UNIVERSIDADE FEDERAL DO MATO GROSSO DO SUL
Instituto de Física

VICTOR FIDELIS FERNANDES

DESENVOLVIMENTO DE ROTINAS EM *PYTHON* PARA SELEÇÃO DE VARIÁVEIS EM DADOS ESPECTROSCÓPICOS

Campo Grande – MS
2024

VICTOR FIDELIS FERNANDES

DESENVOLVIMENTO DE ROTINAS EM *PYTHON* PARA SELEÇÃO DE VARIÁVEIS EM DADOS ESPECTROSCÓPICOS

Dissertação apresentada ao Curso de Graduação, em Física, da Universidade Federal de Mato Grosso do Sul como requisito final à obtenção do título de Bacharel em Física.

Orientador: Prof. Dr. Bruno S. Marangoni

Victor Fidelis Fernandes

Campo Grande – MS

2024

VICTOR FIDELIS FERNANDES

DESENVOLVIMENTO DE ROTINAS EM PYTHON PARA SELEÇÃO DE VARIÁVEIS EM DADOS ESPECTROSCÓPICOS

Dissertação apresentada ao Curso de Graduação, em Física, da Universidade Federal de Mato Grosso do Sul como requisito final à obtenção do título de Bacharel em Física.

Campo Grande, MS, 27 de Fevereiro de 2024

COMISSÃO EXAMINADORA

_____

Prof. Dr. Bruno Spolon Marangoni
Universidade Federal de Mato Grosso do Sul

_____

Prof. Dr. Cicero Rafael Cena da Silva
Universidade Federal de Mato Grosso do Sul

_____

Dr. Matheus Cicero da Silva Ribeiro
Universidade Federal de Mato Grosso do Sul

RESUMO

A sexagem de aves, crucial para a avicultura, criações domésticas e preservação de espécies, enfrenta o desafio de identificar o sexo em aves jovens, muitas vezes sem dimorfismo sexual. Diversas técnicas são utilizadas, cada uma com suas vantagens e desvantagens.

Com isso em mente, o presente projeto aborda a identificação de gênero em aves das espécies bicudo, calopsita, curió e *ring neck* através da análise espectral do infravermelho por transformada de Fourier – FTIR, e linguagem *python* algoritmo *Random Forest*. Integrando a espectroscopia para detectar sutis diferenças nas penas, o *Random Forest* é treinado para distinguir entre machos e fêmeas. O sucesso demonstrado nessa abordagem, validado por amostras externas em 100% de acurácia para os bicudos, destaca seu potencial para aplicações interdisciplinares e realça a interseção entre análise de dados e física aplicada.

Palavras-chave: *Python. FTIR.* Espectroscopia. Inteligencia Artificial. *Random Forest*.

ABSTRACT

Bird sexing, crucial for poultry farming, domestic breeding, and species preservation, faces the challenge of identifying the sex in young birds, often without sexual dimorphism. Various techniques are used, each with its advantages and disadvantages.

With this in mind, the present project addresses the gender identification in birds of the species bicudo, cockatiel, curió, and ring neck through the spectral analysis of infrared by Fourier transform – FTIR, and python language Random Forest algorithm. Integrating spectroscopy to detect subtle differences in feathers, the Random Forest is trained to distinguish between males and females. The success demonstrated in this approach, validated by external samples with 100% accuracy for the bicudos, highlights its potential for interdisciplinary applications and emphasizes the intersection between data analysis and applied physics.

Keywords: Python. FTIR. Spectroscopy. Artificial Intelligence. Random Forest.

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

A identificação correta do sexo das aves é fundamental para o sucesso a indústria avícola e criações de aves domésticas, visto que muitas espécies não apresentam distinção sexual em suas características externas, principalmente quando ainda jovens [1][2]. Além de contribuir com a conservação biológica e a divulgação científica, a sexagem também gera dados que ajudam a mitigar os crimes ambientais e o comércio ilegal de aves silvestres [3].

Além disso, a sexagem de aves é extremamente importante para os criadores, não apenas pela reprodução, mas também pelo fato de que em algumas espécies apenas os machos manifestam certas características, como o canto [1].

Algumas técnicas utilizadas atualmente são:

- **Sexagem Molecular por PCR:** Esta técnica se baseia na reação de cadeia de polimerase. É simples, conveniente, barata, rápida e segura [4]. Ela reduz os custos de manutenção de aves jovens, evita a formação de casais do mesmo sexo ou entre parentes próximos, e facilita o manejo genético em criações com sistema de produção de aves por separação por sexo [5].
- **Sexagem por Análise de DNA:** A análise do DNA permite a sexagem de aves com 99,9% de confiabilidade [3]. Além disso, traz outras vantagens em relação a outras técnicas, como a diminuição do estresse e do risco de danos à ave e também o fato de ser aplicada a partir de qualquer idade [6].
- **Laparoscopia:** A laparoscopia tem sido empregada em diversas espécies animais desde a década de 1970 [7]. No entanto, é um procedimento cirúrgico, o que se torna inviável em diversas espécies que são extremamente pequenas.

Além dessas, a espectroscopia é uma técnica que tem se tornado frequente na sexagem, pois é uma técnica não invasiva que pode ser bastante precisa. Embora a sexagem por PCR e a análise de DNA também sejam precisas [4], elas apresentam ainda assim, um custo elevado a longo prazo, principalmente comparado em grandes volumes, em relação a espectroscopia [8].

Porem, a espectroscopia retorna uma quantidade muito grande de dados [9]. Para ajudar nesse processo, uma das abordagens que tem ganhado notoriedade na área é o aprendizado de máquinas [10].

Vários grupos vêm desenvolvendo técnicas com análises multivariadas agregado a espectroscopia para a classificação de comidas [11], para análise de degradação de baterias de lítio [12], para determinar o vigor de sementes de soja [13] e até mesmo para diagnostico de leishmaniose [14].

Isso sugere que a espectroscopia, em conjunto com o aprendizado de máquina, configura-se como uma ferramenta promissora para sexagem. A técnica oferece uma série de vantagens, como precisão, custo-benefício e natureza não invasiva, tornando-a uma alternativa viável às técnicas tradicionais.

2. MATERIAIS E MÉTODOS

A espectroscopia é uma técnica que se baseia na utilização da luz para estudar a composição, a estrutura e as propriedades da matéria. A raiz da palavra, do latim spectrum (imagem, aparição), remete para algo como "observação da imagem oculta". Em uma medida espectroscópica, a amostra a ser estudada é irradiada com um feixe de luz incidente e a análise é feita à luz transmitida, emitida ou difundida pela amostra [15].

A Espectroscopia de Infravermelho por Transformada de Fourier (FTIR) é uma técnica analítica poderosa e amplamente utilizada em diversos setores industriais para identificação, caracterização e quantificação de materiais. O termo FTIR significa "*Fourier transform infrared"* (infravermelho com transformada de Fourier) e é a forma mais comum de espectroscopia de infravermelho.

Em uma experiência de espectroscopia FTIR, quando a radiação infravermelha (IV) passa por uma amostra, parte da radiação é absorvida. Essa absorção ocorre porque os sistemas atômico-moleculares que constituem a matéria apresentam estados de energia discretos. Sistemas atômico-moleculares diferentes apresentam níveis de energia com separações diferentes, dependendo da natureza e composição da amostra.

A análise por espectroscopia FTIR fornece evidências da presença de grupos funcionais presentes na estrutura de uma substância, podendo ser utilizada na identificação de um composto ou para a investigar sua composição química.

Os dados gerados durante uma medida FTIR são geralmente apresentados como um espectro, que é um gráfico da intensidade da luz em função do número de onda (a unidade inversa do comprimento de onda, geralmente expressa em cm^{-1}). O número de onda é uma medida da energia da luz e é diretamente proporcional à frequência, pela Equação 1. Cada pico no espectro representa uma frequência específica de luz que foi absorvida pela amostra e corresponde a uma transição de energia específica dos sistemas atômico-moleculares na amostra.

$$k = \frac{1}{c}\,\nu \tag{1}$$

Onde k representa o número de onda, c é a velocidade da luz e ν é a frequência da onda.

Os intervalos de número de onda que são tipicamente medidos em uma experiência de FTIR, variam dependendo do tipo de amostra e do objetivo do estudo. No entanto, um intervalo comum para a maioria das aplicações de FTIR é de cerca de 4000 cm^{-1} a 400 cm^{-1}, com o intervalo de medida de 0,5 cm^{-1}. Este intervalo cobre a região do infravermelho médio, onde ocorrem a maioria das transições vibracionais moleculares.

Nesse domínio espectral, cada espectro FTIR gerado é uma combinação intrincada de picos espectrais, cada um correspondendo a uma característica molecular específica da amostra. A complexidade da interpretação dos dados é exacerbada pela presença de ruído espectral e pela sobreposição de picos.

A inteligência artificial (IA), especificamente o aprendizado de máquina, pode oferecer uma solução robusta para esses desafios. O aprendizado de máquina é uma subárea da IA que se concentra na construção de sistemas que podem aprender a partir de dados. No contexto da espectroscopia FTIR, um modelo de aprendizado de máquina pode ser treinado em um conjunto de espectros de amostras de penas de aves com seu sexo já conhecido. Durante o treinamento, o modelo aprende a mapear padrões espectrais para as características correspondentes da amostra.

Uma vez treinado, o modelo de aprendizado de máquina pode ser usado para analisar novos espectros e prever os sexos das amostras correspondentes. Isso pode ser feito de maneira eficiente, mesmo para grandes volumes de dados, graças à capacidade do aprendizado de máquina de processar rapidamente grandes conjuntos de dados.

Além disso, o aprendizado de máquina pode identificar padrões sutis nos dados que podem ser difíceis de detectar por meio de métodos de análise convencionais. Isso pode resultar em uma maior precisão na interpretação dos dados espectrais [9].

O pré-processamento de dados é uma etapa fundamental no treinamento de um modelo de aprendizado de máquina, como o *Random Forest*. A limpeza de dados é essencial para garantir que o modelo esteja aprendendo a partir de informações precisas, pois os dados brutos podem conter ruído. Além disso, a normalização é outra etapa crucial que coloca todas as características na mesma escala, evitando viés no modelo devido a diferentes escalas de características.

O *Standard Normal Variate* (SNV) é uma técnica de pré-processamento de dados que é frequentemente usada em aprendizado de máquina para lidar com variações espectrais. O

SNV realiza uma normalização em cada espectro individualmente, o que pode melhorar a precisão do modelo ao lidar com dados espectrais [16].

A técnica SNV é útil na espectroscopia FTIR, onde é usada para corrigir variações de luz dispersa nos dados espectrais [16]. A abordagem SNV efetivamente remove as interferências além de reduzir a variância dentro da classe [17].

Essas técnicas de pré-processamento de dados, como o SNV, são essenciais para construir a maioria dos tipos de modelos de calibração em FTIR. Com uma etapa de pré-processamento bem projetada, o desempenho do modelo pode ser muito melhorado [16].

A detecção de *outliers* também é importante, pois *outliers* são pontos de dados que são significativamente diferentes do restante dos dados e podem distorcer o modelo se não forem tratados adequadamente.

O *Spectrum Angle Mapping* (SAM) é uma técnica eficaz para a remoção de *outliers*. Ele avalia a variância nos ângulos entre os vetores de diferença de um ponto para os outros pontos. Isso pode ser particularmente útil em conjuntos de dados de alta dimensão, onde a mera consideração das distâncias pode ser insuficiente [18].

Comparando com o espectro médio, o SAM tem a vantagem de não depender de qualquer seleção de parâmetro que influencie a qualidade do ranking alcançado. Isso torna o SAM uma ferramenta robusta e confiável para a detecção de *outliers [19]*.

A redução de dimensionalidade é uma técnica importante que pode ser usada para identificar e manter apenas as características mais informativas dos espectros FTIR, melhorando a eficiência e, possivelmente, a performance do modelo.

A Análise de Componentes Principais (PCA – *Principal Component Analysis*) é uma técnica estatística que pode ser usada para reduzir a dimensionalidade. Ela transforma um conjunto de variáveis possivelmente correlacionadas em um conjunto menor de variáveis não correlacionadas chamadas componentes principais. O primeiro componente principal é responsável pela maior parte da variabilidade nos dados, e cada componente subsequente é responsável pela maior parte da variabilidade restante [20].

O PCA funciona calculando a matriz de covariância dos dados para entender a variabilidade. Em seguida, os autovetores e autovalores da matriz de covariância são calculados. Os autovetores (componentes principais) determinam as direções do novo espaço de eixos, e os autovalores determinam suas magnitudes. Em outras palavras, os autovalores explicam a variância dos dados ao longo dos novos eixos [21].

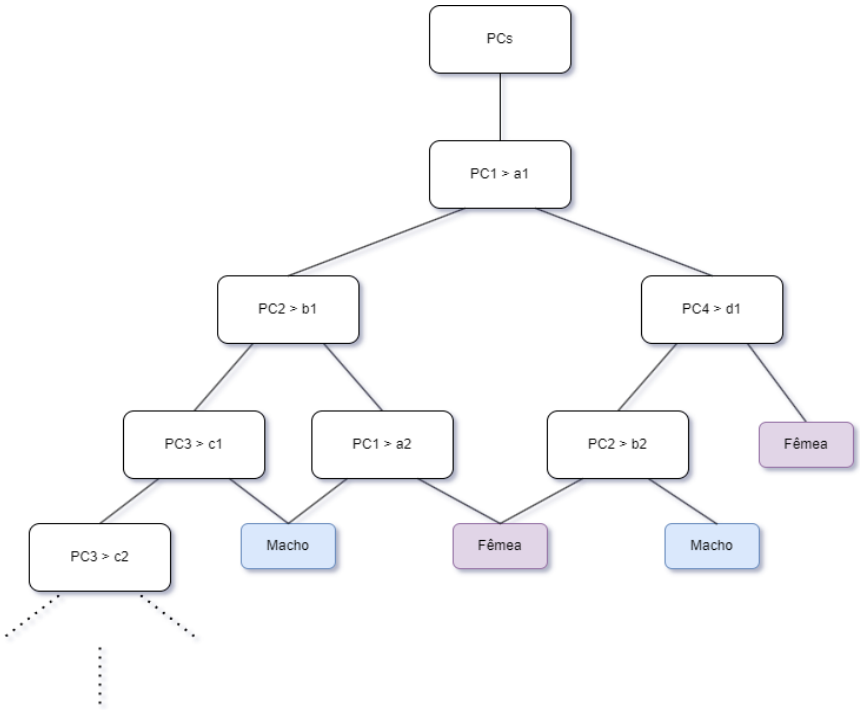
Cada componente principal é uma combinação linear das variáveis originais. Os componentes são ordenados por quantidade de variância original que eles conseguem explicar. Portanto, o PCA permite manter apenas os primeiros componentes principais que explicam a maior parte da variância, reduzindo efetivamente a dimensionalidade dos dados. Isso pode melhorar a eficiência computacional e, possivelmente, a performance do modelo, minimizando o risco de *over-fitting* [20].

O *score plot* é uma ferramenta visual importante no contexto do PCA. Ele é uma representação gráfica que permite visualizar a variabilidade dos dados. Cada ponto no gráfico representa uma observação do conjunto de dados, posicionada de acordo com seus valores nos componentes principais.

O algoritmo *Random Forest* é uma ferramenta poderosa para a classificação do sexo de aves com base em dados de espectroscopia FTIR, especialmente após a redução da dimensionalidade dos dados usando PCA. Ele é capaz de lidar com a complexidade e a alta dimensão dos dados, oferecendo ao mesmo tempo uma medida de importância das variáveis que pode ser útil para a interpretação dos resultados.

Se trata de um método de aprendizado de máquina que constrói várias árvores de decisão durante o treinamento. Cada árvore é construída a partir de uma amostra dos dados de treinamento. Além disso, em cada nó da árvore, um componente principal aleatório é selecionado para a divisão. Isso introduz aleatoriedade no processo de construção da árvore e ajuda a evitar o over-fitting [22]. Um exemplo de árvore é apresentado na Figura 1.

Figura 1: *Diagrama representativo de arvore de decisão para a sexagem.*



Fonte: *Autor*

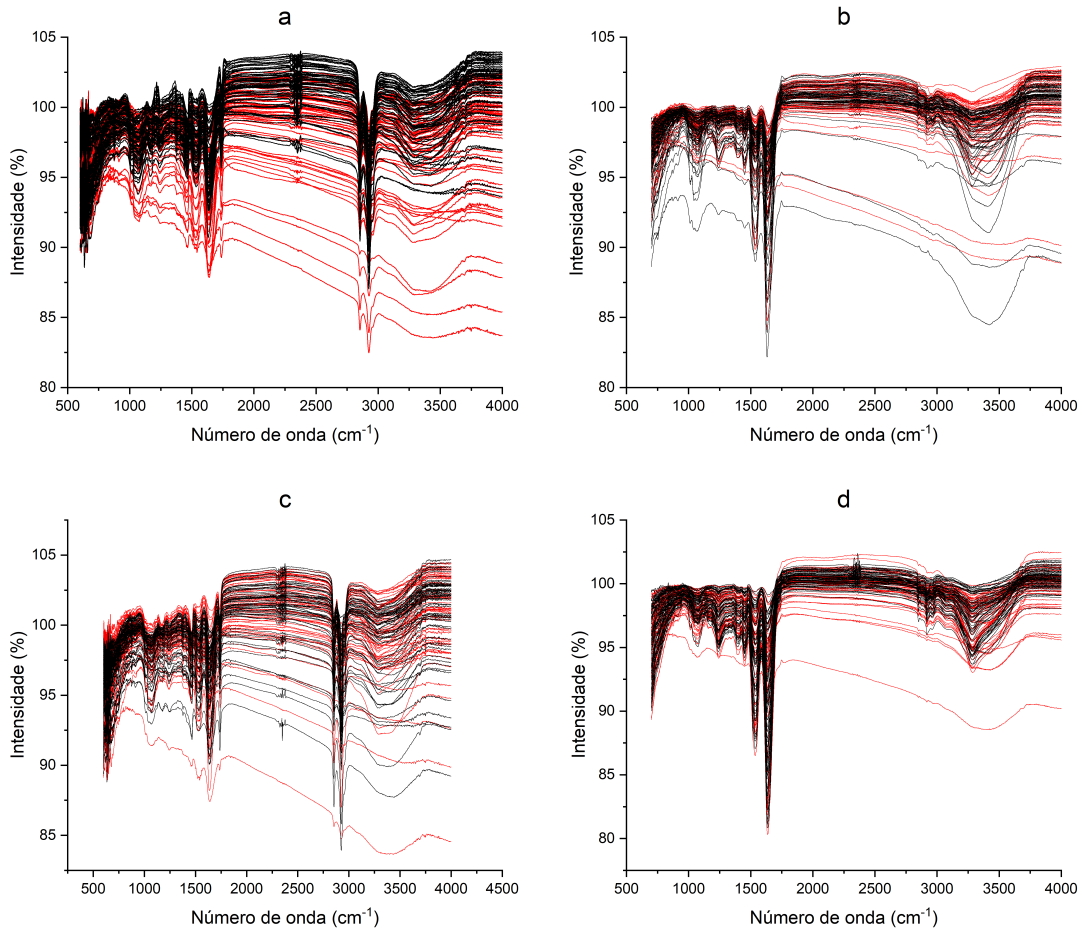
Quando uma nova instância de dados é apresentada ao modelo, cada árvore na floresta dá uma previsão independente. No caso da classificação, a classe com a maioria dos votos é a previsão do modelo [22].

4. RESULTADOS E DISCUSSÕES

Após a obtenção dos espectros FTIR, anteriormente medidos pelo

As medidas espectroscópicas retornaram os dados usados para construir os gráficos da Figura 2 do espectro sem nenhum pré-tratamento nos dados.

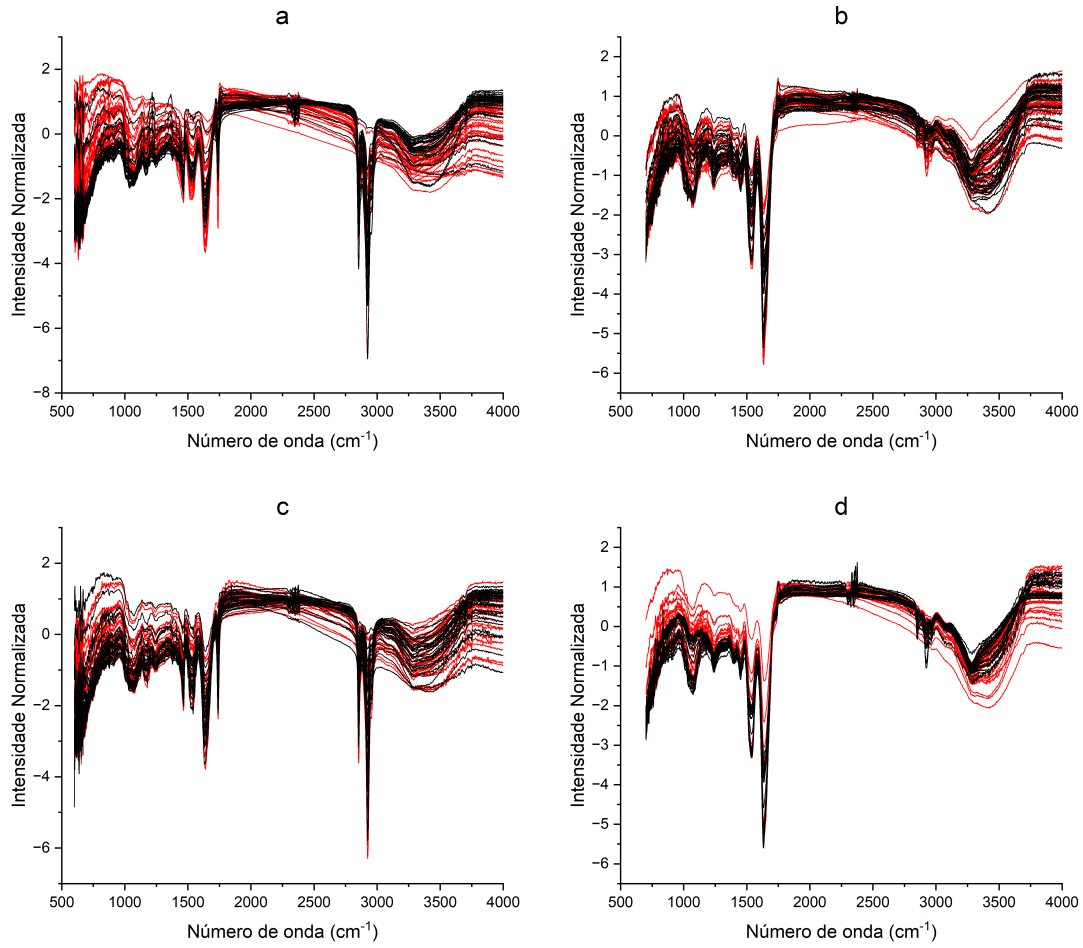
Figura 2: *Espectro sem tratamento das especies (a) Bicudo, (b) Calopsita, (c) Curió e (d) Ring Neck. Fêmeas em vermelho e machos em preto.*



Fonte: *Autor.*

A partir da normalização dos dados, obteve-se os gráficos da Figura 3.

Figura3: *Espectro normalizado das especies (a) Bicudo, (b) Calopsita, (c) Curió e (d) Ring Neck. Fêmeas em vermelho e machos em preto.*

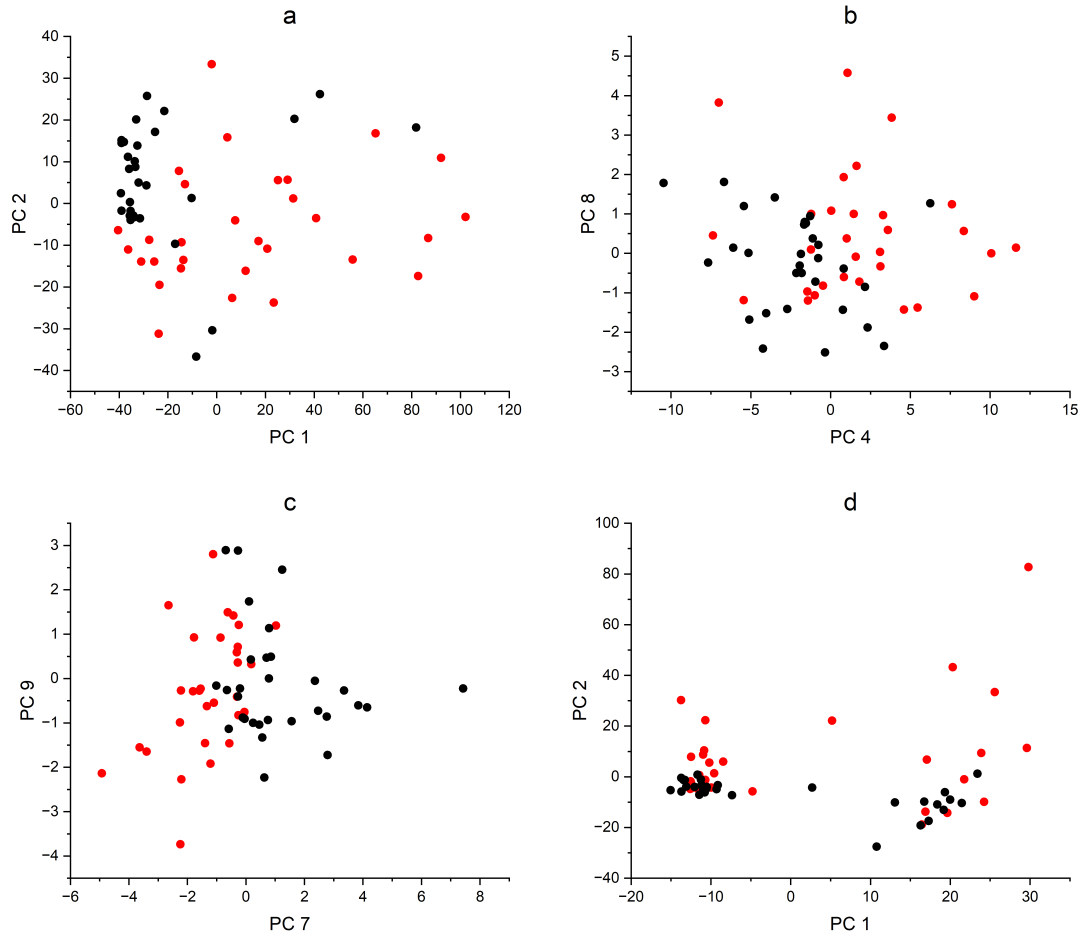


Fonte: *Autor.*

É importante destacar, nesse momento, a importância da normalização para os dados, visto que para todas as espécies, os espectros ficaram mais concentrados em torno de um valor. Isso é útil pois representa com mais fidelidade as diferenças de características individuais, ou seja, faz com que fique mais fácil de perceber a real distinção entre os indivíduos.

Esses dados foram utilizados no processo de análise de componentes principais (PCA), foram selecionados os componentes, que melhor separe os conjuntos, utilizadas no *random forest* e construiu-se os *score plots* da Figura 4.

Figura 4: *score plot que melhor separou cada especie (a) Bicudo, (b) Calopsita, (c) Curió e (d) Ring Neck. Fêmeas em vermelho e machos em preto.*



Fonte: *Autor.*

Vale destacar, que mesmo sendo os componentes que melhor separam cada os conjuntos de cada espécie, todos ainda estão muito agrupados para uma seleção simples, ou seja, olhando apenas para um componente principal.

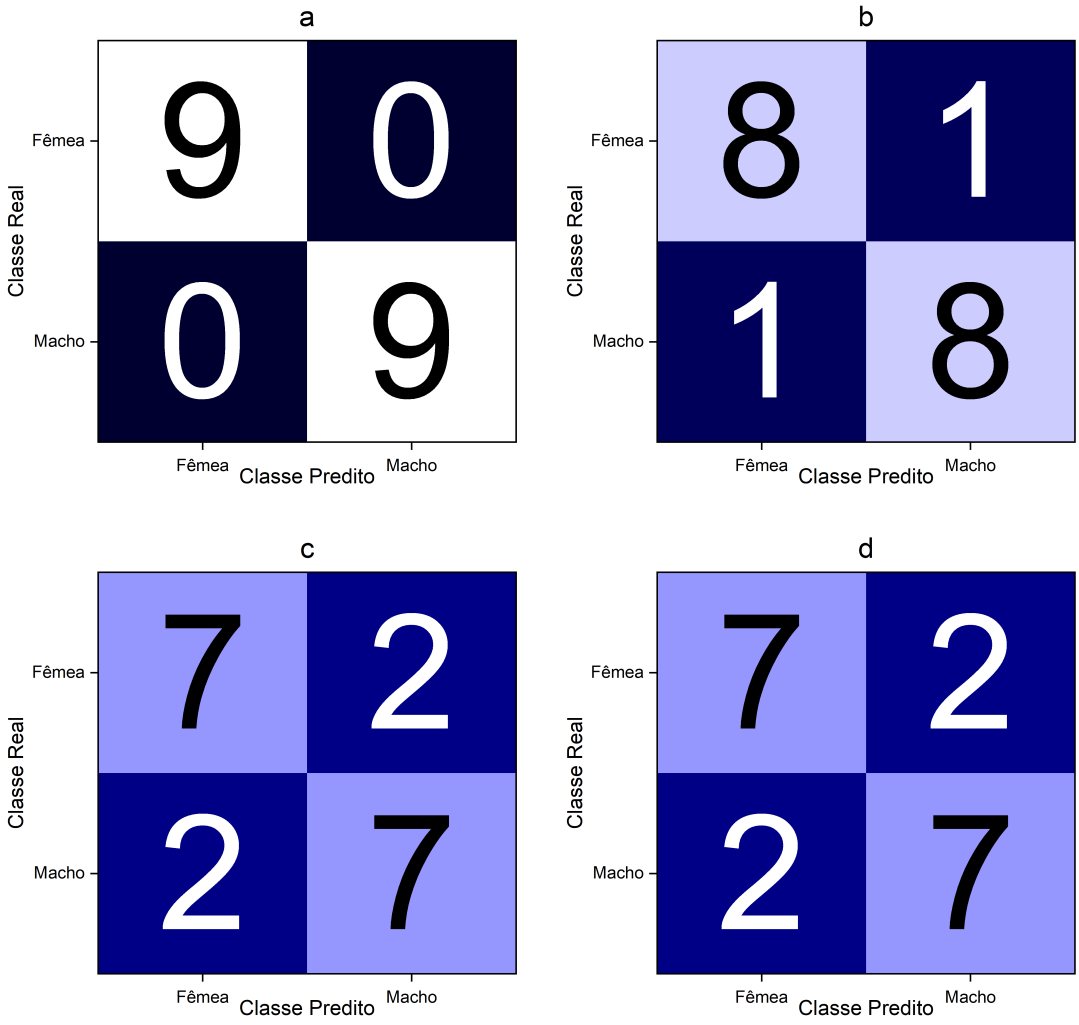
Com isso em mente, o modelo de *random forest* foi treinado, utilizando de 137 árvores para cada classe. Esse valor foi obtido em otimização para menor número de árvores com o melhor resultado.

Treinou-se inicialmente com 100, 200, 250 e 300 árvores para cada espécie e com base no resultado, o número de árvores foi reduzido até que atingisse a mesma acurácia apresentada pelos modelos de 200, 250 e 300.

O modelo atingiu, em validação interna, as acurácias de 95 % para o bicudo, 78,57 % para calopsitas, 95,23 % nos curiós e 76,19 % para os *ring neck.*

O treinamento levou 21 das 30 amostras de cada classe. 9 foram utilizadas para a validação externa. Essas amostras foram tratadas de maneira isolada, a fim de não influenciar o processo de treinamento. O resultado da validação é apresentado na Figura 5.

Figura 5: *Matriz de confusão das especies (a) Bicudo, (b) Calopsita, (c) Curió e (d) Ring Neck. Fêmeas em vermelho e machos em preto.*



Fonte: *Autor.*

Com base na validação externa calculou-se a acurácia do algoritmo em 100% para o Bicudo, 88,89% para o Calopsita e 77,78% para Curió e *Ring Neck.*

4. CONCLUSÃO

A validação externa apresenta uma acurácia de 100% para o Bicudo, 88,89% para o Calopsita e 77,78% para Curió e *Ring Neck*. Com base nos resultados alcançados o método se mostra eficiente em seleção de sexo de algumas espécies testadas. Contudo, deve ser feito um estudo mais aprofundado.

Um dos caminhos possíveis é a análise das árvores. Podendo analisar, por exemplo, a porcentagem de árvores que classificaram a espécie, estipular um número mínimo de árvores, e, caso a classificação da amostra não respeite esse valor, classificá-la como indefinido.

Essa abordagem pode, também, trazer problemas, pois, por um lado a acurácia do algorítimo pode aumentar, visto que algumas amostras classificadas erroneamente serão classificadas como indefinido. Porem, adiciona mais um grau de liberdade no treinamento do algorítimo, trazendo mais uma otimização a ser realizada.

Uma outra abordagem é encontrar amostras “difíceis” de serem classificadas, e criar uma classe de indefinidos e treinar o algorítimo com as 3 classes. Essa abordagem tem a vantagem de ter a etapa de treinamento e de validações idênticas. Porem apresenta um problema de número de amostras, precisaria de um número relativamente grande de amostras “difíceis”.

Por fim, o protocolo adotado se mostra promissor para a classificação de algumas espécies de aves, mas requer aperfeiçoamento para trabalhar com outras espécies.

5. REFERÊNCIAS BIBLIOGRÁFICAS

[1]. F. B. Knackfuss Et Al., Sexagem de Aves da Espécie Amazona aestiva (Papagaio Verdadeiro) por meio de técnica de PCR.14(6) Pubvet, 2020.
[2]. R. Griffiths, M. C. Double, K. Orr E R. J. G. Dawson, A DNA test to sex most birds.-() Mol. Eco., 1998.
[3]. B. P. Gonçalves, Sexagem molecular em aves: contribuições à conservação biológica e á divulgação científica.-() Universidade Estadual Paulista Júlio de Mesquita Filho, Instituto de Biociências de Botucatu, 2013.
[4]. J. N. Vieira, SEXAGEM MOLECULAR EM AVES VIA PCR –AVALIAÇÃO DE TRÊS TÉCNICAS DE EXTRAÇÃO DE DNA .-() UFMG, 2009.
[5]. J. N. Vieira, E. G. A. Coelho, D. A. A. Oliveira, Sexagem molecular em aves silvestres.33(2) Rev. Bras. Reprod. Anim., 2009.
6: Artymiak, Jacek, LibreOffice Calc Functions and Formulas Tips, 2011
[7]. T. F. Raso, K. Werther, Sexagem cirúrgica em aves silvestres.56(2) Arq. Bras. Med. Vet. Zootec., 2004.
[8]. G. Steiner, G. Preuse, C. Zimmerer, M. E. Krautwald-junghanns, V. Sablinskas, H. Fuhrmann, E. Koch, T. Bartels, Label free molecular sexing of monomorphic birds using infrared spectroscopic imaging.-(150) Elsevier, 2016.
[9]. A. Angulo, L. Yang, E. S. Aydil, A. Modestino, Machine learning enhanced spectroscopic analysis: towards autonomous chemical mixture characterization for rapid process optimization.1() Digital Discovery, 2021.
[10]. C. A. M. Ramirez, M. Greenop, L. Ashton, I. Ur Rehman, Applications of machine learning in spectroscopy.8-10(56) Applied Espectroscopy Reviews, 2021.
[11]. I. Magnus, M. Virte, H. Thienpont, L. Smeesters, Combining optical spectroscopy and machine learning to improve food classification.-(130) Elsevier, 2021.
[12]. Y. Zhang, Q. Tang, Y. Zhang, J. Wang, U. Stimming, A. A. Lee, Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning.11(1706) Nature Communications, 2020.
[13]. G. Larios, G. Nicolodelli, M. Ribeiro, T. Canassa, A. R. Reis, S. L. Oliveira, C. Z. Alves, B. S. Marangoni, C. Cena, Soybean seed vigor discrimination by using infrared spectroscopy and machine learning algorithms.35(-) Analytical Methods, 2020.
[14]. G. Larios, M. Ribeiro, C. Arruda, S. L. Oliveira, T. Canassa, M. J. Baker, B. Marangoni, C. Ramos, C. Cena, A new strategy for canine visceral leishmaniasis diagnosis based on FTIR spectroscopy and machine learning.11(14) Jornal of Biophotonics, 2021.
[15]. P. R. Claro, Espectroscopia.5(4) Rev. Ciência Elem., 2017.
[16]. M. S. Dhanoa, S. J. Lister, The Link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) Transformations of NIR Spectra.1(2) Journal of Near Infrared Spectroscopy, .
[17]. J. Huang, S. Romero-Torres, M. Moshgbar, Practical Considerations in Data Pre-treatment for NIR and Raman Spectroscopy.-() American Pharmaceutical Review, 2010.
[18]. H. P. Kriegel, M. Schubert, A. Zimek, Angle-based outlier detection in high-dimensional data.-() KDD 'o8, 2008.
[19]. A. Ur Rahman, S. B. Belhaouari, Unsupervised outlier detection in multidimensional data.8(80) Jornal of Big Data, 2021.
[20]. B. Zhao, X. Dong, Y. Guo, X. Jia, Y. Huang, PCA Dimensionality Reduction Method for Image Classification.-(54) Springer, 2022.

[21]. T. S. Arulananth, L. Balaji, M. Baskar, V. Anbarasu, K. S. Rao, PCA Based Dimensional Data Reduction and Segmentation for DICOM Images.-(55) Springer, 2023.
[22]. J. L. S. Ramirez, R. J. Cruz, Y. V. Rey C. Y. Marquez, Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects.16(6) algorithms, 2023.