que os dados seguem uma distribuição específica, como a normal. Um exemplo é o LDA. Já os métodos não paramétricos geram classificadores diretamente a partir dos dados, como no caso da Floresta Aleatória (*Random Forest*) (45).

Neste trabalho, optou-se pela utilização do modelo *Random Forest*, um modelo não linear e não paramétrico, amplamente empregada em diversas áreas do conhecimento, demonstrando eficácia em uma variedade de aplicações (64,58).

3.5 Florestas Aleatórias (*Random Forests*)

O classificador *Random Forest* é uma técnica de aprendizado de máquina baseada em múltiplas árvores de decisão (*Decision Trees*). No método *Decision Tree*, o objetivo é dividir iterativamente o conjunto de dados em subconjuntos mais homogêneos, minimizando a impureza dos grupos resultantes. A impureza em árvores de decisão é uma medida de quão misturados estão os dados dentro de um nó. Quanto maior a impureza, mais diversidade de classes há no conjunto de amostras desse nó. O objetivo da árvore de decisão é dividir os dados de forma a reduzir essa impureza, criando subconjuntos mais homogêneos. Para medir essa impureza, os critérios mais comuns são o Índice de Gini e a Entropia (65).

O *Random Forest* aprimora o processo de classificação ao criar múltiplas árvores de decisão treinadas com subconjuntos aleatórios de dados e variáveis preditoras. O nome deriva desse princípio, pois consiste em um "conjunto de árvores" que trabalham em conjunto para tomar decisões mais robustas. A classificação final é feita por meio de votação majoritária entre as árvores, o que torna o modelo mais resistente a *overfitting* e melhora sua capacidade de generalização. Para avaliar seu desempenho, as métricas mais utilizadas incluem Acurácia, que mede a proporção de classificações corretas, e o F1-Score, especialmente útil para problemas com classes desbalanceadas. Além disso, a Área sob a Curva ROC (AUC-ROC) é frequentemente empregada para medir a capacidade do modelo de distinguir entre diferentes classes (65).

Neste trabalho, o *Random Forest* foi utilizado empregando a impureza de Gini como critério de divisão das árvores de decisão. A impureza de Gini mede a probabilidade de uma amostra ser classificada incorretamente se atribuída aleatoriamente a uma classe dentro de um subconjunto específico. Matematicamente, é definida como o somatório do produto da frequência relativa de uma classe pela frequência relativa das demais classes dentro do mesmo subconjunto. A frequência relativa é calculada pela proporção do número de amostras de uma classe em relação ao total de amostras presentes no subconjunto. No contexto do *Random Forest*, essa métrica é aplicada em cada árvore individual para selecionar as divisões mais informativas, contribuindo para a criação de um modelo robusto e capaz de lidar com variações nos dados. A agregação dos resultados de múltiplas árvores, por meio de votação majoritária, melhora a

precisão do modelo e reduz o risco de *overfitting*, tornando o classificador mais generalizável e eficaz (65).

A estrutura de uma árvore de decisão é composta por três elementos principais: nós, galhos e folhas. Os nós representam os pontos de decisão onde uma variável é selecionada para dividir os dados com base em um critério de impureza, como o Índice de Gini. Os galhos conectam os nós e indicam os possíveis caminhos resultantes de cada decisão, conduzindo as amostras aos subconjuntos subsequentes. Já as folhas contêm os subconjuntos finais e representam as classificações atribuídas pelo modelo. O processo de divisão ocorre de maneira recursiva, segmentando os dados em grupos cada vez mais homogêneos, até que um critério de parada seja atingido. Esse critério pode ser a pureza total dos subconjuntos, um número mínimo de amostras por nó, a profundidade máxima da árvore ou um limite para a redução da impureza (65).

No modelo de *Random Forest* utilizado neste trabalho, o critério de parada para o crescimento das árvores de decisão é determinado por alguns parâmetros principais. O parâmetro *max_depth* é configurado como *None*, o que significa que as árvores de decisão crescerão até que os nós atinjam um ponto em que as folhas contenham menos amostras do que o valor definido em *min_samples_split*. Ou seja, as árvores continuarão a se expandir até que não seja possível realizar divisões adicionais de maneira significativa.

O parâmetro *min_samples_split* está configurado como 2, o que define o número mínimo de amostras necessárias para dividir um nó. Quando o número de amostras em um nó é menor que esse valor, a divisão é interrompida e o nó se torna uma folha. Além disso, o parâmetro *min_samples_leaf*, configurado como 1, estabelece o número mínimo de amostras para que um nó seja considerado uma folha. Dessa forma, as árvores são capazes de crescer até que os critérios de divisão e o número mínimo de amostras sejam atingidos, garantindo que a divisão dos dados seja eficaz e evitando o crescimento excessivo das árvores. Como o parâmetro *max_depth* não está limitado, a profundidade das árvores não é restringida diretamente, permitindo que elas se expandam até atingirem os critérios de parada estabelecidos. Essas configurações ajudam a controlar o *overfitting*, permitindo que as árvores cresçam de maneira equilibrada e adequada às características do conjunto de dados.

O modelo baseado em *Random Forest* é relativamente simples e interpretável, mas pode apresentar instabilidade, já que pequenas mudanças na formação dos subgrupos podem gerar árvores substancialmente diferentes, o que resulta em variações significativas nos resultados a cada ciclo de treinamento e teste. Além disso, esses modelos podem sofrer com *overfitting*, especialmente quando o número de variáveis (atributos ou características dos dados) é maior do que o número de amostras (exemplos ou observações), o que pode levar o modelo a aprender padrões específicos aos dados de treinamento e não generalizáveis. Para mitigar esses problemas, técnicas de “poda” podem ser aplicadas, como limitar a profundidade das árvores, estabelecer um número mínimo de amostras em cada subdivisão e restringir a quantidade de variáveis consideradas em cada nó (66,58).

Para superar as limitações das árvores de decisão individuais, podem ser utilizadas técnicas de combinação de múltiplas árvores, como *bagging* e *boosting*. O *bagging* (*Bootstrap Aggregation*) cria múltiplos modelos independentes de árvores de decisão, treinados com subconjuntos de dados gerados por reamostragem aleatória com reposição. As previsões dessas árvores são então combinadas, o que reduz a variabilidade e o *overfitting* do modelo. Já o *boosting* gera modelos sequenciais de árvores de decisão, ajustando os pesos das amostras para corrigir os erros das iterações anteriores, o que ajuda a reduzir o viés. Enquanto o *bagging* foca em reduzir a variância, o *boosting* busca diminuir o viés, tornando ambos métodos poderosos para melhorar a robustez e a precisão das árvores de decisão (66).

A Floresta Aleatória é um método de aprendizado de máquina baseado na técnica de *bagging*, no qual a classificação final é obtida por meio da votação das classificações feitas por cada árvore de decisão individual no *ensemble*. Para criar cada árvore, o processo começa com a reamostragem aleatória do conjunto de dados, ou seja, amostras são selecionadas com reposição, o que significa que algumas amostras podem ser escolhidas mais de uma vez enquanto outras podem ser descartadas. Esse procedimento ajuda a aumentar a diversidade entre as árvores e reduz a variabilidade do modelo final. Além disso, a Floresta Aleatória introduz uma camada adicional de aleatoriedade: em cada divisão (nó) de uma árvore, é considerado apenas um subconjunto aleatório de variáveis para decidir a melhor divisão, em vez de considerar todas as variáveis possíveis. Isso promove uma maior independência entre as árvores, evitando que elas se tornem altamente correlacionadas. Esse processo de introdução de variabilidade tanto nos dados quanto nas variáveis para cada árvore ajuda a reduzir a variância do modelo final,

melhorando a robustez e a capacidade de generalização do modelo, especialmente em relação a conjuntos de dados ruidosos ou complexos. (64).

A interpretação da importância das variáveis na Floresta Aleatória é feita através da métrica de decréscimo médio da impureza de Gini (*Mean Decrease in Gini Impurity*), que avalia a contribuição de cada variável para a redução da impureza de Gini ao longo de todas as árvores, sendo ponderada pelo número de vezes que a variável foi utilizada para dividir um nó. Essa métrica fornece uma estimativa quantitativa da relevância de cada variável no modelo. Além disso, é possível identificar as amostras mais decisivas observando aquelas que são mais frequentemente utilizadas em divisões importantes ou em ramos críticos das árvores, ajudando a entender quais dados influenciam mais na classificação final e contribuindo para a explicabilidade do modelo. (67).

4 METODOLOGIA

4.1 Coleta e Preparação das Amostras

As amostras analisadas consistem em penas de aves jovens, machos e fêmeas, das espécies *Oryzoborus maximiliani* (Bicudo), *Nymphicus hollandicus* (Calopsita), *Oryzoborus angolensis* (Curió) e *Psittacula krameri* (Ring Neck). Foram recebidas 30 amostras de cada sexo para cada espécie, totalizando 240 amostras, sendo cada uma proveniente de um indivíduo diferente. As amostras foram gentilmente fornecidas pela empresa de sexagem de aves “Codex Gen”, localizada em Campo Grande, MS. O sexo de todas as amostras foi previamente identificado por meio da técnica de referência de análise de DNA. As amostras vieram embaladas individualmente em sacos plásticos, todas devidamente identificadas. Após o recebimento, foram colocadas em um recipiente plástico e condicionadas em refrigerador com temperatura controlada em torno de 6 °C.

4.2 Procedimento para realização das medidas

O conjunto de amostras selecionados para análise foi organizado de forma que as medidas fossem realizadas no mesmo dia (machos e fêmeas), diminuindo vies proveniente das condições experimentais. As penas foram retiradas da geladeira e deixada sobre a bancada em ambiente com ar-condicionado em temperatura em torno de 23° C por 15 minutos antes da realização das medidas. Antes da análise, as amostras foram higienizadas com lenço de papel para retirada de possíveis contaminantes, como poeira. As medidas foram realizadas na região da pena conhecida como vexilo, indicada na Figura 1:

4.3 Medidas por FTIR

As medições de espectroscopia no infravermelho por transformada de Fourier (FTIR, do inglês *Fourier Transform Infrared Spectroscopy*) foram realizadas em um espectrômetro da marca “Perkin Elmer”, modelo “Spectrum 100 Series”, na região do infravermelho médio (MIR, do inglês *Mid-Infrared*). Os espectros foram obtidos na faixa de 3800 a 800 cm^{-1} , com incremento de número de onda de 0,5 cm^{-1} , no modo de transmitância. Cada leitura consistiu em 10 varreduras (scans) com resolução espectral de 4 cm^{-1} .

As amostras foram analisadas diretamente no acessório de reflexão total atenuada (ATR, do inglês *Attenuated Total Reflection*), utilizando um cristal de seleneto de zinco (ZnSe). A

superfície do cristal foi limpa com acetona antes de cada medição para evitar contaminações. Cada observação corresponde à média de duas leituras realizadas em duplicata para cada amostra, resultando em um total de 240 observações representativas por classe (Fêmea e Macho).

4.4 Pré processamento e Análise de Componentes Principais

Primeiramente, foi realizada uma análise visual dos espectros brutos de cada espécie, com o objetivo de identificar possíveis inconsistências nas leituras do equipamento. Em seguida, foram calculadas as médias das duplicatas, e os dados resultantes passaram por pré-processamento e normalização utilizando a técnica SNV (*Standard Normal Variate*). A técnica SNV diminui o deslocamento (*offset*) dos espectros e padroniza suas intensidades, ao centralizar os dados pela média de cada espectro e normalizá-los pelo respectivo desvio padrão, de acordo com a equação (1) descrita anteriormente.

Em seguida, foi calculada a média de cada espécie, assim como a média dos espectros de cada classe (macho e fêmea) por espécie. As médias foram comparadas para identificar semelhanças visuais entre os espectros das espécies, assim como a existência de bandas espectrais que possibilitem identificar visualmente o sexo, o que impossibilitaria o uso de *machine learning*.

Os espectros de cada espécie foram submetidos à Análise de Componentes Principais (PCA) com o objetivo de avaliar a formação de grupos por classe e identificar as bandas que mais contribuíram para a variância observada. Inicialmente, foi analisado o espectro completo, abrangendo a faixa de 3800 a 800 cm^{-1} . Em seguida, foram investigados intervalos espectrais específicos com base na magnitude dos *loadings* observados: de 3000 a 2800 cm^{-1} e de 1800 a 800 cm^{-1} . O PCA é uma técnica estatística que transforma um conjunto de variáveis possivelmente correlacionadas em um novo conjunto de variáveis não correlacionadas, denominadas componentes principais (PC's). Essas PC's são combinações lineares das variáveis originais, ordenadas de forma que a primeira componente explica a maior parte da variância dos dados, a segunda explica a segunda maior variância, e assim por diante (68). Esse processo reduz a dimensionalidade dos dados, preservando as informações mais relevantes para a análise (69). Os *scores* obtidos nos intervalos espectrais indicados foram então utilizados na construção e avaliação dos modelos de *machine learning*.

4.5 Aprendizado de Máquina (*Machine Learning*)

O conjunto de dados foi dividido em 70% para treinamento e 30% para teste, e o classificador utilizado foi o *Random Forest*, um algoritmo de aprendizado de máquina baseado em um conjunto de árvores de decisão. O *Random Forest* utiliza a técnica de *bagging* (*bootstrap aggregating*), na qual cada árvore é treinada com subconjuntos aleatórios de amostras, selecionadas com reposição (*bootstrap*), e um subconjunto aleatório de variáveis para cada divisão interna da árvore (70,71). Esse processo promove diversidade entre as árvores, aumentando a robustez e a capacidade de generalização do modelo.

No presente trabalho, o número de árvores foi ajustado em um intervalo de 100 a 200, variando em incrementos de 10, para otimizar o desempenho do classificador. Além disso, para evitar o *overfitting*, foram realizados testes controlando o número de Componentes Principais (PC's), utilizando um teste incremental com as 20 primeiras PC's. Essa abordagem resultou na construção e avaliação de 110 modelos, permitindo analisar o impacto da dimensionalidade sobre a performance do classificador.

A acurácia de treinamento para cada modelo foi obtida por validação do tipo *Leave-One-Out Cross-Validation* (LOOCV), onde uma amostra foi deixada de fora e as 20 restantes foram utilizadas para realizar o PCA. Os *scores* obtidos foram tratados por SNV e usados no treinamento. Em seguida a amostra deixada de fora foi projetada no espaço do PCA de treinamento e os *scores* normalizados por SNV foram usados para testar seu desempenho (validação). Este procedimento foi repetido até que todas as amostras tenham sido deixadas de fora (21 vezes). A acurácia de validação foi obtida pela média das acurácias de cada validação.

Uma das principais vantagens da LOOCV é a utilização eficiente do conjunto de dados, pois em cada iteração um único ponto de dados é usado para teste, e todos os outros são utilizados para treinamento. Isso é especialmente vantajoso em conjuntos de dados pequenos, onde cada amostra é importante. Além disso, o LOOCV proporciona uma estimativa com baixa variância, uma vez que o modelo é validado em todas as amostras, permitindo uma avaliação mais precisa de sua capacidade de generalização. O método também possibilita uma avaliação detalhada do desempenho do modelo em cada ponto de dados, sendo útil para identificar falhas específicas ou dificuldades de generalização.

Os parâmetros ótimos foram definidos selecionando o modelo que apresentou a maior acurácia no LOOCV utilizando a menor quantidade de PC's. Finalmente, as 42 amostras de treino foram usadas para treinar um modelo com os melhores parâmetros obtidos pelo LOOCV e as 18 amostras de teste que não participaram das etapas anteriores foram usadas para testar o modelo. A partir das predições de treinamento (LOOCV) e de teste do modelo ótimo, foram construídas as matrizes de confusão, permitindo avaliar o desempenho da classificação por sexo. Além disso, foi possível comparar o desempenho do modelo entre o treinamento e o teste, avaliando sua capacidade de generalização.

5 RESULTADOS E DISCUSSÃO

Na Figura 7 a média dos espectros de cada espécie são comparados e os picos das bandas mais evidentes são destacados.

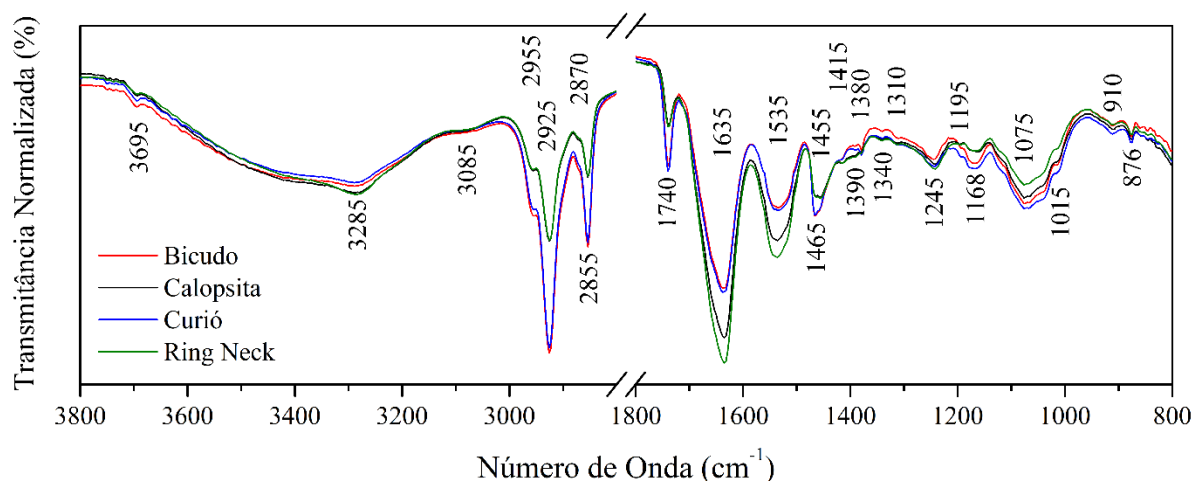


Figura 7: Média dos espectros FTIR de cada espécie.

As bandas observadas estão de acordo com aquelas encontradas na literatura para o vexilo de penas, que são detalhadas na Tabela 1 (72).

Tabela 1: Atribuição de picos e seus modos vibracionais em espectros FTIR de vexilos de penas.

Número de Onda (cm ⁻¹)	Pico Atribuído	Modos Vibracionais
1075	C-C	Alongamento simétrico
1168	C-C	Alongamento simétrico
1245	(CN) Amida III	Alongamento simétrico
1310	C-H ₂	Vibração de flexão simétrica
1380	C-H ₃	Vibração de flexão simétrica
1436	C-H ₃	Vibração de flexão simétrica
1455	C-H ₃	Vibração de flexão simétrica
1535	(N-H) Amida II, folha β	Vibração de flexão simétrica
1655	(C=O) Amida I, hélice α	Alongamento simétrico
1666	(C=O) Amida I, folha β	Alongamento simétrico
1680	C=O	Alongamento simétrico
2870	CH ₂	Alongamento simétrico
2925	CH ₃	Alongamento simétrico
2955	CH ₃	Vibrações de alongamento assimétrico
3085	NH Amida B	Vibrações de alongamento assimétrico
3285	NH Amida A, hélice α	Alongamento simétrico

Com base nos resultados obtidos, é possível observar uma semelhança entre os espectros de Bicudo e Curió, assim como entre Calopsita e *Ring Neck*. Essas diferenças podem ser observadas nas bandas distintas em torno de 1415, 1390, e 1195 cm⁻¹, bem como nas intensidades das demais bandas.

5.1 *Oryzoborus maximiliani* (Bicudo)

Na Figura 8 a média e o desvio padrão dos espectros do pássaro Bicudo são comparados por classe (Fêmea e Macho). Nota-se bastante semelhança entre os espectros, sendo a aplicação de rotinas de *machine learning* fundamentais para identificar e esclarecer suas diferenças.

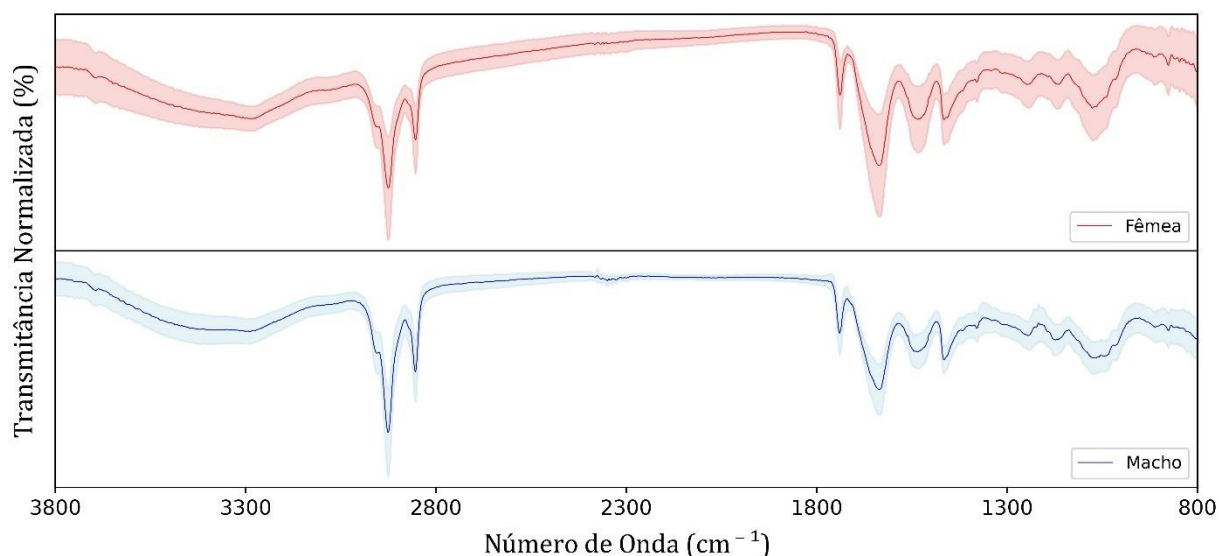


Figura 8: Média e desvio padrão dos espectros FTIR separados por sexo para o pássaro Bicudo.

Comparando as classes “Fêmea” e “Macho” para o pássaro “Bicudo” foram observadas diferenças nos espectros médios nas regiões em torno de 1700, 1000 e 900 cm^{-1} além das bandas previamente identificadas em 1390, 1415, e 1195 cm^{-1} . Contudo, não é possível atribuir essas variações exclusivamente ao sexo das aves, sendo necessária a aplicação de técnicas mais robustas de classificação, como as fornecidas por algoritmos de *machine learning*. Foi observado um desvio padrão maior nas amostras da classe Fêmea, que pode influenciar a variância dos dados e consequentemente a precisão do modelo.

A Figura 9 apresenta os *scores* e *loadings* da Análise de Componentes Principais (PCA) para o intervalo de 3800 a 800 cm^{-1} para o pássaro Bicudo. Essa análise permite visualizar como as amostras se relacionam entre si, identificar a formação de grupos de acordo com suas classes e determinar quais bandas contribuem de maneira mais significativa para a variância observada. Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

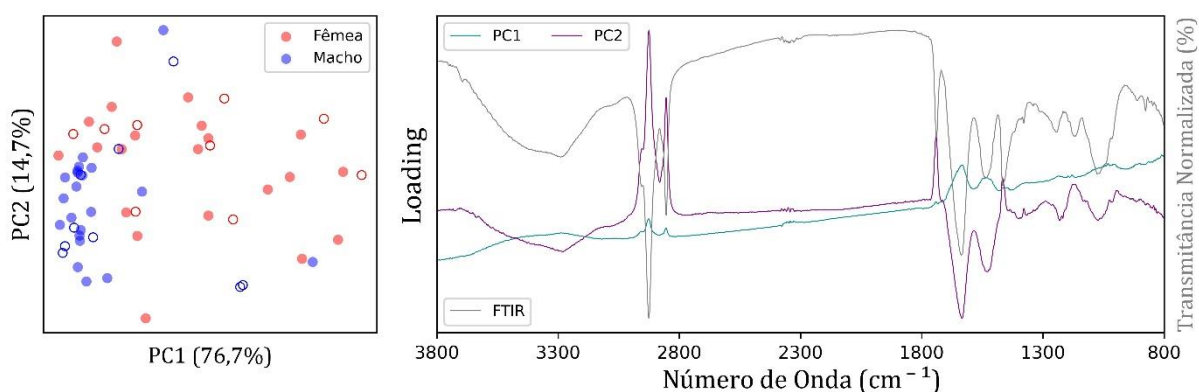


Figura 9: *Scores* e *Loadings* para o pássaro Bicudo (3800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

Nos *scores*, observa-se uma grande variância nos dados, principalmente da classe “Fêmea” em relação a classe “Macho”. Ambas as classes formam subgrupos, indicando a influência de algum fator desconhecido nos resultados. Os *loadings* indicam uma contribuição de grande magnitude nas bandas localizadas entre 3000 e 2800 cm^{-1} , bem como na região compreendida entre 1800 e 1300 cm^{-1} . Foram realizados treinamentos nos intervalos de entre 3000 e 2800 cm^{-1} e 1800 até 800 cm^{-1} , com o objetivo de identificar o intervalo mais promissor.

As Figuras a seguir apresentam as acurácias de treinamento obtidas pelo classificador *Random Forest* por validação do tipo *Leave-One-Out Cross-Validation* (LOOCV), para diferentes quantidades de árvores e Componentes Principais (PC's), para os intervalos: 3800 a 800 cm^{-1} (Figura 10), 3000 a 2800 cm^{-1} (Figura 11) e 1800 a 800 cm^{-1} (Figura 12) para o pássaro Bicudo.

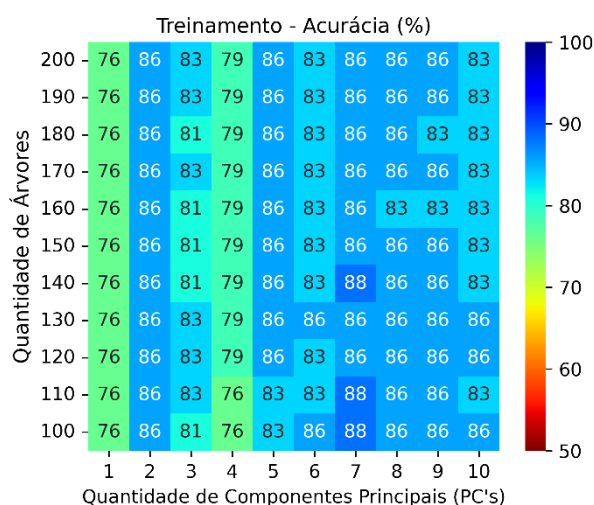


Figura 10: Acurácias em função da quantidade de PC's e árvores para o pássaro Bicudo (3800 a 800 cm^{-1}).

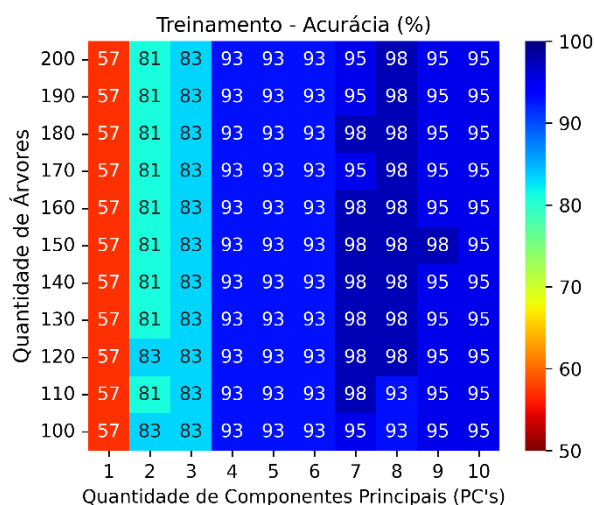


Figura 11: Acurácias em função da quantidade de PC's e árvores para o pássaro Bicudo (3000 a 2800 cm^{-1}).

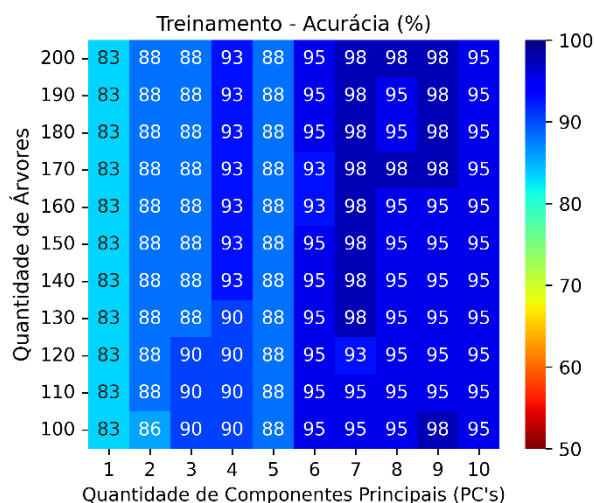


Figura 12: Acurácias em função da quantidade de PC's e árvores para o pássaro Bicudo (1800 a 800 cm^{-1}).

Ao comparar os resultados, observa-se que, de modo geral, o intervalo de 3000 a 2800 cm^{-1} apresentou melhor desempenho em relação aos demais, com uma acurácia máxima de 97,6%, alcançada com 110 árvores e 7 PC's. No entanto, o intervalo de 1800 a 800 cm^{-1} também obteve a mesma acurácia máxima, destacando-se por atingir 83% com apenas uma componente principal (PC1), enquanto o intervalo de 3000 a 2800 cm^{-1} alcançou apenas 57% sob a mesma condição. A Figura 13 apresenta os *scores* e *loadings* do PCA para o intervalo de 3000 até 2800 cm^{-1} , região com melhor desempenho para análise de sexagem do pássaro “Bicudo”. Os círculos

são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

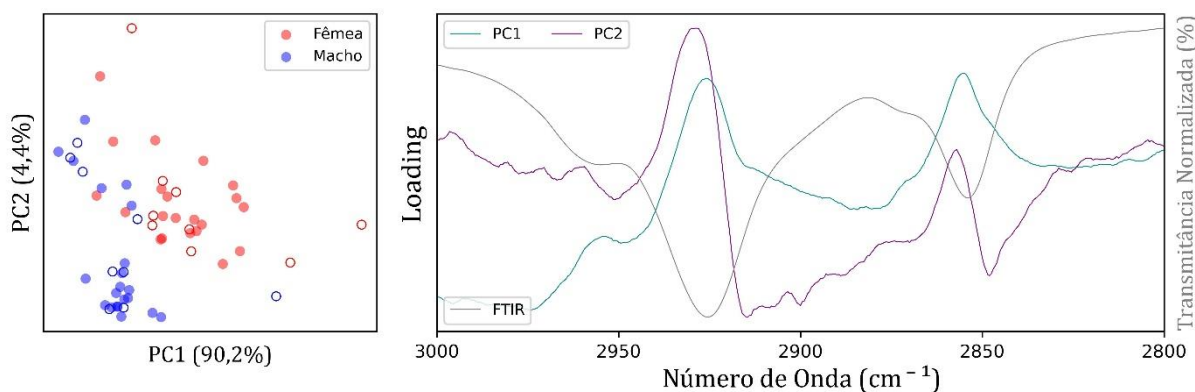


Figura 13: Scores e Loadings para o pássaro Bicudo (3000 a 2800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

Observando os *scores*, verifica-se uma menor variância em ambas as classes, com destaque para a classe “Macho” que possui um agrupamento mais denso que a classe “Fêmea”, quando comparada com o intervalo espectral de 3800 a 800 cm^{-1} . As amostras da classe “Macho” formam dois subgrupos, assim como no intervalo de 3800 a 800 cm^{-1} . Porém, a formação dos grupos corresponde às classes estudadas, apesar de alguma sobreposição. Os *loadings* indicam contribuição das três principais bandas deste intervalo com picos em torno de 2955, 2925 e 2855 cm^{-1} . As penas de aves são compostas basicamente de queratina, a principal proteína estrutural desse material (73). As bandas vibracionais características dos grupos funcionais da queratina, associadas às vibrações de grupos alifáticos (CH_2 e CH_3), foram detectadas nessa região espectral (73,74). Os resultados indicam que a composição ou estrutura das penas apresenta diferenças entre machos e fêmeas, permitindo sua classificação.

Na Figura 14 as matrizes de confusão de treinamento por LOOCV (a) e teste (b) são comparadas para a melhor configuração do classificador. Cada linha corresponde ao conjunto real e as colunas são as previsões feitas pelo classificador. A acurácia é a média das acurácias apresentadas na diagonal.

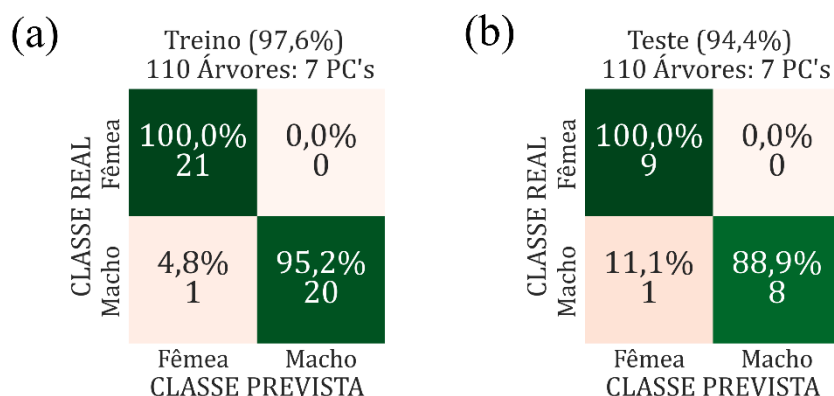


Figura 14: Matrizes de confusão de validação LOOCV (a) e teste (b) para o pássaro Bicuado (3000 a 2800 cm^{-1}).

Os resultados indicam que a acurácia de teste foi de 94,4%, um valor muito próximo da acurácia de treino (97,6%), pelo uso de 110 árvores e 7 PC's. O modelo foi capaz de prever com 100% de acurácia a classe “Fêmea”, enquanto a classe “Macho” enfrentou dificuldades devido à formação de subgrupos, o que comprometeu a construção de um modelo mais robusto. Ainda assim, o resultado é promissor.

5.2 *Nymphicus hollandicus* (Calopsita)

Na Figura 15 a média e o desvio padrão dos espectros do pássaro “Calopsita” são comparados por classe (“Fêmea” e “Macho”). Nota-se bastante semelhança entre os espectros, sendo a aplicação de rotinas de *machine learning* fundamentais para identificar e esclarecer suas diferenças.

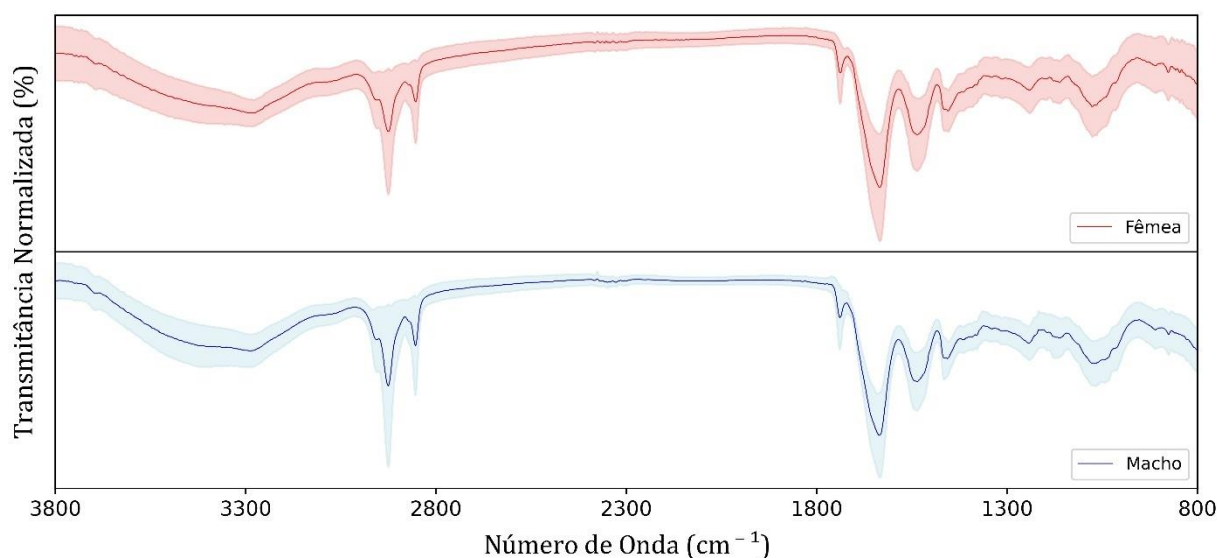


Figura 15: Média e desvio padrão dos espectros FTIR separados por sexo para o pássaro Calopsita.

Comparando as classes “Fêmea” e “Macho” para o pássaro “Calopsita”, verifica-se que as médias têm aspecto semelhante, porém, foram observadas diferenças nos espectros médios nas regiões em torno de 1700, 1200, 1050, 1000 e 850 cm^{-1} . Contudo, não é possível atribuir essas diferenças exclusivamente ao sexo das aves. Assim, novamente, torna-se necessária a aplicação de técnicas de *machine learning*. Foi observado novamente um desvio padrão maior nas amostras da classe “Fêmea” que pode influenciar a variância dos dados e consequentemente a precisão do modelo de *machine learning*.

A Figura 16 apresenta os *scores* e *loadings* do PCA para o intervalo de 3800 a 800 cm^{-1} para o pássaro “Calopsita”. Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

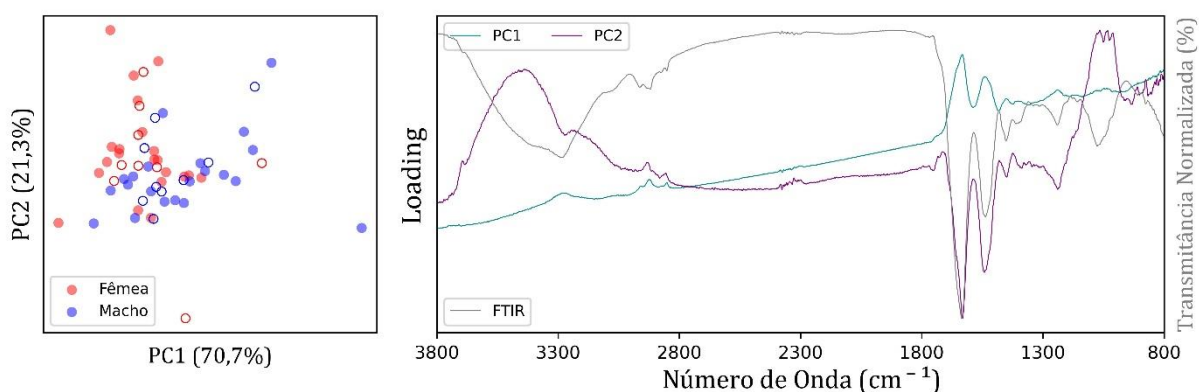


Figura 16: *Scores* e *Loadings* para o pássaro Calopsita (3800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

Nos *scores*, observa-se uma tendência de formação de grupos, apesar da sobreposição. A variância nos dados é menor que a obtida para o pássaro “Bicudo” e não há formação de subgrupos. Os *loadings* apresentam características diferentes do caso estudado para o pássaro “Bicudo”, indicando contribuições mais evidentes nas bandas em torno de 3500 cm^{-1} e 1000 cm^{-1} , com redução do peso na banda em torno de 3000 até 2800 cm^{-1} . Foram realizados treinamentos nos intervalos entre 3000 e 2800 cm^{-1} e 1800 até 800 cm^{-1} , com o objetivo de identificar o intervalo espectral mais adequado.

As Figuras a seguir apresentam as acurácias obtidas pelo classificador *Random Forest* por treinamento LOOCV, para diferentes quantidades de árvores e PC's, para os intervalos: 3800 a 800 cm^{-1} (Figura 17), 3000 a 2800 cm^{-1} (Figura 18) e 1800 a 800 cm^{-1} (Figura 19) para o pássaro Calopsita.

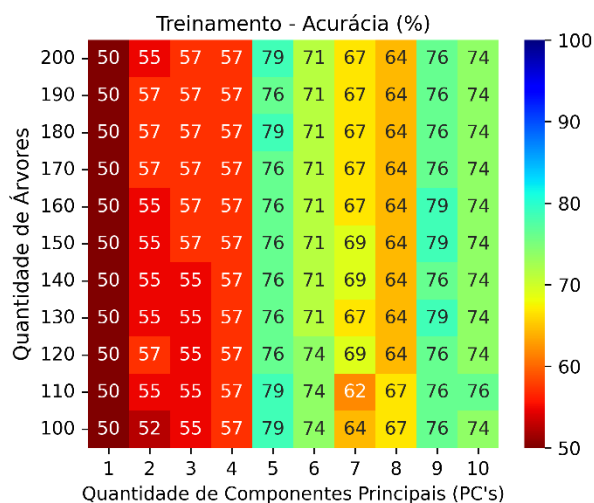


Figura 17: Acurácias em função da quantidade de PC's e árvores para o pássaro Calopsita (3800 a 800 cm^{-1}).

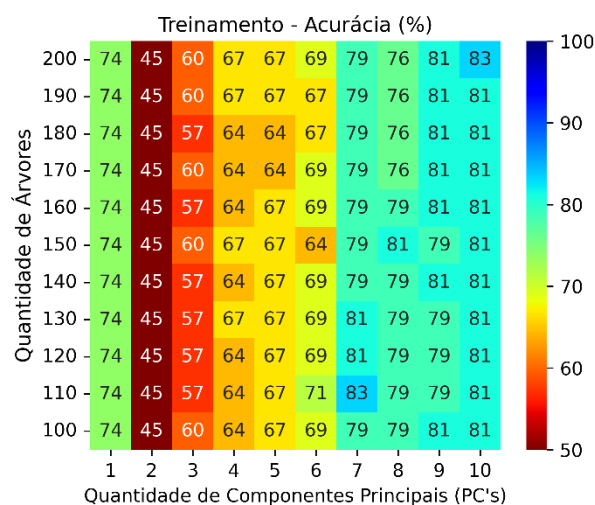


Figura 18: Acurácias em função da quantidade de PC's e árvores para o pássaro Calopsita (3000 a 2800 cm^{-1}).

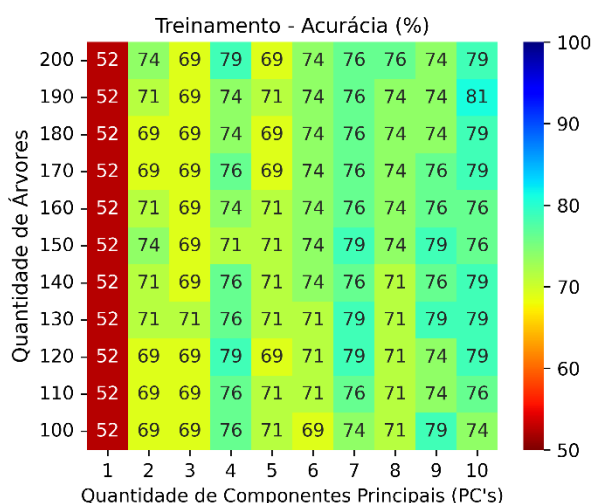


Figura 19: Acurácias em função da quantidade de PC's e árvores para o pássaro Calopsita (1800 a 800 cm^{-1}).

Ao comparar os resultados, observa-se que, de modo geral, o intervalo de 3000 a 2800 cm^{-1} apresentou o melhor desempenho em relação aos demais. Nesse intervalo, a maior acurácia de treinamento obtida foi de 83,3% com 110 árvores e 7 PC's.

A Figura 20 apresenta os *scores* e *loadings* do PCA para o intervalo de 3000 a 2800 cm^{-1} , região com o melhor desempenho para análise de sexagem do pássaro “Calopsita”. Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

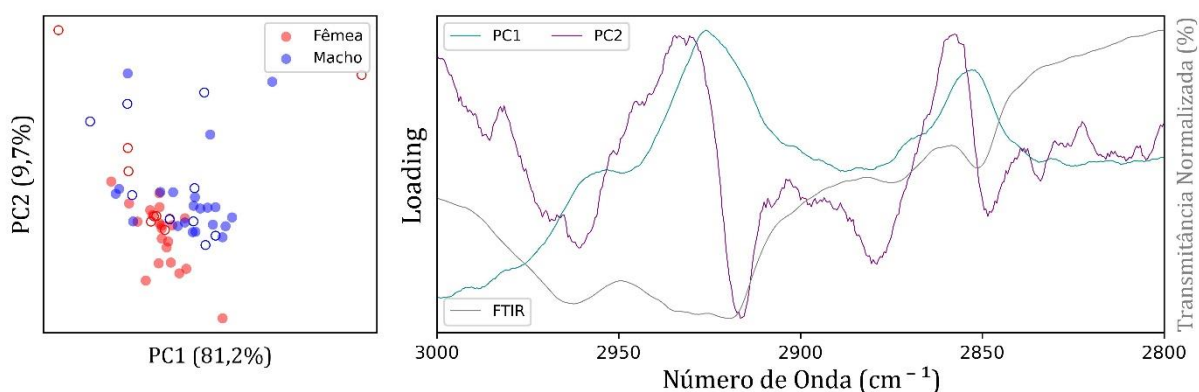


Figura 20: *Scores* e *Loadings* para o pássaro Calopsita (3000 a 2800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

Observando os *scores*, verifica-se um agrupamento claro e condensado, porém com a presença de um subgrupo na classe “Macho”, o que pode ter afetado o desempenho do classificador - assim como nos resultados para o pássaro “Bicudo”. Os *loadings* indicam contribuição das três principais bandas deste intervalo com picos em torno de 2955, 2925 e 2855 cm^{-1} , enquanto na PC2 os *loadings* indicam a contribuições de pesos em torno dessas bandas. Os resultados indicam que a composição ou estrutura das penas apresenta diferenças entre machos e fêmeas, principalmente em relação à queratina (73,74).

Na Figura 21 as matrizes de confusão de validação por LOOCV (a) e teste (b) são comparadas. Cada linha corresponde ao conjunto real e as colunas são as previsões feitas pelo classificador. A acurácia é a média das acurácias apresentadas na diagonal.

(a)		Treino (83,3%) 110 Árvores: 7 PC's	
CLASSE REAL	Fêmea	81,0% 17	19,0% 4
	Macho	14,3% 3	85,7% 18
		Fêmea	Macho
		CLASSE PREVISTA	

(b)		Teste (77,8%) 110 Árvores: 7 PC's	
CLASSE REAL	Fêmea	66,7% 6	33,3% 3
	Macho	11,1% 1	88,9% 8
		Fêmea	Macho
		CLASSE PREVISTA	

Figura 21: Matrizes de confusão de validação LOOCV (a) e teste (b) para o pássaro Calopsita (3000 a 2800 cm^{-1}).

Os resultados indicam que a acurácia de teste foi de 77,8%, um valor muito próximo da acurácia de validação (83,3%). O modelo foi capaz de prever melhor a classe “Macho” do que a classe “Fêmea”. Logo, novamente enfrentou dificuldades na classificação devido à formação de subgrupos. No entanto, o resultado é promissor.

5.3 *Oryzoborus angolensis* (Curió)

Na Figura 22 a média e o desvio padrão dos espectros do pássaro Curió são comparados por classe (“Fêmea” e “Macho”). Nota-se bastante semelhança entre os espectros, sendo a aplicação de rotinas de *machine learning* fundamentais para identificar e esclarecer suas diferenças.

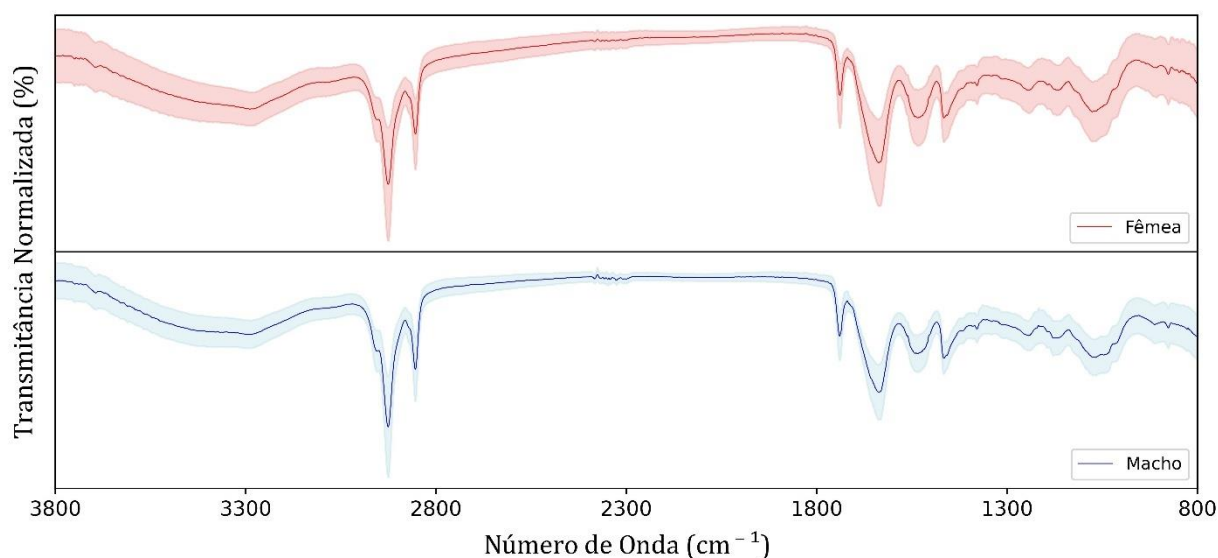


Figura 22: Média e desvio padrão dos espectros FTIR separados por sexo para o pássaro Curió.

Comparando as classes “Fêmea” e “Macho” para o pássaro “Curió”, verifica-se que as médias têm aspecto semelhante, porém, foram observadas diferenças nos espectros médios nas regiões em torno de 1200, 1050, 1000 e 850 cm^{-1} . Contudo, não é possível atribuir essas variações exclusivamente ao sexo das aves. Assim, novamente torna-se necessária a aplicação de técnicas de *machine learning*. Foi observado um desvio padrão maior nas amostras da classe Fêmea, assim como nas amostras anteriores que pode influenciar a variância dos dados e consequentemente a precisão do modelo.

A Figura 23 apresenta os *scores* e *loadings* do PCA para o intervalo de 3800 a 800 cm^{-1} para o pássaro “Curió”. Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

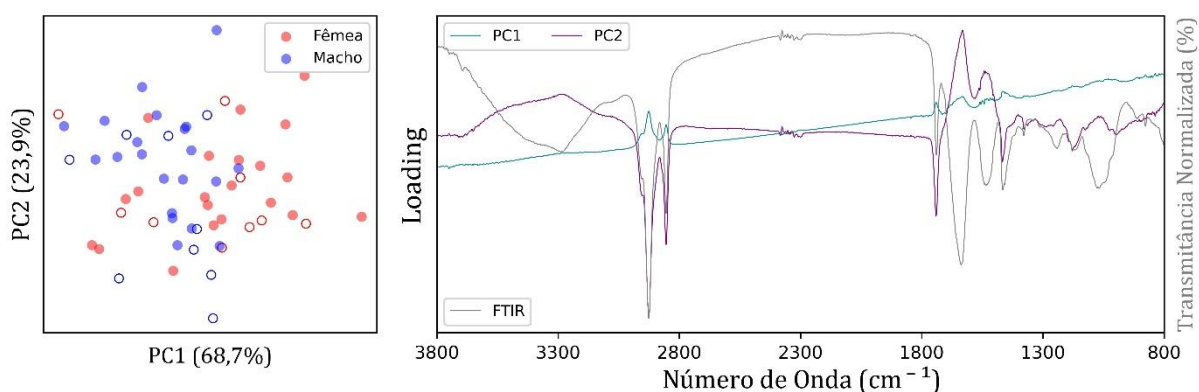


Figura 23: *Scores* e *Loadings* para o pássaro Curió (3800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

Nos *scores*, observa-se uma grande variância nos dados em ambas as classes, sem agrupamento definido. Os *loadings* apresentam características diferentes dos casos anteriores, indicando contribuições mais evidentes nas bandas em torno de 2900 cm^{-1} e 1600 cm^{-1} . Os *loadings* da PC2 apresentam uma contribuição de magnitudes significativamente menores em comparação com os *loadings* da PC1. Assim, foram realizados treinamentos nos intervalos de entre 3000 e 2800 cm^{-1} e 1800 até 800 cm^{-1} , com o objetivo de identificar o intervalo espectral mais adequado.

As Figuras a seguir apresentam as acurácias obtidas pelo classificador *Random Forest* por treinamento LOOCV, para diferentes quantidades de árvores e PC's, para os intervalos: 3800 a 800 cm^{-1} (Figura 24), 3000 a 2800 cm^{-1} (Figura 25) e 1800 a 800 cm^{-1} (Figura 26) para o pássaro “Curió”.

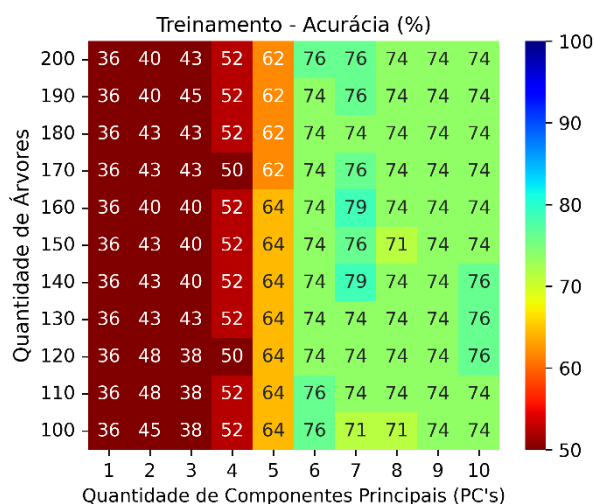


Figura 24: Acurácias em função da quantidade de PC's e árvores para o pássaro Curió (3800 a 800 cm^{-1}).

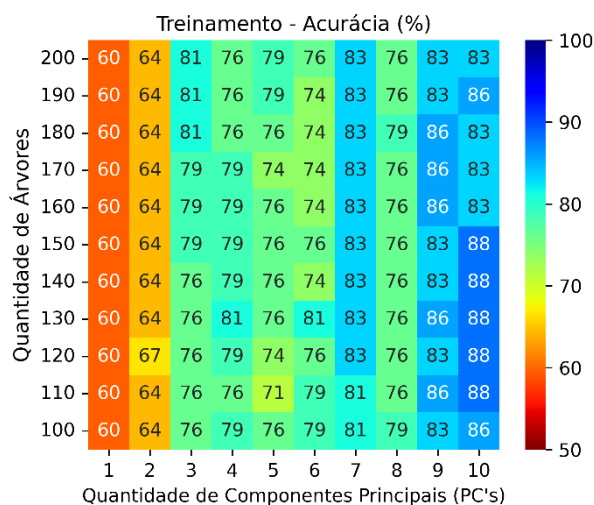


Figura 25: Acurácias em função da quantidade de PC's e árvores para o pássaro Curió (3000 a 2800 cm^{-1}).

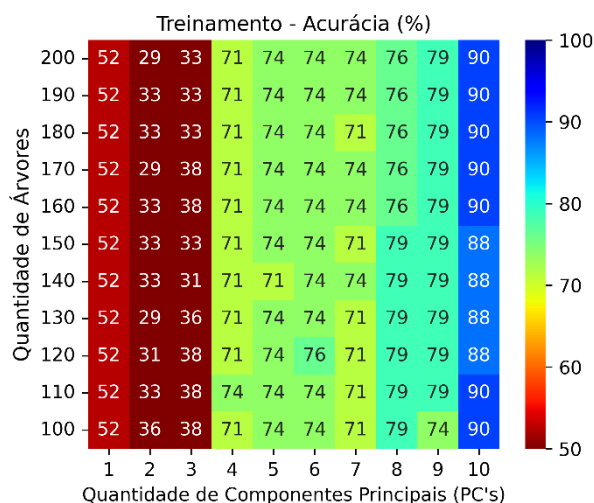


Figura 26: Acurácias em função da quantidade de PC's e árvores para o pássaro Curió (1800 a 800 cm^{-1}).

Ao comparar os resultados, observa-se que, de modo geral, diferente para os pássaros “Bicudo” e “Calopsita”, o intervalo de 1800 a 800 cm^{-1} foi o que apresentou o melhor desempenho em relação aos demais, esse intervalo foi considerado o ideal, com acurácia máxima de 90,5 % com 110 árvores e 10 PC's.

A Figura 27 apresenta os *scores* e *loadings* do PCA para o intervalo de 1800 a 800 cm^{-1} , região com melhor desempenho para análise de sexagem do pássaro “Curió”. Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

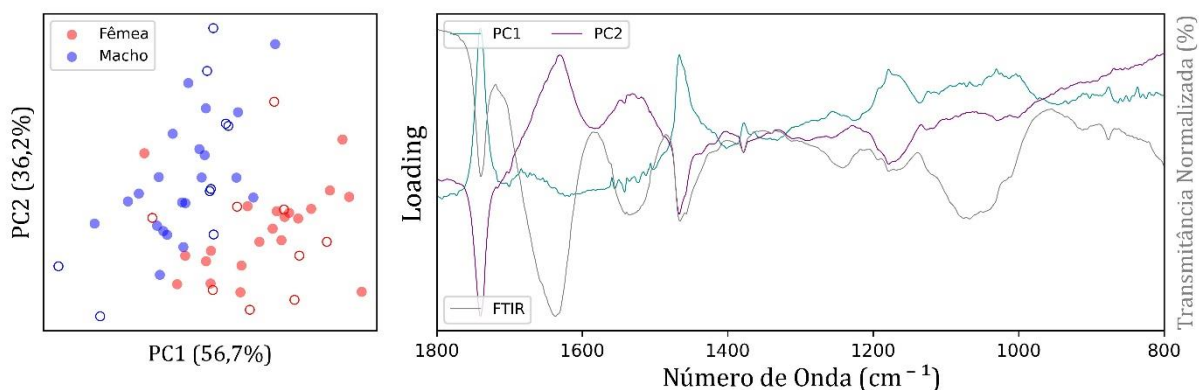


Figura 27: *Scores* e *Loadings* para o pássaro Curió (1800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

Observando os *scores*, verifica-se uma variância maior da classe “Macho” em relação a classe “Fêmea”. As amostras de treinamento da classe “Fêmea” estão no agrupamento referente a classe “Macho”. Apesar da formação de grupos, não há uma definição clara de fronteira entre eles. Os *loadings* descrevem contribuição em aproximadamente 1740 cm^{-1} e da região de amidas por volta de 1650 cm^{-1} , 1450 cm^{-1} e 1200 cm^{-1} indicando que a variabilidade está mais associada à concentração relativa de proteínas e lipídios específicos. Os resultados indicam que a composição ou estrutura das penas apresenta diferenças entre machos e fêmeas, principalmente em relação à queratina (73,74).

Na Figura 28 as matrizes de confusão de validação por LOOCV (a) e teste (b) são comparadas. Cada linha corresponde ao conjunto real e as colunas são as previsões feitas pelo classificador. A acurácia é a média das acurácias apresentadas na diagonal.

(a)		Treino (90,5%) 100 Árvores: 10 PC's	
CLASSE REAL	Fêmea	90,5% 19	9,5% 2
	Macho	9,5% 2	90,5% 19
		Fêmea	Macho
		CLASSE PREVISTA	

(b)		Teste (94,4%) 100 Árvores: 10 PC's	
CLASSE REAL	Fêmea	88,9% 8	11,1% 1
	Macho	0,0% 0	100,0% 9
		Fêmea	Macho
		CLASSE PREVISTA	

Figura 28: Matrizes de confusão de validação LOOCV (a) e teste (b) para o pássaro Curió (1800 a 800 cm^{-1}).

Os resultados indicam que a acurácia de teste foi de 94,4 %, um valor relativamente maior em relação a acurácia de treinamento (90,5%), o que sugere que o modelo está generalizando bem, embora a performance de treino abaixo do teste possa indicar um leve subajuste (*underfitting*) inicial. O modelo foi capaz de prever melhor a classe “Macho” no teste do que a classe “Fêmea”, enquanto se manteve equilibrado no treinamento. Logo, enfrentou dificuldades na classificação devido a sobreposição de amostras de classes diferentes no mesmo grupo, assim como o desbalanceamento entre as classes (“Macho” e “Fêmea”). Contudo, apesar da relativa baixa acurácia de treino e do desbalanceamento, o resultado é promissor, devido a tendência de formação de grupos e alta acurácia.

5.4 *Psittacula krameri* (Ring Neck)

Na Figura 29 a média e o desvio padrão dos espectros do pássaro *Ring Neck* são comparados por classe (“Fêmea” e “Macho”). Nota-se bastante semelhança entre os espectros, sendo a aplicação de rotinas de *machine learning* fundamentais para identificar e esclarecer suas diferenças.

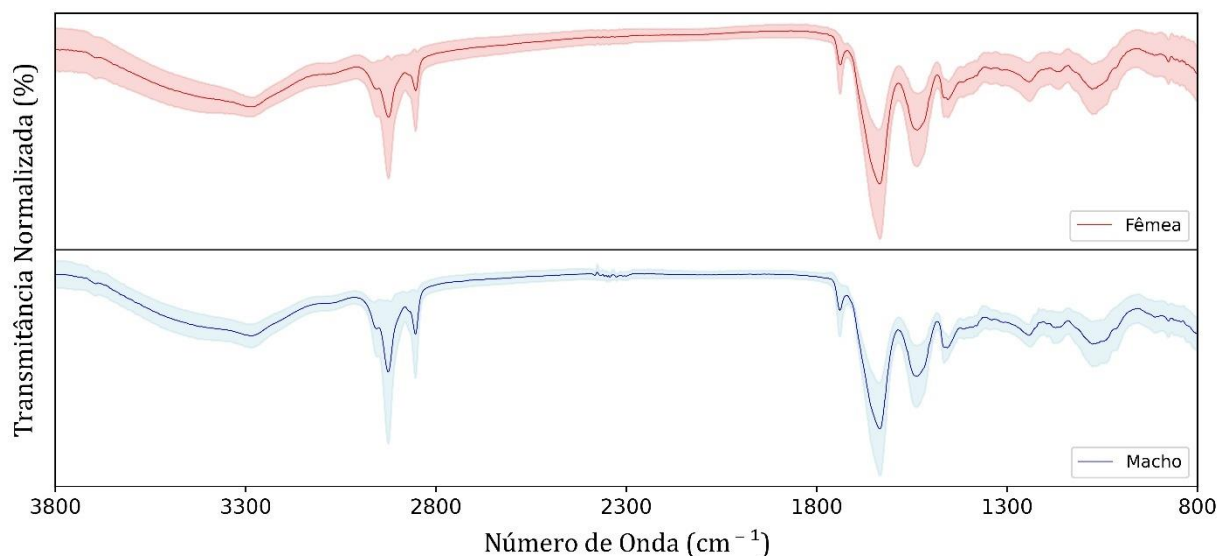


Figura 29: Média e desvio padrão dos espectros FTIR separados por sexo para o pássaro *Ring Neck*.

Comparando as classes “Fêmea” e “Macho” para o pássaro “*Ring Neck*”, verifica-se que as médias têm aspecto semelhante, porém, foram observadas diferenças nos espectros médios nas regiões em torno de 1700, 1400, 1200, 1050, 1000, 900 e 850 cm^{-1} . Contudo, não é possível atribuir essas diferenças exclusivamente ao sexo das aves. Assim, torna-se necessária a aplicação de técnicas de *machine learning* para elucidar essas diferenças. Foi observado um desvio padrão maior nas amostras da classe Fêmea, assim como nas amostras anteriores que pode influenciar a variância dos dados e consequentemente a precisão do modelo.

A Figura 30 apresenta os *scores* e *loadings* do PCA para o intervalo de 3800 a 800 cm^{-1} para o pássaro *Ring Neck*. Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

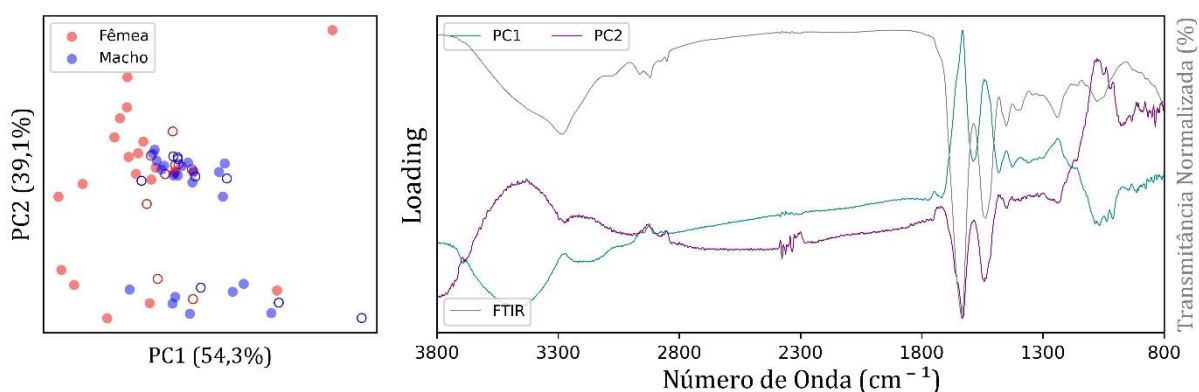


Figura 30: Scores e Loadings para o pássaro *Ring Neck* (3800 a 800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

Nos scores, observa-se a tendência de formação de grupos, apesar da sobreposição das classes “Macho” e “Fêmea”, que para o pássaro “*Ring Neck*” foi muito mais evidente que para os demais. Além disso, verifica-se a formação de subgrupos em ambas as classes. Os loadings apresentam características diferentes ao anteriores, indicando contribuições nas bandas entre 3000 e 2800 cm^{-1} e 1800 e 1300 cm^{-1} . Foram realizados treinamentos nos intervalos de entre 3000 e 2800 cm^{-1} e 1800 até 800 cm^{-1} , com o objetivo de identificar o intervalo espectral mais adequado.

As Figuras a seguir apresentam as acurácias obtidas pelo classificador *Random Forest* por treinamento LOOCV, para diferentes quantidades de árvores e PC's, para os intervalos: 3800 a 800 cm^{-1} (Figura 31), 3000 a 2800 cm^{-1} (Figura 32) e 1800 a 800 cm^{-1} (Figura 33) para o pássaro “*Ring Neck*”.

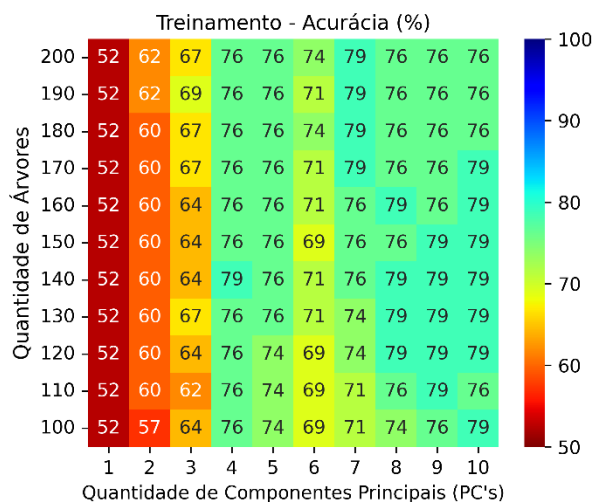


Figura 31: Acurácias em função da quantidade de PC's e árvores para o pássaro *Ring Neck* (3800 a 800 cm^{-1}).

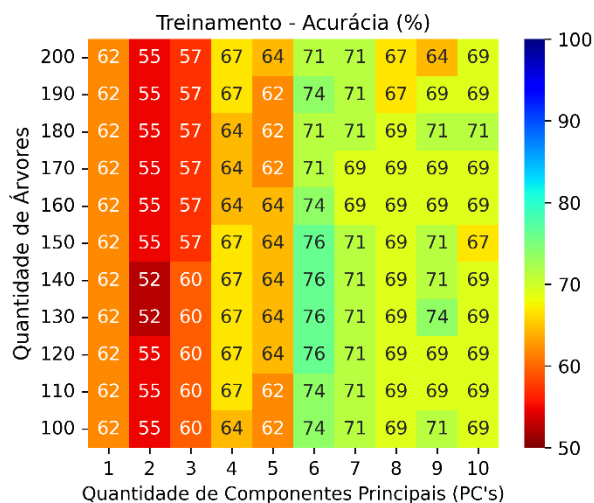


Figura 32: Acurácias em função da quantidade de PC's e árvores para o pássaro *Ring Neck* (3000 a 2800 cm^{-1}).

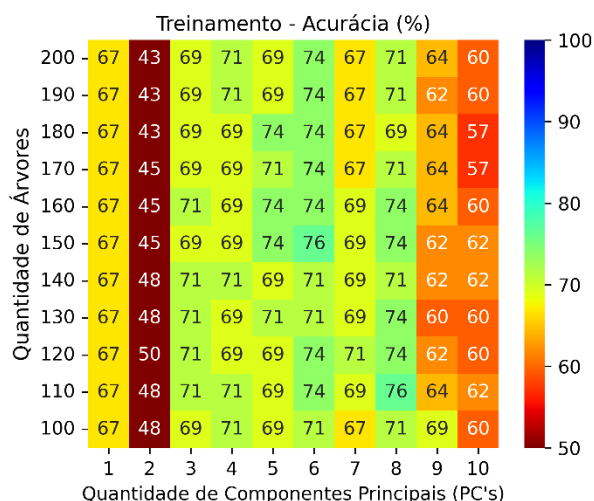


Figura 33: Acurácias em função da quantidade de PC's e árvores para o pássaro *Ring Neck* (1800 a 800 cm^{-1}).

Observa-se que, de modo geral, o intervalo de 3000 a 2800 cm^{-1} apresentou o melhor desempenho em relação aos demais, com a maior acurácia alcançada, de 76,2% com 120 árvores e 6 PC's.

A Figura 34 apresenta os *scores* e *loadings* do PCA para o intervalo de 3000 a 2800 cm^{-1} , região com o melhor desempenho para análise de sexagem do pássaro *Ring Neck*. Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

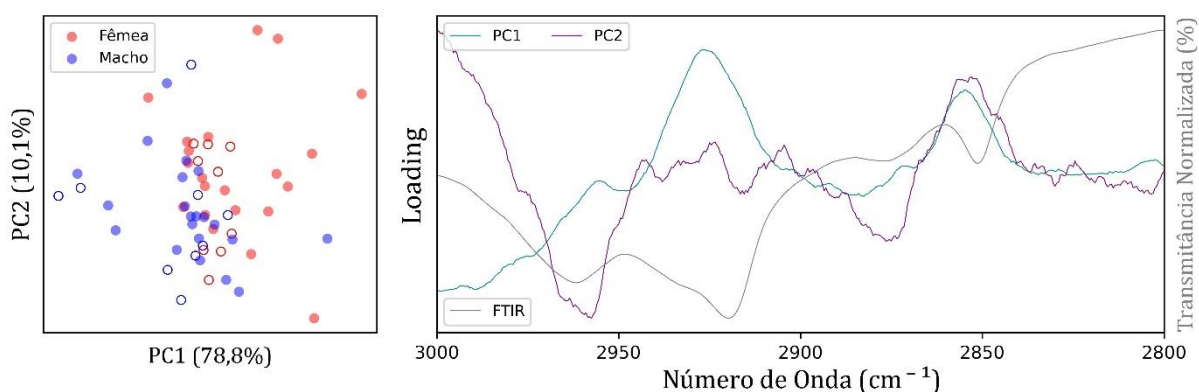


Figura 34: *Scores* e *Loadings* para o pássaro *Ring Neck* (3000 a 2800 cm^{-1}). Os círculos são as amostras de treino utilizadas no PCA, as circunferências são as amostras de teste projetadas neste mesmo espaço e a média dos espectros é exibida na cor cinza.

Observando os *scores*, verifica-se uma grande variância em ambas as classes apesar da tendência de formação de grupos. Os grupos correspondem às classes estudadas, porém com sobreposição. Os *loadings* da PC1 indicam contribuição das três principais bandas deste intervalo com picos em torno de 2955, 2925 e 2855 cm^{-1} . Os *loadings* da PC2 indicam as mesmas contribuições e outras contribuições próximas. Os resultados indicam que a composição ou estrutura das penas apresenta diferenças entre machos e fêmeas, principalmente em relação à queratina (73,74).

Na Figura 35 as matrizes de confusão de validação por LOOCV (a) e teste (b) são comparadas. Cada linha corresponde ao conjunto real e as colunas são as previsões feitas pelo classificador. A acurácia é a média das acurácias apresentadas na diagonal.

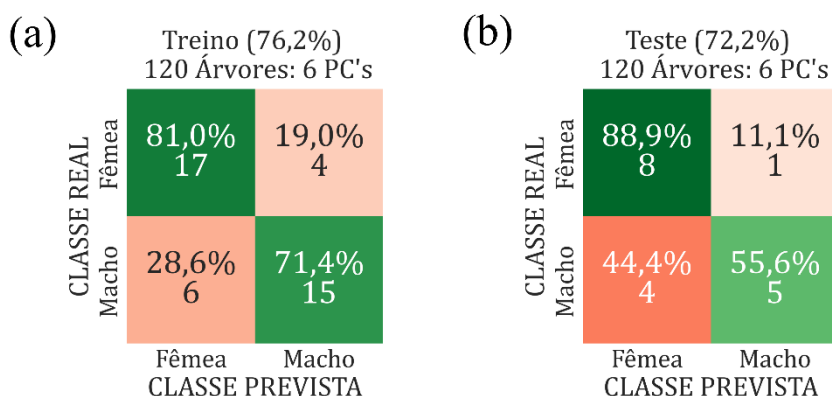


Figura 35: Matrizes de confusão de validação LOOCV (a) e teste (b) para o pássaro *Ring Neck* (3000 a 2800 cm^{-1}).

Os resultados indicam que a acurácia de teste foi de 72,2%, um valor muito próximo da acurácia de treinamento (76,2%), o que sugere que o modelo apresenta boa generalização. Apesar da acurácia ser inferior aos resultados anteriores, o resultado é promissor, considerando a tendência de formação de grupos. Já a formação de subgrupos precisa ser mais bem estudada.

A Tabela 2 apresenta o resumo dos resultados, por pássaro e intervalo analisado. Nela podemos comparar as acurácias e os parâmetros utilizados no treino dos classificadores.

Tabela 2: Resumo dos resultados.

Pássaro	Intervalo (cm ⁻¹)	Treinamento Acurácia (%)	Teste Acurácia (%)	Quantidade de Árvores	Quantidade de PC's
Bicudo	3800 - 800	88,1	77,8	100	7
Bicudo	3000 - 2800	97,6	94,4	110	7
Bicudo	1800 - 800	97,6	77,8	130	7
Calopsita	3800 - 800	78,6	72,2	100	5
Calopsita	3000 - 2800	83,3	77,8	110	7
Calopsita	1800 - 800	81,0	66,7	190	10
Curió	3800 - 800	78,6	88,9	140	7
Curió	3000 - 2800	88,1	66,7	110	10
Curió	1800 - 800	90,5	94,4	100	10
<i>Ring Neck</i>	3800 - 800	78,6	66,7	140	4
<i>Ring Neck</i>	3000 - 2800	76,2	72,2	120	6
<i>Ring Neck</i>	1800 - 800	76,2	72,2	150	6

Na Tabela 2 observa-se que o intervalo de 1800 a 800 cm⁻¹ apresenta resultados promissores, com acurácias de treinamento muito próximas àquela obtidas no intervalo de 3000 até 2800 cm⁻¹. No entanto, as acurácias de teste são muito menores que as de treinamento (exceto para o pássaro Curió), indicando que o modelo não obteve boa generalização nesse intervalo espectral.

Esse trabalho foi publicado na revista ACS Omega em outubro de 2025 com o título “*Early and Noninvasive Bird Gender Identification by ATR-FTIR Spectra Coupled with a Random Forest Algorithm*” – Fator de impacto 4.3 e Qualis A2.

6 CONCLUSÃO

O presente estudo avaliou o potencial da técnica de espectroscopia no Infravermelho por Transformada de Fourier por Reflectância Total Atenuada (ATR-FTIR), combinada com a Análise de Componentes Principais (PCA) e o algoritmo de aprendizado de máquina Floresta Aleatória (RF), para a determinação do sexo de diferentes espécies de aves precoces. As análises foram conduzidas utilizando amostras de quatro espécies de aves: o Bicudo (*Oryzoborus maximiliani*), o Curió (*Oryzoborus angolensis*), a Calopsita (*Nymphicus hollandicus*) e o Periquito-de-colar (*Psittacula krameri*, também citado como *Ring Neck*).

A avaliação espectral revelou que os espectros do Bicudo e do Curió apresentaram maior semelhança entre si nas bandas em torno de 1390, 1415 e 1195 cm^{-1} . A Calopsita e o Periquito-de-colar também demonstraram semelhança mútua nessa mesma região. De forma geral, o intervalo mais promissor em todas as análises situou-se entre 3000 e 2800 cm^{-1} , região na qual foram detectadas as bandas vibracionais características dos grupos funcionais da queratina, com picos em 2955, 2925 e 2870 cm^{-1} . Embora esse intervalo apresente poucas bandas, ele revelou informações biológicas altamente relevantes para a diferenciação biológica. Por outro lado, a região entre 1800 e 800 cm^{-1} mostrou resultados iniciais promissores, porém com baixa capacidade de generalização nos testes, demandando otimizações futuras.

Nos testes de validação externa, os modelos de aprendizado de máquina treinados com as regiões espectrais otimizadas conseguiram distinguir o sexo das aves de forma eficaz na maioria dos casos. Os melhores resultados foram obtidos para as espécies *O. maximiliani* (Bicudo) e *O. angolensis* (Curió), que alcançaram uma acurácia de 94,4%. Desempenhos satisfatórios e promissores também foram observados para *N. hollandicus* (Calopsita), com 77,8% e para *P. krameri* (Periquito-de-colar), com 72,2%.

Apesar do sucesso dos classificadores, alguns fatores desconhecidos impactaram o desempenho do sistema e exigem investigação adicional. No PCA, observou-se uma maior variância na classe “Fêmea” e a formação de subgrupos distintos na classe “Macho”, nuances que precisam ser compreendidas para permitir a expansão segura do método.

Em suma, os resultados consolidam a espectroscopia ATR-FTIR como uma forte candidata para complementar ou até substituir os métodos moleculares tradicionais na identificação do sexo de aves. Para transformar essa técnica em uma aplicação comercial viável,

robusta e escalável, recomenda-se a realização de estudos adicionais. O avanço desta pesquisa deve focar na expansão do banco de dados, na amostragem de diversos tipos de penas e no teste de estratégias alternativas de aprendizado de máquina.

7 SUGESTÕES DE TRABALHOS FUTUROS

- Uma possível linha de investigação envolve a análise das árvores de decisão, permitindo a avaliação da porcentagem de árvores que classificam corretamente uma espécie. Poder-se-ia estabelecer um número mínimo de árvores necessárias para a classificação; caso a amostra não atenda a esse critério, ela seria considerada "indefinida";
- Outra estratégia seria identificar amostras com alta dificuldade de classificação e criar uma classe adicional para "indefinidos", treinando o algoritmo com três classes. Essa abordagem tem a vantagem de manter as fases de treinamento e validação consistentes, sem a necessidade de pós-processamento. No entanto, ela requer um número substancial de amostras difíceis para treinar o modelo de forma eficaz.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] ELNOMROSY, S. M. et al. Application of Loop-Mediated Isothermal Amplification (LAMP) in Sex Identification of Parrots Bred in Egypt. **Biology**, v. 11, n. 4, p. 565, 2022.
- [2] TURCU, M.-C. et al. A comparison of feathers and oral swab samples as DNA sources for molecular sexing in companion birds. **Animals**, v. 13, n. 3, p. 525, 2023.
- [3] ELLEGREN, H.; FRIDOLFSSON, A.-K. Male-driven evolution of DNA sequences in birds. **Nature Genetics**, v. 17, n. 2, p. 182–184, out. 1997.
- [4] GRIFFITHS, R. et al. A DNA test to sex most birds. **Molecular ecology**, v. 7, n. 8, p. 1071–1075, 1998.
- [5] KNACKFUSS, F. Sexagem de Aves da Espécie Amazona aestiva (Papagaio Verdadeiro) por meio de técnica de PCR. **Pubvet**, v. 14, n. 06, 20 jul. 2020.
- [6] DEL PUERTO, F. et al. Identificación molecular del sexo en 9 especies de aves del Centro de Investigación en Animales Silvestres de la hidroeléctrica de ITAIPU, lado paraguayo. **Memorias del Instituto de Investigaciones en Ciencias de la Salud**, v. 15, n. 3, p. 89–92, 2017.
- [7] LEITE, A. A.; FERREIRA, J. L.; DE OLIVEIRA, R. O. R. G. Técnica de sexagem por laparoscopia em arara-canindé (Ara ararauna): Laparoscopic sexing technique in the canindé macaw (Ara ararauna). **Brazilian Journal of Animal and Environmental Research**, v. 5, n. 4, p. 3641–3643, 24 out. 2022.
- [8] STEINER, G. et al. **Bird sexing by Fourier transform infrared spectroscopy**. . Em: BIOMEDICAL VIBRATIONAL SPECTROSCOPY IV: ADVANCES IN RESEARCH AND INDUSTRY. SPIE, 2010.
- [9] STEINER, G. et al. Sexing of turkey poults by Fourier transform infrared spectroscopy. **Analytical and Bioanalytical Chemistry**, v. 396, n. 1, p. 465–470, jan. 2010.
- [10] STEINER, G. et al. Label free molecular sexing of monomorphic birds using infrared spectroscopic imaging. **Talanta**, v. 150, p. 155–161, abr. 2016.
- [11] JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255–260, 2015.
- [12] RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [s.l.] Pearson, 2016.
- [13] MARKET.US. **Birdwatching Tourism Market Size, Growth, Trends and Forecast 2024–2033**. Disponível em: <https://market.us/report/birdwatching-tourism-market/>. Acesso em: 18 jan. 2026.

- [14] PEREIRA, D. C. et al. Molecular sexing in the formation of pairs of blue-and-yellow macaw (*Ara ararauna*) in reintroduction programs. **Research, Society and Development**, v. 10, n. 10, p. 570101019330–570101019330, 2021.
- [15] MAROCHI, M. Z. et al. Dimorfismo sexual em *Hepatus pudibundus* (Crustacea, Decapoda, Brachyura). **Iheringia. Série Zoologia**, v. 106, 2016.
- [16] GRANDO, A. P. Utilização de tomografia por ressonância magnética nuclear para sexagem de aves silvestres sem dimorfismo sexual. 2002.
- [17] SANTOS, M. H. B. DOS et al. Sexagem fetal em ovelhas Santa Inês por ultrasonografia. **Ciência Rural**, v. 36, p. 573–578, 2006.
- [18] OLIVEIRA, R. et al. Sexagem fetal em equinos. **Rev. Bras. Reprod. Anim.**, Belo, 2014.
- [19] SILVA, N. S. R. et al. Determinação de proteínas do plasma seminal usando cromatografia de exclusão molecular em touros jovens. **Brazilian Journal of Animal and Environmental Research**, v. 4, n. 2, p. 1874–1882, 2021.
- [20] VIEIRA, J.; COELHO, E.; OLIVEIRA, D. Sexagem molecular em aves silvestres. **Revista Brasileira de Reprodução Animal**, v. 33, n. 2, p. 66–70, 2009.
- [21] ANTONIO, E.; FRAGA, R. E.; TOMAZI, L. Sexagem molecular em araras vermelhas e Centros de Triagem de Animais Silvestres: Revisão. **PubVet**, v. 15, n. 11, p. 1–10, 2021.
- [22] AMARO JÚNIOR, E.; YAMASHITA, H. Aspectos básicos de tomografia computadorizada e ressonância magnética. **Brazilian Journal of Psychiatry**, v. 23, p. 2–3, 2001.
- [23] DIAS, E. A.; OLIVEIRA, C. DE. Psittacine sex determination by radioimmunoassay (RIA) of sex steroids using fecal samples. 2006.
- [24] KANCHANAPHUM, P. Identification of human DNA by loop-mediated isothermal amplification (LAMP) technique combined with white ring precipitation of Cu (OH)₂. **Songklanakarin Journal of Science and Technology**, v. 40, n. 4, p. 738–742, 2018.
- [25] RASO, T.; WERTHER, K. Sexagem cirúrgica em aves silvestres. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 56, p. 188–192, 2004.
- [26] CERIT, H.; AVANUS, K. Sex determination by CHDW and CHDZ genes of avian sex chromosomes in *Nymphicus hollandicus*. **Turkish Journal of Veterinary & Animal Sciences**, v. 31, n. 6, p. 371–374, 2007.
- [27] HE, Y. et al. Deep learning image segmentation reveals patterns of UV reflectance evolution in passerine birds. **Nature Communications**, v. 13, n. 1, p. 5068, 2022.

- [28] MEHDIZADEH, S. A. et al. Early determination of pharaoh quail sex after hatching using machine vision. **Bulletin of Environment, Pharmacology and Life Sciences**, p. 5–11, 2014.
- [29] HARZ, M. et al. Minimal invasive gender determination of birds by means of UV-resonance Raman spectroscopy. **Analytical chemistry**, v. 80, n. 4, p. 1080–1086, 2008.
- [30] HOLLAS, J. M. **Modern spectroscopy**. [s.l.] John Wiley & Sons, 2004.
- [31] SALA, O. Fundamentos da espectroscopia raman no infravermelho. n. 2, 2008.
- [32] LARKIN, P. **Infrared and Raman spectroscopy: principles and spectral interpretation**. [s.l.] Elsevier, 2017.
- [33] BUENO, W. A. Manual de espectroscopia vibracional. 1990.
- [34] OLIVEIRA, R. A. DE; MARTINS, A. F. P.; SILVA, A. P. B. DA. Thomas Young e a teoria ondulatória da luz no início do século XIX: aspectos conceituais e epistemológicos. **Revista Brasileira de Ensino de Física**, v. 41, 2018.
- [35] ROGALSKI, A. History of infrared detectors. **Opto-electronics review**, v. 20, p. 279–308, 2012.
- [36] PAVIA, D. L. et al. **Introdução à espectroscopia**. [s.l.] Cengage Learning, 2010.
- [37] STUART, B. H. **Infrared spectroscopy: fundamentals and applications**. [s.l.] John Wiley & Sons, 2004.
- [38] VALAND, R. et al. A review of Fourier Transform Infrared (FTIR) spectroscopy used in food adulteration and authenticity investigations. **Food Additives & Contaminants: Part A**, v. 37, n. 1, p. 19–38, 2020.
- [39] MILEWSKA, A. et al. In-line Monitoring of Protein Concentration With MIR Spectroscopy During UFDF. **Engineering in Life Sciences**, v. 23, n. 2, 2022.
- [40] ABRAMOVICH, A.; SHULZINGER, A. Diagnostic and Analysis of Human Sperm Characteristics Using Fourier Transform Infrared Spectroscopy. **Open Journal of Urology**, v. 05, n. 06, p. 97–101, 2015.
- [41] ZHANG, S. et al. Fiber Optic Probe-based ATR-FTIR Spectroscopy for Rapid Breast Cancer Detection: A Pilot Study. **Journal of Biophotonics**, v. 16, n. 11, 2023.
- [42] CHRISTENSEN, J. B. et al. Intrinsic Spectral Resolution Limitations of QEPAS Sensors for Fast and Broad Wavelength Tuning. **Sensors**, v. 20, n. 17, p. 4725, 2020.

- [43] HASHIMOTO, K.; IDEGUCHI, T. Phase-Controlled Fourier-Transform Spectroscopy. **Nature Communications**, v. 9, n. 1, 2018.
- [44] FERREIRA, M. et al. Quimiometria I: calibração multivariada, um tutorial. **Química Nova**, v. 22, p. 724–731, 1999.
- [45] FERREIRA, M. M. C. **Quimiometria: conceitos, métodos e aplicações**. [s.l.] Editora da UNICAMP, 2015.
- [46] GEMPERLINE, P. **Practical guide to chemometrics**. [s.l.] CRC press, 2006.
- [47] BARROS NETO, B. DE; SCARMINIO, I. S.; BRUNS, R. E. 25 anos de quimiometria no Brasil. **Química Nova**, v. 29, p. 1401–1406, 2006.
- [48] KOWALSKI, B. R. Chemometrics: views and propositions. **Journal of Chemical Information and Computer Sciences**, v. 15, n. 4, p. 201–203, 1975.
- [49] BRERETON, R. G. **Applied chemometrics for scientists**. [s.l.] John Wiley & Sons, 2007.
- [50] LYRA, W. DA S. et al. Classificação periódica: um exemplo didático para ensinar análise de componentes principais. **Química Nova**, v. 33, p. 1594–1597, 2010.
- [51] CANEVAROLO JR. Técnicas de caracterização de polímeros. **São Paulo: Artliber Editora**, 2007.
- [52] RIBEIRO, M. C. et al. Discrimination of maize transgenic and non-transgenic varieties by laser induced spectroscopy (LIBS) and machine learning algorithms. **Microchemical Journal**, v. 203, p. 110898, 2024.
- [53] AMIGO, J. M. Data mining, machine learning, deep learning, chemometrics: Definitions, common points and trends (Spoiler Alert: VALIDATE your models!). **Brazilian Journal of Analytical Chemistry**, v. 8, n. 32, p. 45–61, 2021.
- [54] PIERCE, K. M.; MOHLER, R. E. A review of chemometrics applied to comprehensive two-dimensional separations from 2008–2010. **Separation & Purification Reviews**, v. 41, n. 2, p. 143–168, 2012.
- [55] DE LIMA, M. D.; BARBOSA, R. Methods of authentication of food grown in organic and conventional systems using chemometrics and data mining algorithms: A review. **Food Analytical Methods**, v. 12, p. 887–901, 2019.
- [56] ROSENBERG, L. H. et al. Multivariate Meta-Analysis of Proteomics Data From Human Prostate and Colon Tumours. **BMC Bioinformatics**, v. 11, n. 1, 2010.
- [57] XU, Y.; GOODACRE, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. **Journal of Analysis and Testing**, v. 2, n. 3, p. 249–262, 2018.

- [58] MÜLLER, A. C.; GUIDO, S. **Introduction to machine learning with Python: a guide for data scientists**. [s.l.] O'Reilly Media, Inc., 2016.
- [59] WESTAD, F.; MARINI, F. Validation of chemometric models—a tutorial. **Analytica Chimica Acta**, v. 893, p. 14–24, 2015.
- [60] KORJUS, K.; HEBART, M. N.; VICENTE, R. An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable. **Plos One**, v. 11, n. 8, p. e0161788, 2016.
- [61] KELLEHER, J. D.; MAC NAMEE, B.; D'ARCY, A. **Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies**. [s.l.] MIT press, 2020.
- [62] REIS, W. et al. Enhancing early identification of high-fertile cattle females using infrared blood serum spectra and machine learning. **Scientific Reports**, v. 14, n. 1, p. 19446, 2024.
- [63] BREIMAN, L. Random forests. **Machine learning**, v. 45, p. 5–32, 2001.
- [64] BREIMAN, L. **Classification and regression trees**. New York: Routledge, 2017.
- [65] GONÇALVES, V. P. Monitoramento da estabilidade oxidativa de biodiesel empregando espectroscopia vibracional associada a ferramentas quimiométricas. 2022.
- [66] BROWN, S. Introduction to multivariate statistical analysis in chemometrics. **Appl. Spectrosc.**, v. 64, n. 4, p. 112A-112A, 2010.
- [67] LOUPPE, G. Understanding random forests: From theory to practice. **arXiv preprint arXiv:1407.7502**, 2014.
- [68] CIOCCIA, G. et al. Probabilistic-based identification of gunshot residues (GSR) using Laser-Induced Breakdown Spectroscopy (LIBS) and Support Vector Machine (SVM) algorithm. **Microchemical Journal**, v. 207, p. 112142, 2024.
- [69] BEATTIE, J. R.; ESMONDE-WHITE, F. W. Exploration of principal component analysis: deriving principal component analysis visually using spectra. **Applied Spectroscopy**, v. 75, n. 4, p. 361–375, 2021.
- [70] SALMAN, H. A.; KALAKECH, A.; STEITI, A. Random forest algorithm overview. **Babylonian Journal of Machine Learning**, v. 2024, p. 69–79, 2024.
- [71] SCHONLAU, M.; ZOU, R. Y. The random forest algorithm for statistical learning. **The Stata Journal**, v. 20, n. 1, p. 3–29, 2020.
- [72] TESFAYE, T.; SITHOLE, B.; RAMJUGERNATH, D. Valorisation of chicken feather barbs: Utilisation in yarn production and technical textile applications. **Sustainable Chemistry and Pharmacy**, v. 8, p. 38–49, 2018.

- [73] PAN, Y. et al. Bio-molecular analyses enable new insights into the taphonomy of feathers. **PNAS nexus**, v. 3, n. 9, p. pgae341, 2024.
- [74] SOUZA, G. E. DE Q. et al. Recovery of keratin from feather meal: a new route to valorize an agro-industrial co-product. **Environmental Technology**, p. 1–11, 2024.

APÊNDICE – MATERIAL SUPLEMENTAR

Tabela 1 – Acurácia (%) dos diferentes classificadores aplicados às espécies e regiões espectrais avaliadas

Espécie	Região (cm-1)	LDA	QDA	KNNF	KNNN	KNNC	KNNG6	KNN3	KNNW	SVML	SVM2	SVM3	SVMF	SVMN	SVMC
Bicudo	(3800–800)	73,9	80,4	76,1	80,4	78,3	82,6	78,3	73,9	76,1	63	63	73,9	60,9	4,3
	(3000–2800)	84,8	82,6	71,7	87	82,6	84,8	80,4	80,4	87	69,6	69,6	76,1	73,9	65,2
	(1800–800)	89,1	91,3	91,3	84,8	84,8	87	89,1	84,8	91,3	63	71,7	84,8	63	26,1
Calopsita	(3800–800)	54,3	67,4	71,7	73,9	67,4	67,4	65,2	71,7	69,6	71,7	69,6	56,5	37	13
	(3000–2800)	60,9	58,7	67,4	54,3	52,2	52,2	47,8	58,7	60,9	60,9	65,2	60,9	52,2	0
	(1800–800)	67,4	65,2	80,4	78,3	71,7	73,9	71,7	80,4	71,7	71,7	63	73,9	60,9	28,3
Curió	(3800–800)	30,4	47,8	39,1	30,4	34,8	34,8	32,6	30,4	30,4	41,3	45,7	19,6	8,7	0
	(3000–2800)	45,7	41,3	52,2	47,8	43,5	43,5	43,5	52,2	47,8	13	23,9	50	45,7	8,7
	(1800–800)	34,8	41,3	39,1	34,8	21,7	19,6	32,6	30,4	30,4	45,7	34,8	32,6	52,2	8,7
Ring Neck	(3800–800)	69,6	65,2	54,3	58,7	52,2	58,7	43,5	58,7	71,7	50	60,9	54,3	54,3	17,4
	(3000–2800)	60,9	63	50	50	60,9	58,7	63	54,3	60,9	56,5	54,3	54,3	50	15,2
	(1800–800)	60,9	60,9	67,4	47,8	45,7	43,5	43,5	52,2	58,7	54,3	56,5	67,4	60,9	19,6

Sigla	Descrição	Sigla	Descrição
LDA	<i>Linear Discriminant Analysis</i>	SVML	<i>Support Vector Machine Linear</i>
QDA	<i>Quadratic Discriminant Analysis</i>	SVM2	<i>Support Vector Machine Quadratic</i>
KNNF	<i>K-Nearest Neighbors Fine</i>	SVM3	<i>Support Vector Machine Cubic</i>
KNNN	<i>K-Nearest Neighbors Normal</i>	SVMF	<i>Support Vector Machine Fine Gaussian</i>
KNNC	<i>K-Nearest Neighbors Coarse</i>	SVMN	<i>Support Vector Machine Medium Gaussian</i>
KNNG6	<i>KNN Cosine / Grau 6</i>	SVMC	<i>Support Vector Machine Coarse Gaussian</i>
KNN3	<i>KNN com 3 vizinhos</i>		
KNNW	<i>Weighted KNN</i>		

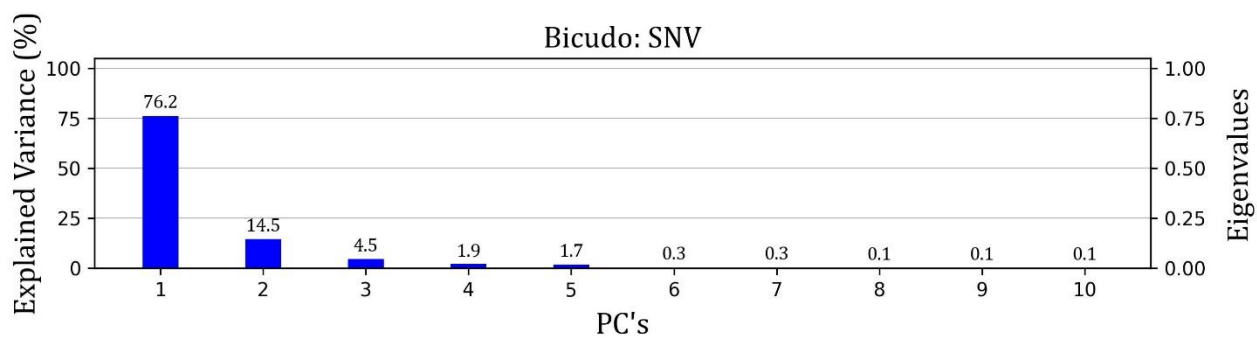


Figura 01 : *Scree Plot* para o intervalo de 3800 até 800 cm^{-1} (Bicudo).

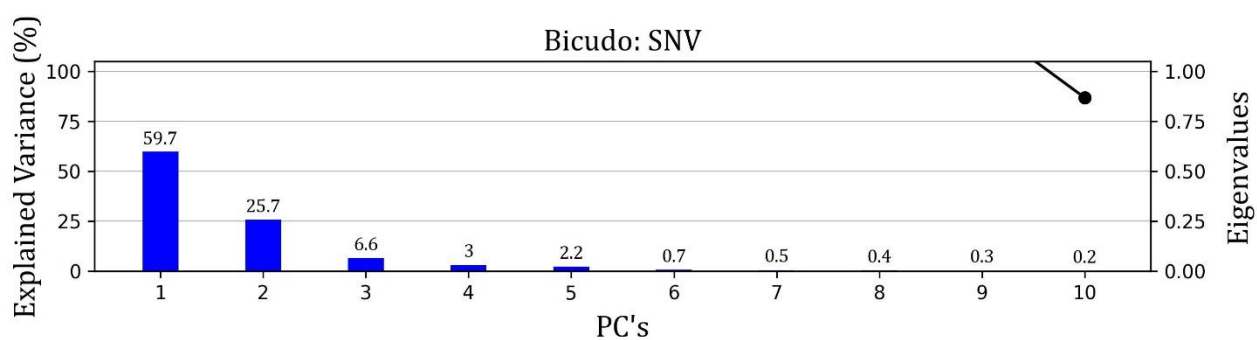


Figura 02 : *Scree Plot* para o intervalo de 1800 até 800 cm^{-1} (Bicudo).

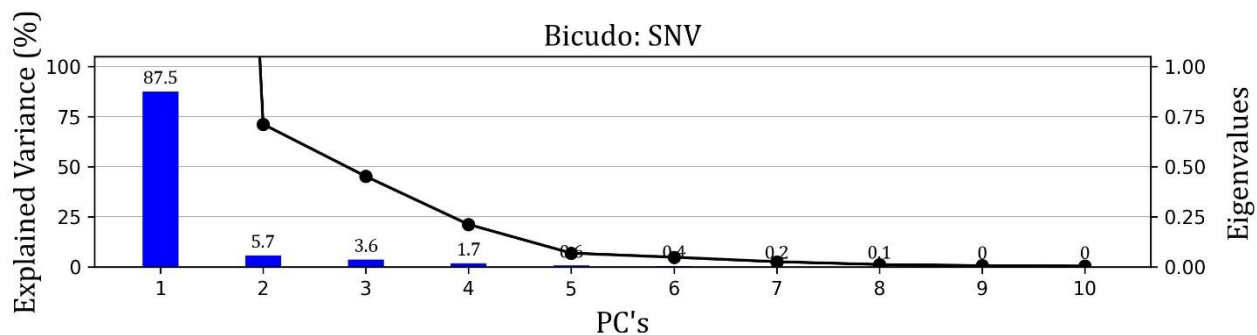


Figura 03 : *Scree Plot* para o intervalo de 3000 até 2800 cm^{-1} (Bicudo).

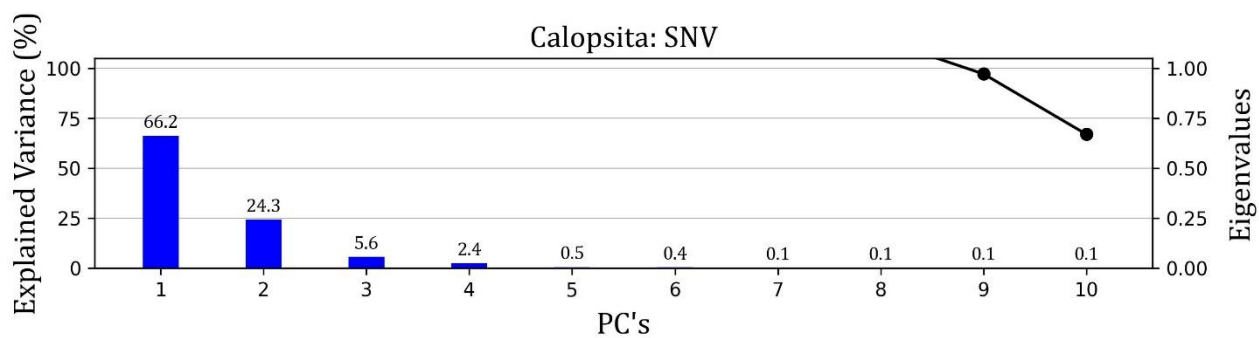


Figura 04 : *Scree Plot* para o intervalo de 3800 até 800 cm^{-1} (Calopsita).

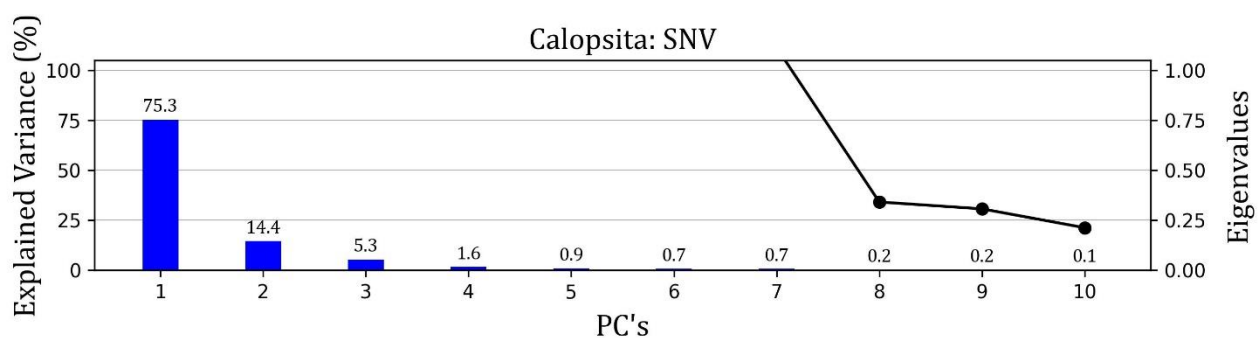


Figura 05 : *Scree Plot* para o intervalo de 1800 até 800 cm^{-1} (Calopsita).

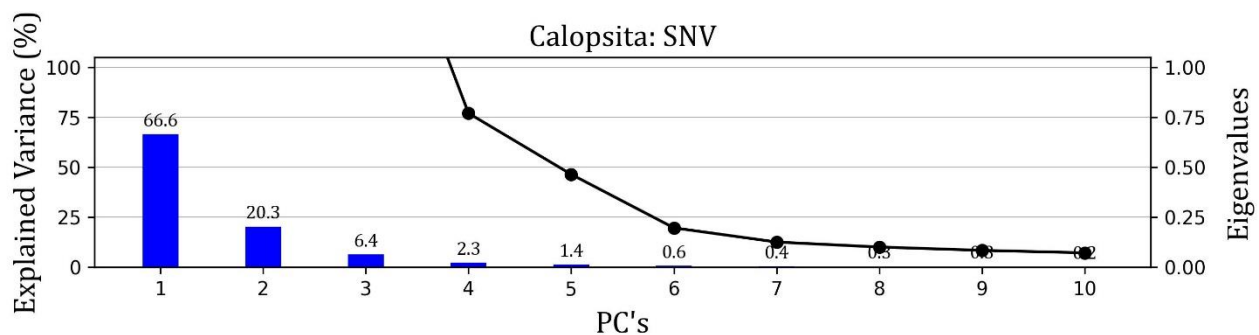


Figura 06 : *Scree Plot* para o intervalo de 3000 até 2800 cm^{-1} (Calopsita).

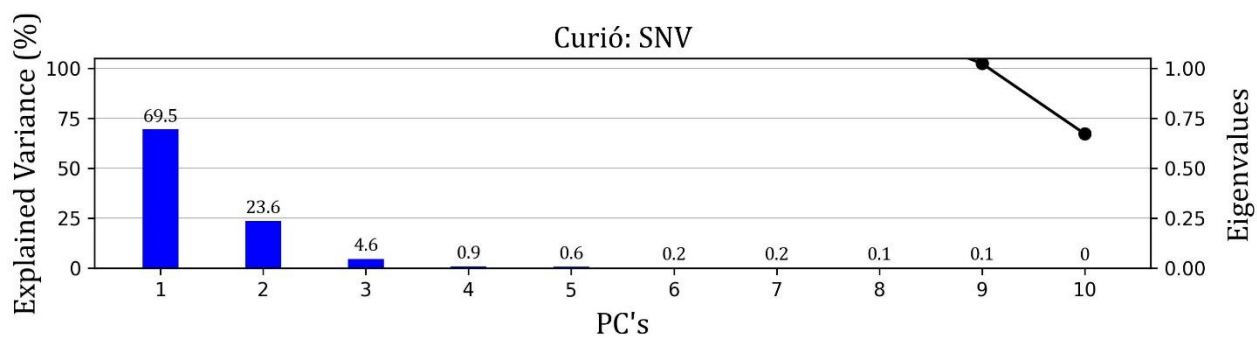


Figura 07 : *Scree Plot* para o intervalo de 3800 até 800 cm^{-1} (Curió).

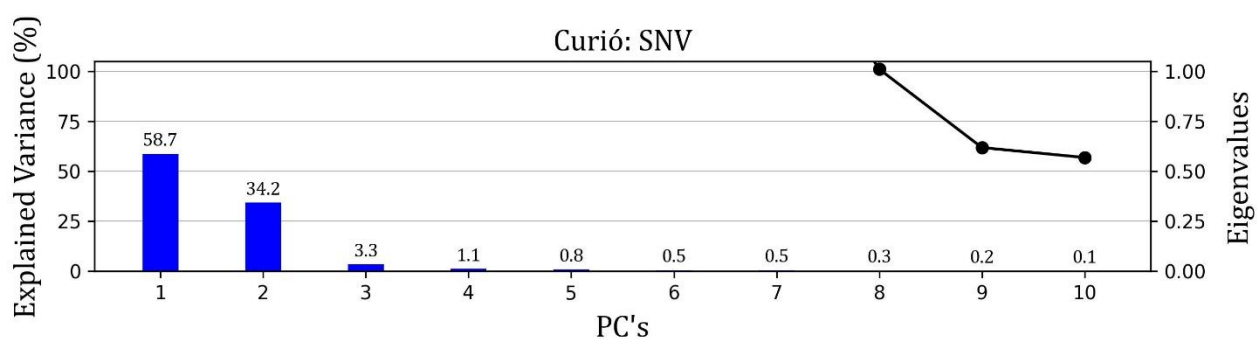


Figura 08 : *Scree Plot* para o intervalo de 1800 até 800 cm^{-1} (Curió).

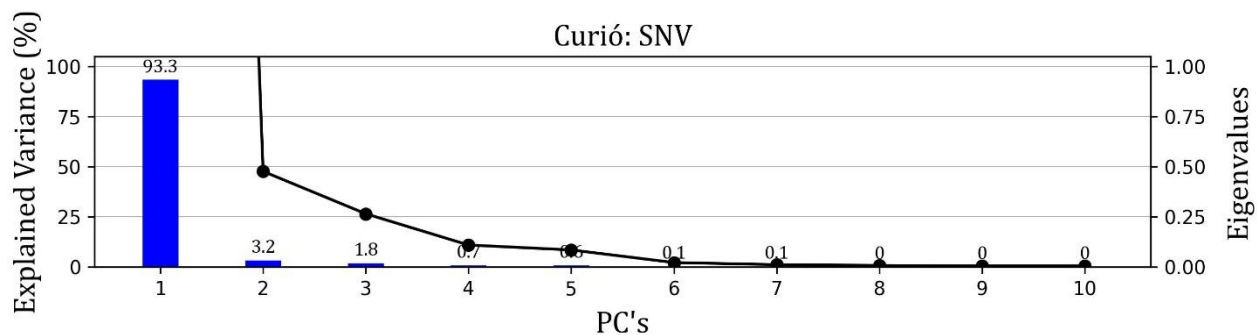


Figura 09 : *Scree Plot* para o intervalo de 3000 até 2800 cm^{-1} (Curió).

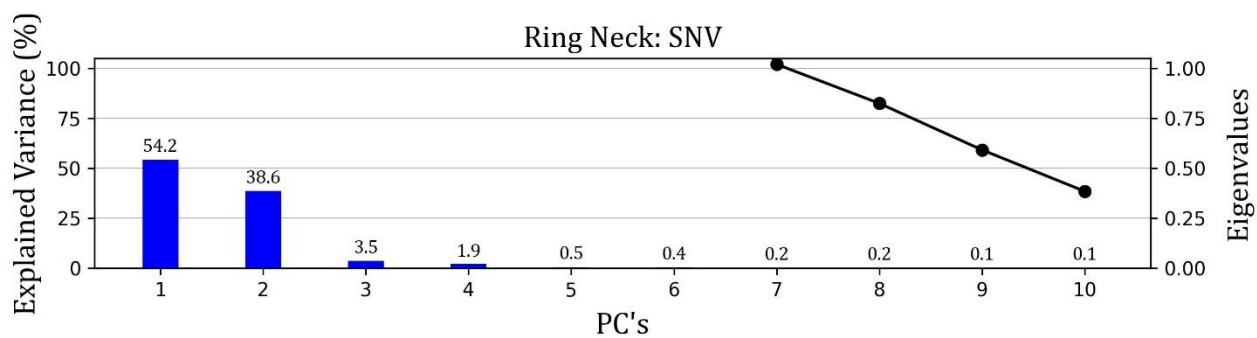


Figura 10 : Scree Plot para o intervalo de 3800 até 800 cm⁻¹ (Ring Neck).

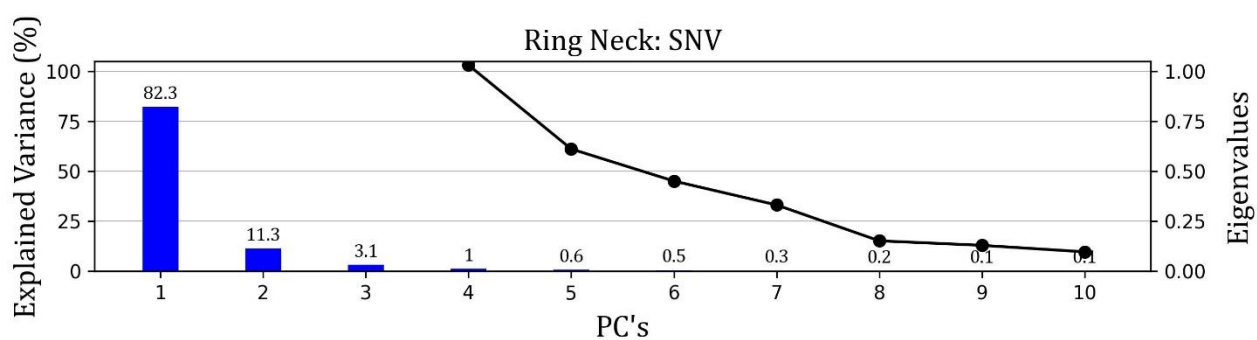


Figura 11 : Scree Plot para o intervalo de 1800 até 800 cm⁻¹ (Ring Neck).

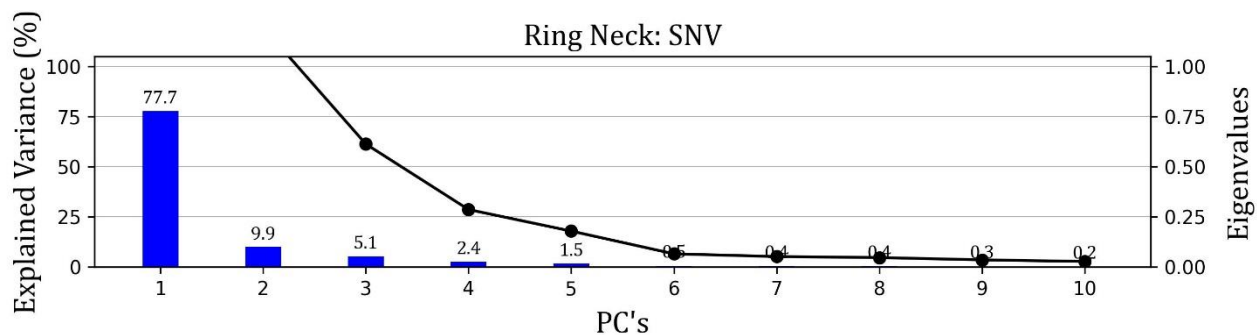


Figura 12 : Scree Plot para o intervalo de 3000 até 2800 cm⁻¹ (Ring Neck).

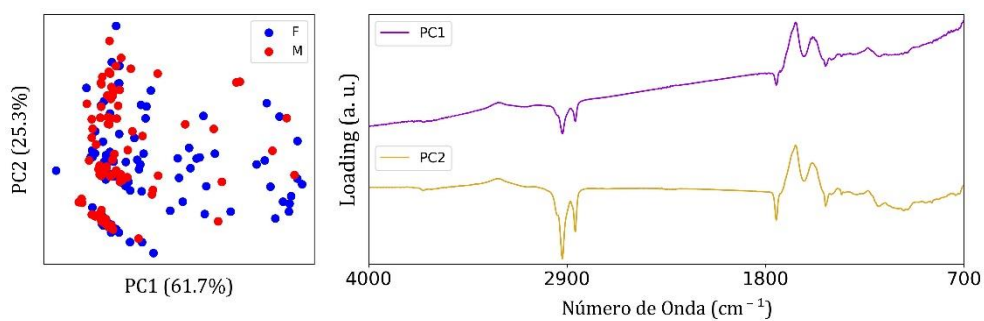


Figura 13 : PCA para o intervalo de 4000 até 800 cm^{-1} (Macho vs. Fêmea).

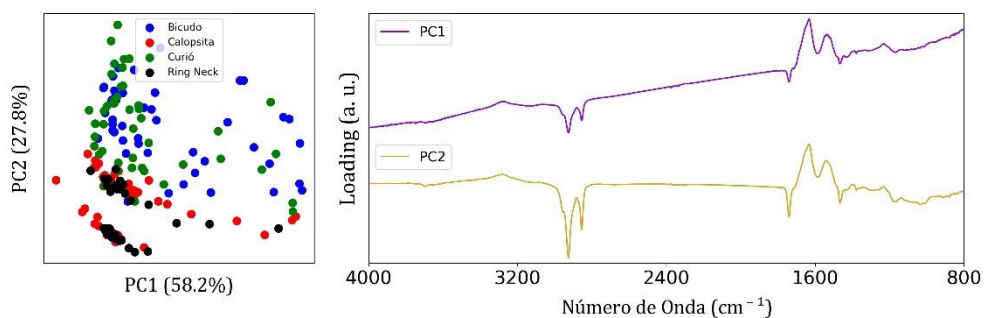


Figura 14 : PCA para o intervalo de 4000 até 800 cm^{-1} (Espécies).

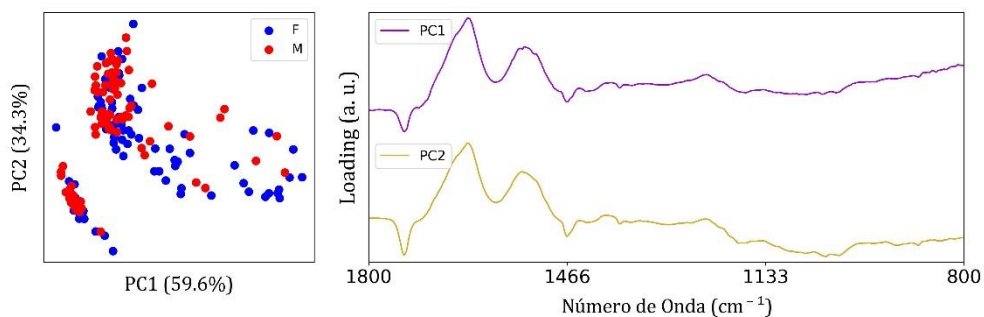


Figura 15 : PCA para o intervalo de 1800 até 800 cm^{-1} (Macho vs. Fêmea).

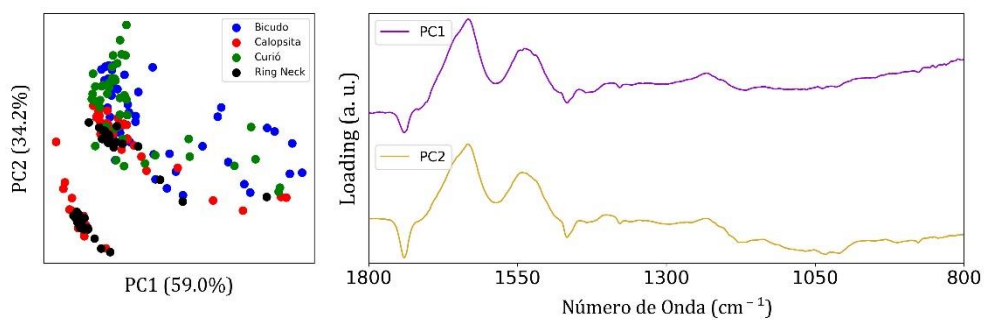


Figura 16 : PCA para o intervalo de 1800 até 800 cm^{-1} (Espécies).

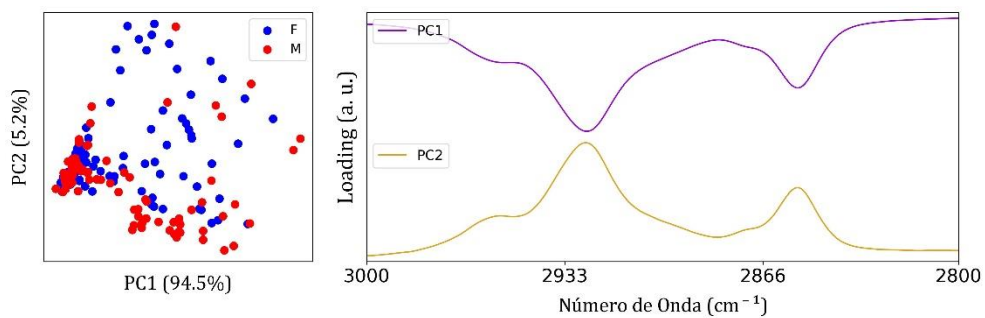


Figura 17 : PCA para o intervalo de 3000 até 2800 cm^{-1} (Macho vs. Fêmea).

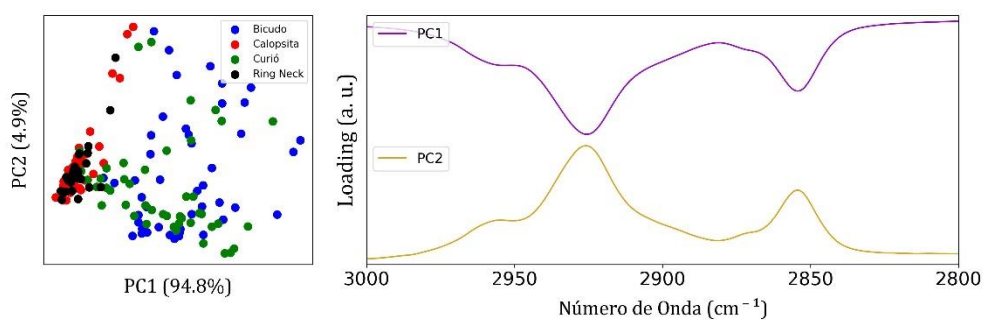


Figura 18 : PCA para o intervalo de 3000 até 2800 cm^{-1} (Espécies).