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

Curso de Física Bacharelado

Trabalho de Conclusão de curso

Aprendizado de Máquina Aplicado à Espectroscopia: Uma Abordagem Não Invasiva para Sexagem de Aves

ELISA PINHEIRO DE ALMEIDA

CAMPO GRANDE, MS

2025

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

Aprendizado de Máquina Aplicado à Espectroscopia: Uma Abordagem Não Invasiva para Sexagem de Aves

Elisa Pinheiro de Almeida

Orientador: Bruno Spolon Marangoni

Trabalho de Conclusão de Curso apresentado como parte das atividades para obtenção do título de Física, do curso de Física Bacharelado da Fundação UFMS.

Campo Grande, MS

2025

É com todo meu amor que dedico este trabalho a minha mãe, Miracy.

AGRADECIMENTOS

Agradeço inicialmente aos meus pais, Miracy e Antônio, que acreditaram em mim, em minha jornada e moveram mundos para que eu tivesse condições de trilhar meu caminho em Campo Grande. Principalmente, a minha mãe que ao meu lado nos bons e maus momentos não largou minha mão, que me acalmou quando tudo o que eu mais queria era larga tudo e voltar para minha cidade. A minha avó, que mesmo não estando mais presente, aparecia em sonhos torcendo por mim. Aos meus tios, que tiveram papeis fundamentais na minha criação.

Agradeço também aos amores da minha vida, meus gatos, que apareceram no meu caminho para acalentar meus dias bons e ruins. Principalmente, ao meu gatinho Alcione, que infelizmente partiu cedo demais, mas que deixou lembranças boas e felizes ao seu lado. Estará sempre em minhas memórias.

As minhas amigas de Belém, que mesmo a distância continuamos apoiando uma à outra em nossas jornadas. Sempre estarão no meu coração, meninas. Aos melhores amigos que a faculdade me presenteou: Marina Peregrinelli, Mariana Brito, Maria Victória, Gustavo Sgurscow, Mayara Fernandes, Vinicius Thiago, Sofia Barbosa, Stéfani Torres e Denis Rodrigues, por dividirem os melhores momentos da faculdade, aonde sempre apoiamos uns aos outros. Obrigada por terem permanecido.

Por fim, agradeço ao meu orientador, Bruno S. Marangoni, por me auxiliar durante o processo de escrita, auxiliar com dúvidas.

RESUMO

Identificar o sexo da ave é de máxima importância para a comercialização, veterinários, criadores. Metodologias baseadas em DNA (ácido desoxirribonucleico) foram desenvolvidas. Contudo, por mais que forneça uma identificação altamente sensível, é um método que demanda tempo e muitas vezes laboratórios específicos. Nesse sentido, alguns trabalhos apresentam técnicas de espectroscopia associada a aprendizagem de máquina para uma análise mais precisa. Este trabalho apresenta análises de medidas espectroscópicas por Espectroscopia no Infravermelho por Transformada de Fourier (FTIR) associada a algoritmos de aprendizagem de máquina como técnica não invasiva para determinação do sexo de ave. Os espectros analisados foram da espécie *Oryzoborus maximiliani* (Bicudo) de 4000 - 600 cm^{-1}. A metodologia envolveu a coleta de espectros de penas de pássaros, tratamento de dados pelo SNV (*Standard Normal Variate*), submetidos à Análise de Componentes principais (PCA) para reduzir e identificar variáveis significativas para passarem pelo algoritmo de aprendizado não supervisionado, sendo o *K-means* utilizado para selecionar as melhores componentes. A seleção de PC combinado ao método *leave-one-out cross-validation* (LOOCV) para os classificadores LDA (*Linear Discriminant Analysis*), SVM (*Support Vector Machine*) e KNN (*K-Nearest Neighbors*) mostrou resultados significativos e promissores, com destaque para o SVM, obtendo acurácia de 99,17%. Esse resultado demonstrou a eficiência da metodologia aplicada.

Palavra-chave: FTIR, PCA, K-means, SVM, LDA, KNN.

ABSTRACT

Identifying the sex of birds is of utmost importance for marketing, veterinarians, and breeders. DNA (deoxyribonucleic acid) based methodologies have been developed. However, although it provides highly sensitive identification, it is a time-consuming method that often requires specific laboratories. In this regard, some studies present spectroscopy techniques associated with machine learning for more accurate analysis. This study presents analyses of spectroscopic measurements by Fourier Transform Infrared Spectroscopy (FTIR) associated with machine learning algorithms as a non-invasive technique for determining bird sex. The spectra analyzed were from the species Oryzoborus maximiliani (Bicudo) from 4000 to 600 cm^(-1). The methodology involved collecting spectra from bird feathers, processing the data using SNV (Standard Normal Variate), and submitting them to Principal Component Analysis (PCA) to reduce and identify significant variables to be passed through the unsupervised learning algorithm, with K-means being used to select the best components. The selection of PC combined with the leave-one-out cross-validation (LOOCV) method for the LDA (Linear Discriminant Analysis), SVM (Support Vector Machine), and KNN (K-Nearest Neighbors) classifiers showed significant and promising results, with SVM standing out, achieving an accuracy of 99.17%. This result demonstrated the efficiency of the applied methodology.

Keyword: FTIR, PCA, K-means, SVM, LDA, KNN.

LISTA DE FIGURAS E TABELAS

SUMÁRIO

1. INTRODUÇÃO

Em análises espectroscópicas, a identificação de medidas fora do padrão é de extrema importância para a condução adequada das análises. Nesse contexto, a o maior desafio é lidar com padrões que destoam do resto, esses dados possuem alta dimensionalidade e neles facilmente são encontrados ruídos.

A Espectroscopia no Infravermelho por Transformada de Fourier (FTIR), é uma técnica que mede a intensidade do feixe no infravermelho com base no comprimento de onda de forma menos prejudicial para as amostras. Ele é rápido e eficiente para localização de grupos funcionais [1,2]. Nesse sentido, existem diversos estudos em que essa técnica tem sido utilizada para análises químicas de materiais biológicos [3], onde essa técnica associada ao Analise de Componente Principal (PCA), SVM (*Support Vector Machine*), KNN (*K-Nearest Neighbors*), LDA (*Linear Discriminant Analysis*), *K-means* se mostram bastante eficaz [4,5].

Identificar o sexo da ave é de máxima importância para a comercialização, veterinários, criadores [6]. Metodologias baseadas em DNA (ácido desoxirribonucleico) foram desenvolvidas em marcadores específicos considerando as variações entre os sexos, para aves, é comum extrair DNA do sangue e das penas [7,68]. Contudo, por mais que forneça uma identificação altamente sensível, é um método que demanda tempo e muitas vezes laboratórios específicos [8]. Além disso, há outras técnicas para identificação de sexos de aves como medições morfometrias [9], acústica [10], laparoscopia [11], comportamental [12], laparotomia [13], entre outras.

Estudos apontam que o FTIR é uma técnica já utilizada como método moderno, rápido e eficiente para sexagem de aves [14,15,16]. Essa técnica foi utilizada para analisar penas de contorno em filhotes de peru, onde foi observado variações espectrais na faixa de 1000 a 1250 cm^{-1}. A classificação foi realizada pelo método não supervisionado PCA, que com as informações obtidas permitiu classificar os filhotes em macho e fêmea superior a 95% [15]. Além disso, foi visto que espectros de FTIR em conjunto com PCA e ao classificador LDA obteve 100% de acurácia utilizando apenas uma quantidade pequena da polpa de penas, o que torna viável para a sexagem de pombos [16].

Desse modo, ao estudar a literatura, nota-se que a técnica de FTIR já é utilizada para a sexagem de aves [15,16]. Esse trabalho destaca-se pela criação de rotinas especificas para a abordagem de dados espectroscópicos de sexagem de ave da espécie *Oryzoborus maximiliani* (Bicudo). Assim, torna-se viável testar sua aplicabilidade utilizando aprendizado de máquina não supervisionado (PCA e *K-means*) e supervisionado (LDA, SVM e KNN).

2. REVISÃO DA LITERATURA

2.1 Sexagem de ave

Sexagem é o ato de determinar se um indivíduo da mesma espécie é macho ou fêmea [17]. Existem diversas metodologias que podem ser vistas para a sexagem de um indivíduo de mesma espécie, por exemplo, técnicas baseadas na morfologia animal, como laparoscopia [11], entre outras. Além disso, a morfologia pode ser aprofundada por exames como tomografia por ressonância magnética, ultrassonografia [18,19].

Estudos recentes mostram um crescente avanço em técnicas para sexagem de aves, tanto para interesse da indústria, como para aves de corte (sexagem *in ovo*). Isso permite a evolução para manejo, conservação, pesquisa [20, 21]. Para aves sem dimorfismo sexual aparente e criadas em cativeiro, protocolos seguros e eficazes para a sexagem são fundamentais para reprodução e comercio legal de aves [22]. A fim de atender essas demandas, diferentes técnicas foram desenvolvidas com o passar dos anos beneficiando pesquisas e conservações.

Entretanto, essas metodologias apresentam algumas desvantagens, como o risco de dano físico ao animal, elevado custo de equipamentos, manutenção, equipe técnica especializada, além da necessidade de reproduzir as técnicas em larga escala [23,24]. Atualmente, as técnicas se baseiam na extração de material genético e avaliação via PCR (*Polymerase Chain Reaction,* ou Reação em Cadeia da Polimerase) [25,16]. Esse método destaca-se pela sua eficácia, rapidez e precisão, além de não apresentar riscos biológicos para o indivíduo e, também é uma técnica minimamente invasiva [25].

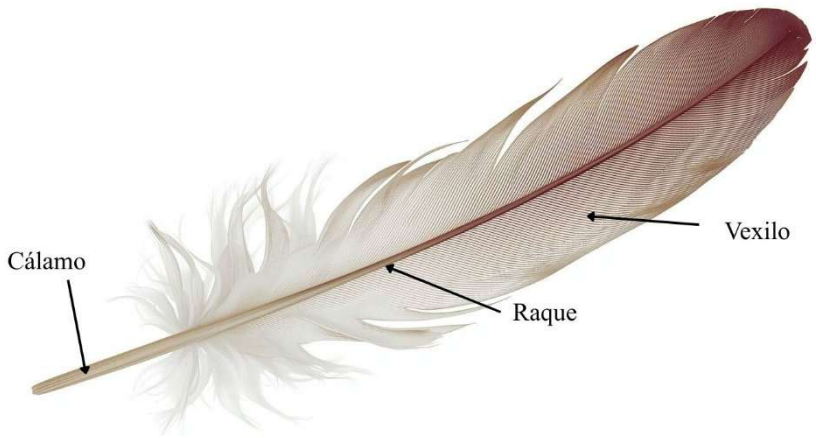
Recentemente, técnicas não invasivas baseadas em espectroscopia e aprendizagem de maquina (*Machine Learning*) foram surgindo para sexagem de aves através de suas penas. Steiner

et al. [8] aplicaram a espectroscopia no infravermelho por transformada de Fourier (FTIR – *Fourier Transform Infrared Spectroscopy*) para realizar a sexagem de pombos domésticos por meio da análise da polpa de suas penas. As imagens espectroscópicas obtidas foram processadas com análise de componentes principais (PCA) e classificadas por meio da análise discriminante linear (LDA – *Linear Discriminant Analysis*), resultando em 100% de acerto na identificação das amostras provenientes de machos.

Harz et al. [26] utilizaram a espectroscopia Raman com ressonância na faixa do UV como técnica minimamente invasiva para a determinação do sexo em aves, analisando amostras de polpa extraídas das penas. A classificação foi realizada por meio de máquinas de vetor de suporte (SVM), alcançando 95% de acurácia. Além disso, aplicou-se a análise de componentes principais (PCA – *Principal Component Analysis*) para identificar as regiões espectrais mais relevantes na discriminação dos espectros Raman obtidos da polpa de penas de frangos.

As penas de aves têm uma estrutura fundamental (Figura 1) constituída pelo cálamo, raque e vexilo. O cálamo é a parte tubular e vazia que se conecta à pele, fixando a pena ao corpo da ave. Do cálamo, se estende o raque, que é um eixo central rígido que dá suporte a toda a pena. Nos lados do raque, localiza-se o vexilo, que é formado por barbas e bárbulas entrelaçadas, criando uma superfície contínua, leve e durável, que é crucial para o voo e o isolamento térmico. Esta configuração estrutural assegura uma boa aerodinâmica e funcionalidade para as aves [71].

Figura 1- Estrutura de uma pena de ave.



Fonte: Autora.

Para estudos com FTIR, a parte mais frequentemente analisada de uma pena é o vexilo, em especial a seção central das barbas. Essa escolha se deve ao fato de que essa área tem uma superfície mais uniforme e menos interferência de estruturas mais duras como o cálamo, resultando em espectros mais consistentes e reproduzíveis. Ademais, o vexilo costuma acumular um menor número de impurezas estruturais, facilitando a identificação mais clara dos grupos funcionais da queratina que estão presentes na pena [72]. Neste trabalho, a região analisada por espectroscopia FTIR, foi a região do vexilo (Figura 1).

2.2 **Análise espectroscópica: Espectroscopia no Infravermelho por Transformada de Fourier (FTIR)**

Espectroscopia no infravermelho (IR), é uma técnica vibracional usada para avaliar a composição química de uma amostra [27]. Esse é um campo cientifico que norteia como átomos e moléculas comportam-se quando entra em contato com a radiação eletromagnética em diferentes comprimentos de onda, observando o quanto absorvem, emitem ou dispersam de energia [28,29]. Baseia-se na absorção da luz pelas moléculas que compõe o material [30].

O espectro eletromagnético é dividido conforme suas diversas áreas de comprimento de onda. O infravermelho inclui comprimentos de onda que variam de 0,78 a 1000 µm, sendo dividido em três categorias principais: o infravermelho próximo (NIR, Infravermelho Próximo), que se estende de 0,78 a 2,5 µm (12800 - 4000 cm^{-1}); o infravermelho médio (MIR, Infravermelho Médio), que vai de 2,5 a 50 µm (4000 - 200 cm^{-1}); e o infravermelho distante (FIR, Infravermelho Distante), abrangendo o intervalo de 50 a 1000 µm (200 - 10 cm^{-1}) [69].

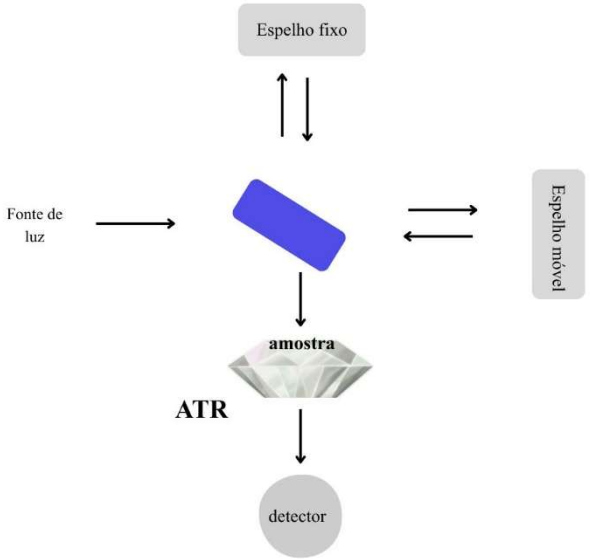
No espectro infravermelho, a captação ou liberação de radiação pelas moléculas acontece por conta das mudanças cíclicas no dipolo elétrico, que refletem a distinção entre as cargas positivas e negativas na eletrosfera. A radiação que é captada ou liberada tem uma frequência que corresponde àquela da oscilação do dipolo [73]. O modo como as moléculas oscilam é influenciado pelas formas que seus átomos têm para se mover, o que determina seus graus de liberdade. Em uma molécula com N átomos, existem 3N tipos distintos de oscilações, que incluem translações, rotações e vibrações. As translações referem-se a movimentações do conjunto da

molécula, as rotações indicam movimentos em torno de eixos determinados, e as vibrações referem-se às oscilações dos átomos em relação às suas ligações químicas quando estão em posição de equilíbrio. Moléculas não lineares apresentam 3N - 6 graus de liberdade vibracional, enquanto as lineares possuem 3N - 5, uma vez que não há rotação ao longo do eixo da molécula [74].

Os tipos de vibrações moleculares são referidos como modos vibracionais e podem ser divididos em duas categorias: estiramento e deformação angular. O estiramento acontece quando os átomos se movem em direção ao eixo da ligação química, seja se aproximando ou afastando. A deformação angular envolve mudanças nos ângulos das ligações químicas, podendo incluir movimentos de curvatura. Os modos vibracionais podem ser classificados como simétricos, quando os átomos se movimentam de maneira semelhante, ou assimétricos, quando os movimentos ocorrem de forma oposta [74].

Os espectrômetros FTIR atuais são baseados no interferômetro de Michelson (Figura 2), que é formado por uma fonte de radiação, um divisor de feixe e dois espelhos, um que permanece fixo e o outro que se move. A radiação que chega é dividida, refletida e então reunida, criando um interferograma que, quando tratado pela transformada de Fourier, produz o espectro da amostra [75].

Figura 2 - Interferometro de Michelson.



Fonte: Autora.

Na espectroscopia FTIR, é possível realizar diversas varreduras (ou *scans*), em que os interferogramas são somados em um processo denominado co-adição. Essa estratégia contribui para o aumento da qualidade do espectro final (31,32). Outro parâmetro fundamental é a resolução espectral, responsável por definir a capacidade do equipamento em separar bandas próximas. Embora resoluções mais altas permitam espectros mais detalhados, também requerem um deslocamento maior do espelho móvel, prolongando o tempo de aquisição e, consequentemente, podendo intensificar a captação de ruído (33,34).

2.3 **Análise de Componentes Principais (PCA)**

Análise de Componentes Principais (PCA, *Principal Component Analysis*) trata-se de uma técnica estatística da análise multivariada. Ela transforma um conjunto original de dados, em um novo conjunto de dados com menor dimensão sem perder as informações originais, denominadas de componentes principais (PC – *Principal Component*) [35].

Do ponto de vista matemático, os dados são estruturados em uma matriz **X**, formada por *n* amostras (linhas) e *m* variáveis (colunas). No PCA, essa matriz é decomposta em duas outras: a matriz **T**, que reúne as projeções das amostras no espaço dos componentes principais (*scores*), e a matriz **P**, que expressa os pesos ou contribuições de cada variável original para a variância explicada por cada componente principal (*loadings*) [36]. Essa decomposição da matriz **X**, gera as PC, que são novas variáveis, realiza-se gráficos de *scores* e *loadings* para saber o quanto cada variável nova contribui [36].

No entanto, em conjuntos de dados obtidos por espectroscopia molecular, é fundamental realizar uma etapa de uniformização antes da aplicação do PCA. Essa etapa visa reduzir possíveis erros durante a aquisição espectral, os quais podem comprometer a qualidade da análise multivariada. Para isso, são empregadas técnicas de pré-processamento que corrigem deslocamentos de linha de base e normalizam as intensidades, favorecendo a comparabilidade

entre os espectros. Entre as estratégias mais utilizadas destacam-se a normalização por comprimento de vetor, a correção de espalhamento múltiplo (MSC), a normalização por variância padrão (SNV) permitem minimizar ruídos e realçar características espectrais relevantes [37,38]. O SNV é definido pela Equação 1:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

Onde x representa o valor da intensidade da transmitância, μ é a média das intensidades, σ é o desvio padrão e z é denota o espectro transformado pelo SNV [39].

2.4 **Aprendizagem de Máquina**

O aprendizado de máquina (*machine learning, ML*) consiste em um conjunto de métodos matemáticos e estatísticos que permitem aos sistemas computacionais identificar padrões e fazer previsões a partir de dados. Esses métodos geram modelos, que podem ser entendidos como funções matemáticas capazes de tomar decisões baseadas em informações previamente analisadas [40]. Os métodos tradicionais de aprendizado de máquina empregam algoritmos de seleção de características para identificar as variáveis mais relevantes e diversos algoritmos de ML, tanto supervisionados quanto não supervisionados, têm sido desenvolvidos para aplicações clínicas. [41,42].

O ML pode ser classificado em métodos supervisionados e não supervisionados. Os métodos supervisionados são aqueles em que temos um conjunto de dados com informações já identificadas ou rotuladas para cada indivíduo ou amostra. Isso ajuda o modelo a aprender como as variáveis de entrada se relacionam com o resultado esperado, permitindo que ele faça previsões. Esses métodos se dividem em duas categorias principais: modelos de classificação, que colocam os dados em categorias específicas, e os modelos de regressão, que fazem previsões de valores numéricos contínuos, com base nos padrões que o modelo aprende durante o treinamento. [43].

Já os métodos não supervisionados são aqueles em que o algoritmo tenta encontrar padrões, estruturas ou grupos nos dados sem precisar de rótulos categorias já definidas. Ao contrário do aprendizado supervisionado, onde o modelo é treinado com exemplos que já têm uma classificação. No aprendizado não supervisionado o algoritmo apenas analisa os dados para identificar

regularidades e relações, sem separar previamente os tipos de informação. Um exemplo comum desse tipo de método é o algoritmo *K-Means*, que é usado para agrupar dados em diferentes clusters [43].

O processo de aprendizado não supervisionado segue etapas fundamentais. Inicialmente, realiza-se a coleta e o pré-processamento dos dados, que abrangem etapas como limpeza, normalização e seleção das variáveis mais relevantes. Em seguida, define-se o algoritmo mais apropriado conforme o objetivo do estudo, seja agrupar amostras semelhantes ou reduzir a dimensionalidade dos dados. O modelo é então aplicado, buscando identificar automaticamente padrões ou estruturas ocultas. Posteriormente, os resultados são analisados e interpretados, o que representa um desafio, já que, na ausência de rótulos, não existe uma referência para comparação direta. Assim, a avaliação costuma basear-se em métricas internas, como a coesão e separação dos grupos formados, além de representações visuais em duas ou três dimensões que auxiliam na interpretação dos achados [44]. Neste trabalho foi utilizado o PCA para redução de dimensionalidade, o *K-means* para identificar as melhores PC e com os classificadores SVM (*Support Vector Machine*), KNN (*K-Nearest Neighbors*), LDA (*Linear Discriminant Analysis*) na identificação dos sexos de aves (macho ou fêmea), respectivamente. O teste foi realizado comparativamente com e sem a presença do K-means para seleção de variáveis.

2.5 **K-means Clustering**

O *K-means* é um dos algoritmos mais populares em aprendizado de máquina não supervisionado, especialmente em processos de agrupamento de dados. Sua finalidade principal é dividir um conjunto de dados em *k* grupos (ou *clusters*), assegurando que os elementos dentro de cada grupo sejam mais semelhantes entre si do que em relação aos elementos de outros grupos. O algoritmo começa definindo *k* centróides iniciais, que podem ser selecionados aleatoriamente ou por meio de técnicas específicas. Em seguida, cada ponto de dado é alocado ao centróide mais próximo, formando clusters temporários. Posteriormente, os centróides são recalculados como a média dos pontos atribuídos a cada grupo, e esse procedimento é repetido de forma iterativa até que os centróides se tornem estáveis, ou seja, até que não haja alterações significativas nas atribuições dos clusters [45,46].

Neste algoritmo, inicialmente é representando o grupo de pontos centrais. Depois, com base na menor distância, os demais pontos são juntados aos centros correspondentes. Se a classificação estiver errada, ela será ajustada, recalculando os pontos centrais de cada grupo, e esse processo será

repetido até que a classificação fique correta. O *K-Means* é conhecido por ser um método rápido e eficiente [47].

Uma das principais vantagens do *K-means* é sua simplicidade e eficiência computacional, o que permite a aplicação em grandes volumes de dados. No entanto, o algoritmo apresenta limitações, como a necessidade de pré-definir o número de clusters *k*, a sensibilidade a valores iniciais dos centróides e a dificuldade em lidar com clusters de formas complexas ou com tamanhos muito diferentes [48,49].

Diante disso, este trabalho visa utilizar o *k-means* para classificar e selecionar as melhores PCs para serem utilizadas pelos classificadores supervisionados.

2.6 Classificadores Supervisionados: SVM (Support Vector Machine), KNN (K-Nearest Neighbors), LDA (Linear Discriminant Analysis)

Os classificadores SVM (*Support Vector Machine*), KNN (*K-Nearest Neighbors*), LDA (*Linear Discriminant Analysis*) são comumente utilizados em aprendizado de máquina supervisionado, especialmente em funções de categorização. Cada um deles aplica princípios diversos para dividir os dados em diferentes categorias, apresentando benefícios e desvantagens que os tornam mais apropriados para distintas situações e características dos dados [50,51].

O SVM é um método de classificação que visa identificar um hiperplano ideal que consiga dividir as categorias de um conjunto de dados, ampliando a distância entre os pontos mais próximos de cada categoria, conhecidos como vetores de suporte. Essa ampliação da distância favorece a capacidade de generalização do modelo. Nos casos em que as categorias não podem ser separadas de forma linear, o SVM recorre a funções de kernel para transferir os dados a um espaço de maior dimensão, onde a separação linear se torna viável. Entre os kernels mais comumente utilizados estão o linear, o polinomial e o radial (RBF). O SVM mostra um desempenho excelente em cenários com grande número de variáveis e quantidade reduzida de amostras, sendo amplamente empregado em setores como bioinformática, reconhecimento de padrões e categorização de imagens [52,53,54,55].

Já KNN, é um classificador não paramétrico que opera com base em instâncias, sem criar um modelo explícito durante o processo de treinamento. A determinação da classificação ocorre por meio

da avaliação dos k vizinhos mais próximos de uma nova entrada, utilizando uma métrica de distância, normalmente a euclidiana, e atribuindo a classe que aparece com maior frequência entre esses vizinhos. Devido à sua simplicidade e facilidade de implementação, o KNN é amplamente empregado em situações de reconhecimento de padrões, identificação de anomalias e sistemas de recomendação [56,57].

O LDA, é uma técnica linear e probabilística que pode ser utilizada para tanto classificação quanto a diminuição de dimensão. Sua função principal é mapear os dados em um espaço de dimensão reduzida, buscando maximizar a separação entre diferentes classes, ao mesmo tempo que otimiza a relação entre variância entre classes e variância dentro das classes. Essa técnica assume que as amostras de cada classe seguem uma distribuição normal multivariada, mantendo matrizes de covariância iguais. Com base nessa premissa, o modelo desenvolve funções discriminantes lineares que ajudam a identificar a classe mais provável para uma nova amostra. É um método eficaz e fácil de entender, apresentando desempenho satisfatório quando as condições de normalidade e homogeneidade de covariância são cumpridas. Contudo, sua eficácia tende a diminuir em situações que envolvem fronteiras não lineares, a presença de outliers, ou quando há mais variáveis do que amostras disponíveis [53,58,59].

Para um grupo restrito de amostras, uma abordagem frequentemente aplicada é a validação cruzada. Um dos métodos desse processo, é o LOOCV *(Leave-One-Out Cross Validation*), onde uma amostra é separada para validação, enquanto as outras são usadas para o treinamento, o processo se repete garantindo que todas serão utilizadas para ambas às ações. É especialmente benéfico para grupos reduzidos de amostras, pois otimiza os dados empregados no aprendizado, apesar de demandar mais recursos computacionais. Ademais, diminui a variação nos subconjuntos de treinamento e teste, o que pode ser útil em procedimentos de ajuste de hiperparâmetros. como o *Grid Search* [60].

Com base nas projeções do modelo, é possível analisar seu desempenho através de critérios conhecidos como medidas de eficiência. Para isso, emprega-se uma tabela denominada matriz de confusão, que facilita a visualização do desempenho do modelo por categoria. Para modelos de classificação, uma das métricas mais frequentes é a acurácia, que indica a fração de previsões corretas em comparação ao total de previsões [61].

Na LOOCV, o SVM é treinado n vezes, cada vez sem uma amostra diferente, e a predição dessa amostra é comparada ao seu rótulo verdadeiro. Essa abordagem é útil para avaliar a robustez

do hiperplano e a sensibilidade do modelo a outliers, já que cada instância tem a oportunidade de ser testada individualmente [62].

No contexto da LOOCV, o LDA é reestimado a cada iteração sem a amostra de teste, recalculando as médias e covariâncias. Esse processo permite avaliar a estabilidade dos parâmetros estatísticos do modelo e fornece uma estimativa precisa da taxa de erro esperada em novos dados [63].

Já para o KNN, o LOOCV é mais simples de ser aplicado, pois cada amostra pode ser classificada considerando todas as outras como possíveis vizinhas, ou seja, sem incluir a própria amostra no conjunto de treino. O LOOCV permite otimizar o valor de *k* escolhendo aquele que maximiza a acurácia média ao longo das iterações [64].

3. MATERIAIS E MÉTODOS

3.1. Coleta e Preparação de amostras

A priori, as amostras referem-se à penas de aves de filhotes, macho e fêmea da espécie *Oryzoborus maximiliani* (Bicudo). Foram recebidas 30 amostras de cada sexo, sendo cada uma de um indivíduo diferente. Foram amostras fornecidas pela empresa de sexagem de aves 'Codex Gen", localizada em Campo Grande, MS. Todas as amostras tiveram o sexo previamente identificado por meio da técnica de referência de análise de DNA [68].

Além disso, as amostras cedidas, foram previamente analisadas pela espectroscopia no infravermelho por transformada de Fourier (FTIR - *Fourier Transform Infrared Spectroscopy*) realizadas no laboratório GOF (Grupo de Óptica e Fotônica), localizado na Universidade Federal de Mato Grosso do Sul (UFMS), em um aparelho espectrômetro da marca "Perkin Elmer", modelo "Spectrum 100 Series", na região do infravermelho médio (MIR – *Mid-Infrared*). Os espectros de infravermelho médio (MIR) exibem diversas bandas de absorção correspondentes às transições fundamentais das moléculas. A análise desses espectros é geralmente feita considerando duas regiões principais: a região dos grupos funcionais (4000–1300 cm^{-1}) e a região de impressão digital (1300–600 cm^{-1}), que fornece uma identificação única para cada molécula [69].

A região dos grupos funcionais pode ser ainda subdividida em zonas características de diferentes tipos de ligações: Estiramentos X–H (4000–2500 cm^{-1}; X = C, N, O ou S); Ligações triplas (2700–1850 cm^{-1}); Ligações duplas (2000–1500 cm^{-1}; C=C, C=N, C=O) [69]. A região de

impressão digital, é caracterizada por um conjunto complexo de vibrações de flexão, apresentando numerosas bandas de absorção, muitas vezes sobrepostas. Essas bandas são altamente específicas da estrutura molecular de cada substância, funcionando como uma "assinatura" única que permite sua identificação [70].

Essa divisão facilita a interpretação das bandas de absorção e a identificação dos grupos funcionais presentes nas moléculas. Foram analisados espectros na faixa de 600 a 4000 cm^{-1}, com incremento de número de onda de 0,5 cm^{-1}, no modo de transmitância. As amostras foram examinadas de forma direta no acessório de reflexão total atenuada, conhecido como ATR (*Attenuated Total Reflection*), com o uso de um cristal de seleneto de zinco (ZnSe). As observações das amostras foram realizadas duas vezes.

3.2. Pré – processamento e Análise de Componente Principal (PCA)

Inicialmente, os dados foram tratados com a metodologia SNV (*Standard Normal Variate*), que tem como função padronizar os espectros e minimizar o ruído, normalizando-os conforme a Equação 1. Após isso, foi determinado o espectro médio para cada categoria (masculino e feminino). Esses dados foram analisados para descobrir eventuais diferenças ou semelhanças nos espectros.

Os algoritmos foram feitos utilizando a linguagem de programação Python. O PCA representa uma metodologia estatística que converte um grupo de variáveis que podem estar interligadas em um novo grupo de variáveis que não possuem correlação entre si, conhecidas como componentes principais (PCs). Essas PCs surgem de combinações lineares das variáveis de partida, organizadas de tal modo que a primeira componente abrange a maior parte da variabilidade dos dados, a segunda cobre a segunda maior parte da variabilidade, e assim por diante. Tal procedimento diminui a complexidade dos dados, mantendo as informações mais significativas para a análise [65]. Com ele foi possível observar a formação de grupos por classe e identificar quais PC mais contribuíram para a variância. Nos dados fornecidos, foram analisados espectros na faixa de 600 a 4000 cm^{-1}.

3.3. Aprendizagem de maquina

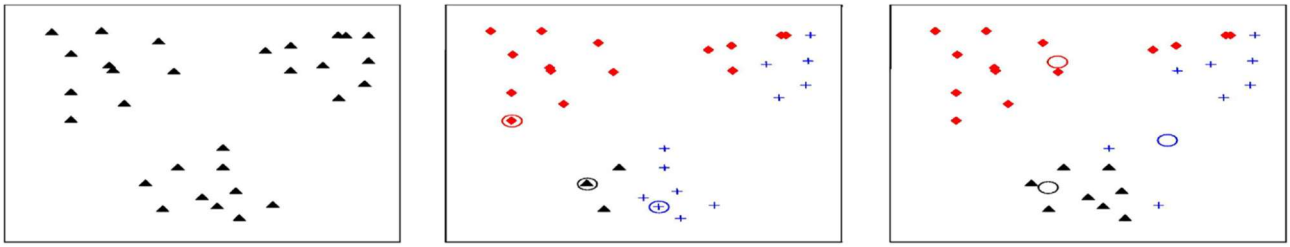
O método de agrupamento *K-means* é uma abordagem de clusterização bastante comum. Esse algoritmo é o recurso de clusterização mais utilizado em contextos científicos. Trata-se de uma técnica de análise de agrupamentos que visa dividir as observações em k grupos, onde cada observação se associa ao grupo cuja média é mais aproximada [66]. O algoritmo genérico é muito simples, ele segue um passo a passo [67]:

1. Escolher o número de K;
2. Cada ponto será atribuído ao K mais próximo, ou seja, o pré-definido;
3. Dessa forma, o sistema de agrupamento *K-means* estará pronto. A fórmula (Equação 2) para encontrar a medida de distância mostra as semelhanças entre dois elementos e tenta influenciar a forma de vários grupos de coisas.

$$D = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2} \quad (2)$$

Onde x_1 e x_2 são os pontos de dados e c_1 e c_2 são os pontos mais próximos dos centroides. A abordagem de agrupamento *K-means* opera separando os itens em diversas categorias de grupos referidas pelo número 'K'. Assim, se definirmos K = 3, o item é classificado no grupo c_1 e no grupo c_2, como mostra a Figura 3 [67].

Figura 3- Representação de como funciona o agrupamento K-means, definido por K = 3 se comporta.



Fonte: Referencia 75, adaptada pela autora.

Após o PCA, 25 PCs foram analisadas pelo *K-means* e foi calculada a acurácia para saber qual PC individual teve a melhor contribuição. A partir dessas acurácias, foram feitos os *loadings* das PCs com as melhores contribuições. Em seguida, foi realizada a validação do tipo *Leave-One-*

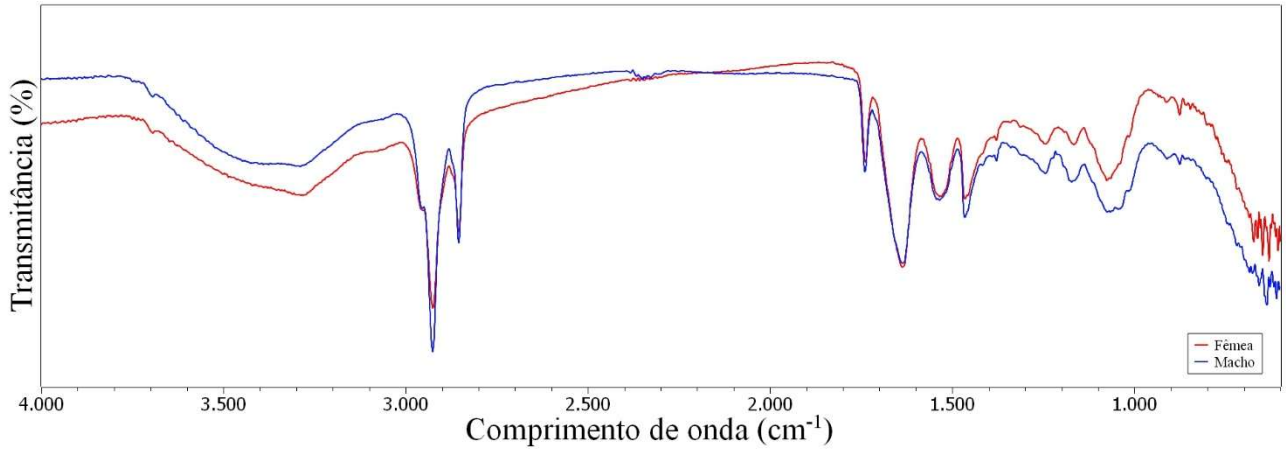
Out Cross-Validation (LOOCV), onde só as melhores PCs selecionadas pelo *K-means* foram utilizadas como variáveis de entrada para os algoritmos SVM, KNN e LDA.

Uma das vantagens de se usar o LOOCV, é a eficiência dele em um conjunto de dados pequeno, onde cada detalhe conta. A partir dessa validação, foram construídas as matrizes de confusão, permitindo comparar a eficiência de cada classificação por sexo de cada classificador.

4. RESULTADOS E DISCUSSÃO

Na Figura 4, as médias dos espectros do pássaro Bicudo são comparadas entre as categorias (Fêmea e Macho). É possível observar uma grande similaridade entre os espectros, sendo que a utilização de técnicas de *machine learning* é essencial para reconhecer e entender suas distinções.

Figura 4 – Espectro médio da sexagem de ave, em preto é a transmitância relacionada ao sexo feminino e em vermelho é a transmitância relacionada ao sexo masculino da ave Bicudo.



Fonte: Autora.

A Tabela 1, adaptada da referência [68], identifica as bandas da amostra analisada.

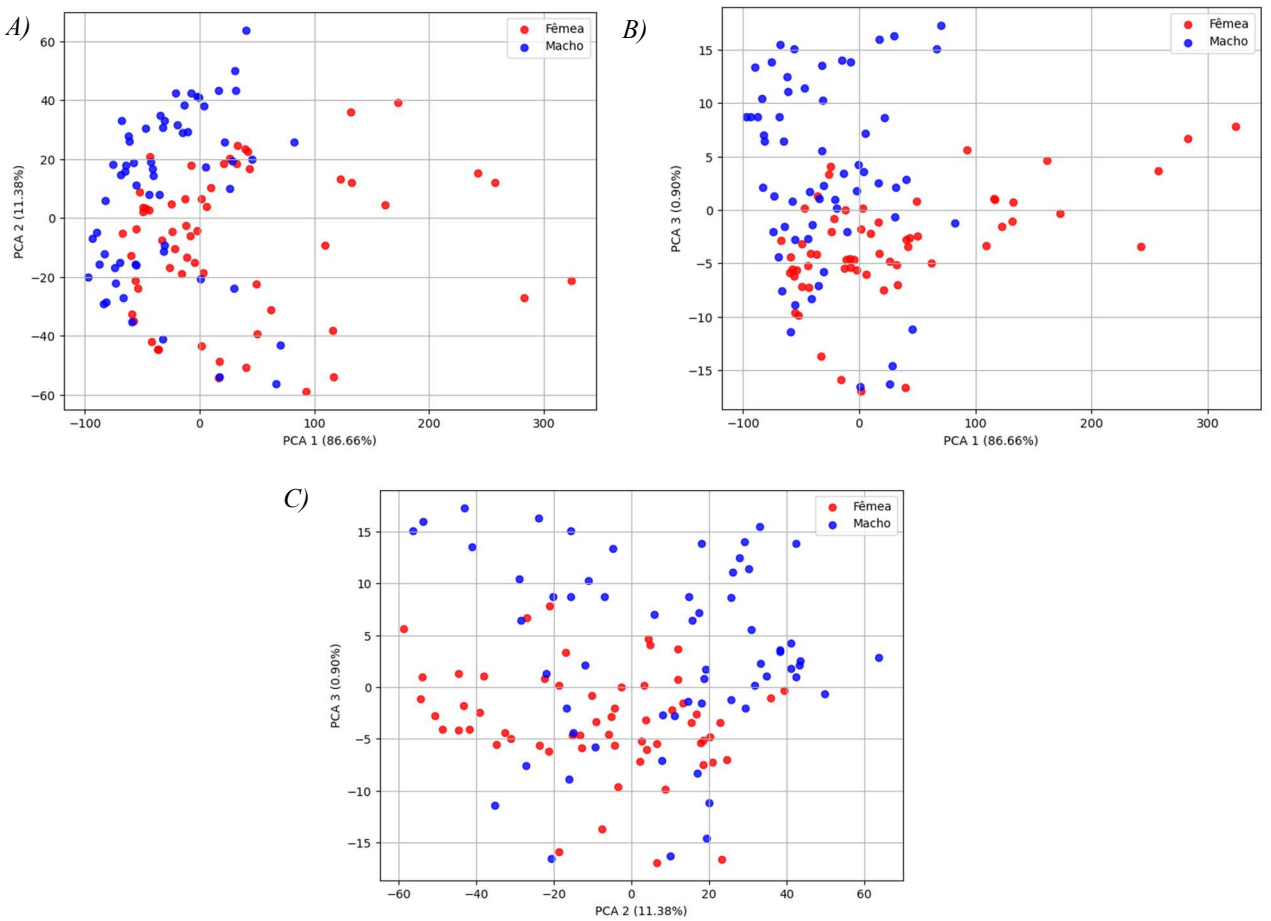
Tabela 1- Atribuição de picos em espectros FTIR de penas.

Número de onda (cm^{-1})	**Pico atribuído**
1075	C-C
1245	(CN) amida III
1535	(N-H) amida II, folha β
1655	(C=O) amida I, hélice α
2870	CH_2
2925	CH_3

Fonte: Referência 68, adaptada pelo autor.

A etapa do PCA proporcionou uma melhor análise dos dados em algo mais preciso e eficiente. Após a redução de dimensionalidade, o PCA facilitou a identificação dos padrões de maneira relevante para a classificação. O PCA distinguiu os espectros de diferentes amostras. Nota-se os espectros da PC1 no eixo positivo, em sua maioria pertencem ao sexo feminino, enquanto no eixo negativo o que predomina é o sexo masculino. Da mesma forma, os gráficos de dispersão da PC1 com a PC3 (Figura 5.B) e da PC2 com a PC3 (Figura 5.C), mostra que também podem ser utilizados para distinção dos sexos da ave.

Figura 5 - Score plot das 3 primeiras PCs. A) PC1 versus PC2. B) PC1 versus PC3. C) PC2 versus PC3.



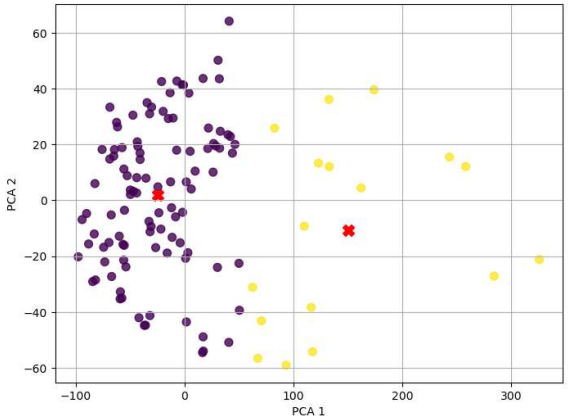
Fonte: Autora.

Os gráficos na Figura 4 mostram o *score plots* das três primeiras componentes principais (PC1, PC2 e PC3), com a diferenciação entre os grupos de fêmeas e machos, analisadas na faixa espectral 4000-600 cm^{-1}. A análise visual demonstra que a PC1, que explica 86,66% da variância total, abriga a maior parte das informações essenciais para distinguir os grupos, indicando que esta componente é a mais significativa na captura das diferenças estruturais entre as amostras. Nota-se uma tendência de dispersão distinta entre fêmeas e machos.

A fase de implementação do *K-means* envolve a divisão das duas categorias com base na localização de seus centroides. O gráfico mostrado na Figura 6, ilustra os resultados do agrupamento *K-means* que foi utilizado nos dados após a aplicação da redução de dimensionalidade através do PCA. Cada ponto simboliza uma amostra, posicionada de acordo com as duas primeiras componentes

principais (PCA 1 e PCA 2), ao passo que as cores refletem os clusters gerados pelo algoritmo, e os símbolos “X” vermelhos destacam os centroides de cada conjunto.

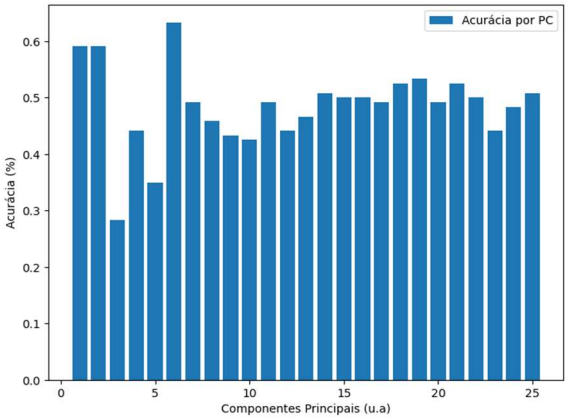
Figura 6 - Aplicação do método de clusterização *K-means* para separação de sexagem de ave.



Fonte: Autora.

A Figura 7 apresenta a acurácia obtida pelo *K-means* quando cada componente principal é analisada individualmente. Nessa etapa, foi realizado a comparação direta de agrupamento realizado pela dispersão de cada PC e a sua classe correspondente (macho ou femea). Dessa forma, temos uma métrica que avalia de maneira não supervisionada a capacidade que cada PC tem em separar de maneira espontânea as classes. Observa-se que a acurácia não permanece constante, algumas PCs apresentam desempenho superior, indicando que possuem maior poder discriminante. Ele mostra como o desempenho do agrupamento varia conforme aumenta a quantidade de PCs consideradas após a redução de dimensionalidade pelo PCA. A comparação entre as Figuras 5 e 6 evidencia que a separação observada visualmente no espaço bidimensional das duas primeiras PCs (Figura 5) está alinhada com a métrica de desempenho mostrada na Figura 7. Em vista disso, a precisão foi determinante para 25 PCs. Consequentemente, o gráfico de barras indica quais PCs obtiveram melhor desempenho com os métodos utilizados.

Figura 7 - Acurácia para 25 PCs.



Fonte: Autora.

Observa-se que a acurácia não cresce de forma linear com o número de componentes, em vez disso, há flutuações no desempenho, com alguns picos de melhor resultado. As componentes 1,2 e 6 apresentam valores relativamente altos, indicando que elas concentram boa parte da variância e da informação mais relevante para a separação dos grupos.

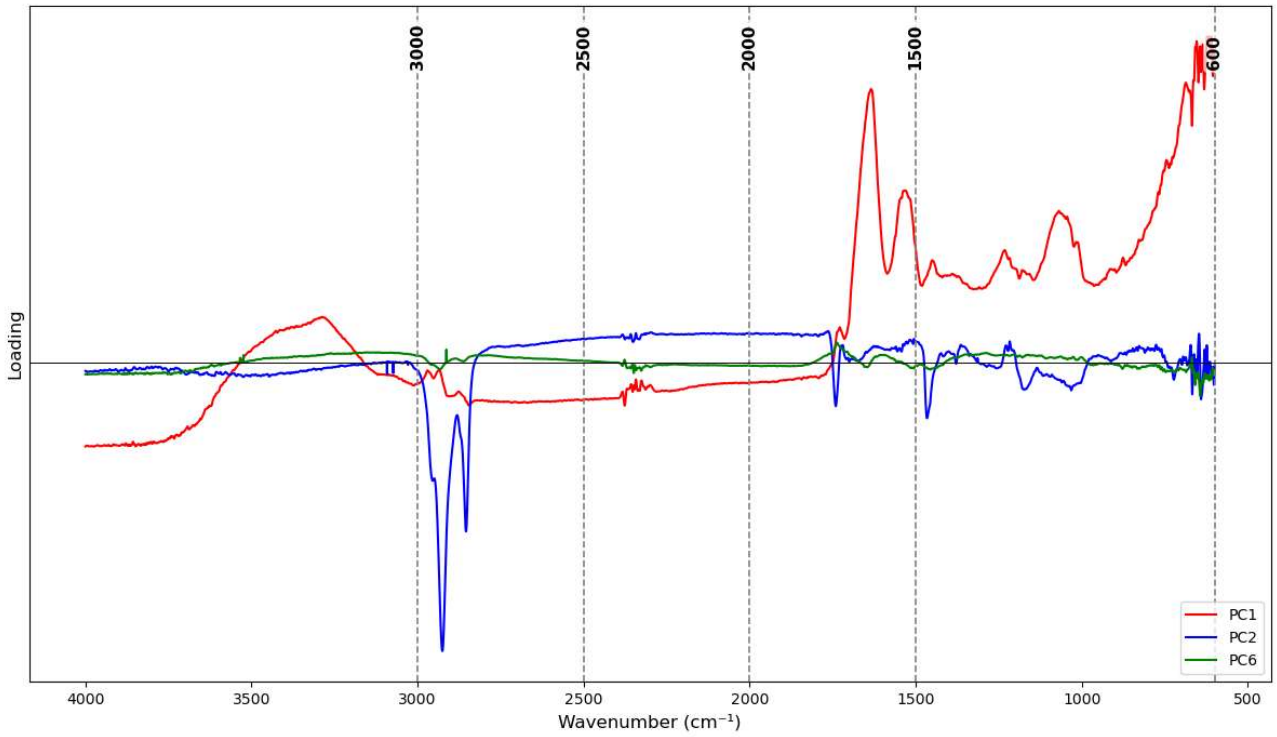
Quando há interesse em identificar as transições mais relevantes para a separação das classes, o gráfico de *loadings* (Figura 8) pode fornecer essas informações de maneira clara. Nesse contexto, os *loadings* correspondem às componentes principais (PCs) que melhor discriminam as amostras. Assim, as informações apresentadas no gráfico estão diretamente correlacionadas às transições mais indicadas para serem utilizadas como marcadores no processo de separação.

Na Figura 8, observamos a eficácia das PC1, PC2 e PC6. Os *loadings* mostram uma influência significativa nas faixas localizadas entre 3000 e 2500 cm^{-1}, assim como na área que varia entre 2000 e 1500 cm^{-1}. O gráfico indicou que essas bandas correspondem a características de lipídios e proteínas, assim como na literatura [68].

A primeira faixa (3000–2500 cm^{-1}) refere-se às vibrações de estiramento C-H encontradas nos grupos metileno e metila, frequentemente ligadas às cadeias de hidrocarbonetos de lipídios, ceras e ácidos graxos. Essas bandas estão profundamente conectadas às camadas superficiais de lipídios das penas, englobando os ésteres de cera liberados pela glândula uropigial, que exercem funções cruciais como impermeabilização, proteção e, possivelmente, comunicação entre membros da mesma espécie

[68]. A segunda faixa (2000–1500 cm^{-1}) inclui bandas ligadas tanto aos lipídios quanto às proteínas, indicando a presença e a interação desses dois principais constituintes moleculares nas amostras.

Figura 8 - *Loadings* para as componentes selecionadas pelo *K-means*. Em vermelho, PC1. Em azul, PC2. Em verde PC6.



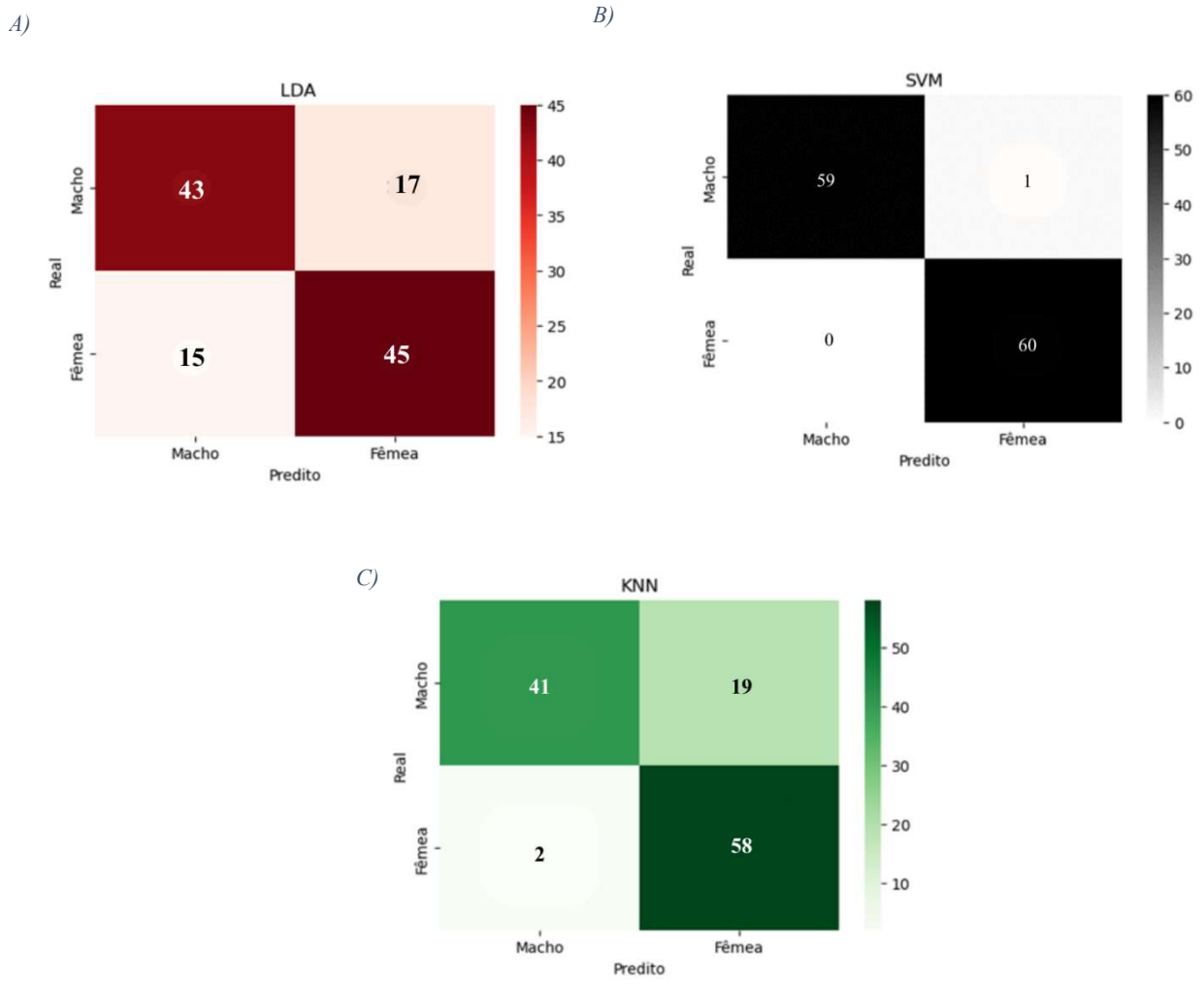
Fonte: Autora.

O gráfico de *loadings* ilustra a participação das componentes principais PC1, PC2 e PC6 na variabilidade dos dados espectrais escolhidos pelo método *K-means*. Verifica-se que a PC1 exibe as maiores flutuações, em especial nas faixas de 600–1800 cm^{-1} e 2800–3000 cm^{-1} (Tabela 1), indicando que essas faixas do espectro são as que mais contribuem para a diferenciação e separação dos grupos moleculares, uma vez que refletem variações estruturais específicas entre as amostras. Essas regiões costumam estar associadas a bandas de vibrações de ligações como C–H, C=O, muito comuns em compostos orgânicos, o que sugere que essas áreas são as mais críticas para diferenciar os grupos e estão ligadas às principais distinções químicas das amostras. A PC2 também evidencia picos significativos nessas regiões, refletindo variações secundárias, mas ainda relevantes, relacionadas à

força de determinadas bandas. Por outro lado, a PC6 mostra valores muitos pequenos em todo o espectro, indicando uma contribuição reduzida para a discriminação dos clusters.

Na Figura 9, são mostradas as aplicações dos classificadores LDA, SVM e KNN, que exibem a quantidade de acertos e erros através da matriz de confusão relacionada às PC1, PC2 e PC6, selecionadas pelo K-means.

Figura 9 – Desempenho dos classificadores para as componentes principais 1, 2 e 6 selecionada pelo *K-means*. A) Classificador LDA. B) Classificador SVM. C) Classificador KNN.



Fonte: Autora.

As PCs 1,2 e 6, foram analisadas em conjunto para cada classificador. O classificador LDA alcançou uma precisão em torno de 73,33%, mostrando um desempenho equilibrado entre as duas categorias. Foram identificados corretamente 43 machos e 45 fêmeas, o que demonstra que o modelo teve razoável eficácia na distinção das amostras, embora ainda apresentasse uma taxa de erro considerável. Em contraste, o classificador SVM se destacou entre os três modelos, obtendo uma precisão de aproximadamente 99,17%. Este modelo corretamente classificou 59 machos e 60 fêmeas, cometendo apenas uma falha em todo o conjunto de dados. Esse resultado indica que o SVM conseguiu identificar uma fronteira ideal no espaço de características, maximizar a margem entre as categorias e assegurar uma alta capacidade de generalização.

Por sua vez, o classificador KNN mostrou um desempenho mediano, com uma precisão de cerca de 82,50%. Levou ao acerto na classificação de 41 machos e 58 fêmeas, revelando uma precisão maior na identificação da classe “Fêmea”. O modelo cometeu apenas dois erros nesse grupo, mas demonstrou mais confusão ao reconhecer machos, apresentando 19 classificações incorretas. Isso sugere que o KNN foi mais responsivo às características das amostras femininas, possivelmente devido à distribuição dos dados ou ao número de vizinhos utilizado, em que k foi igual a 5. Na Tabela 1, apresenta de desempenho nos classificadores LDA, SVM e KNN.

Tabela 2 – Acurácias *29earch29das* aos classificadores relacionadas às PC1, PC2 e PC6.

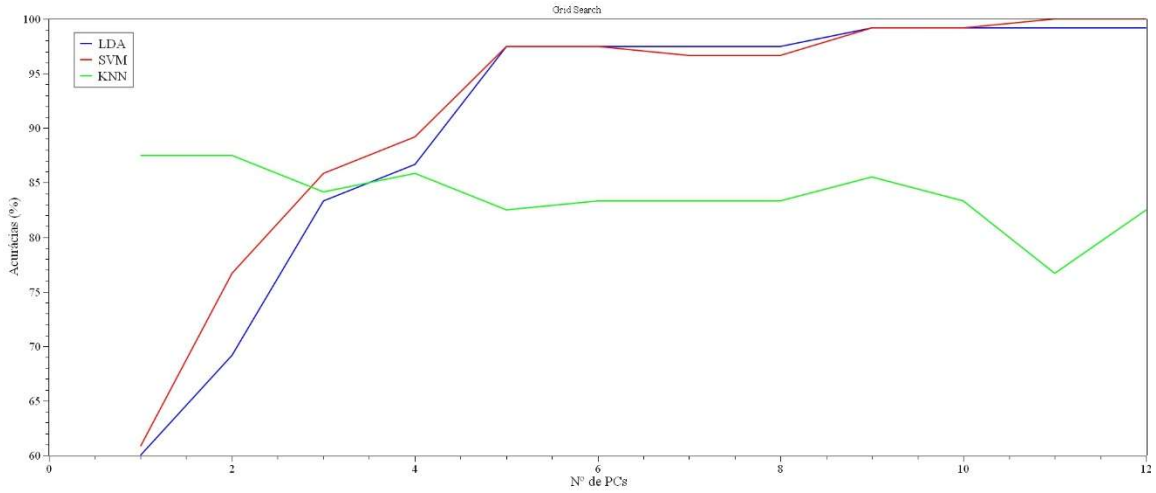
Classificador	**Acurácia**
LDA	73,33%
SVM	99,17%
KNN	82,50%

Fonte: Autora.

Após essa etapa, para efeito comparativo, foi realizado uma otimização manual das PCs nos classificadores. Na Figura 10, observa-se o gráfico do Grid Search. Nele, apresenta a acurácia para 12 PCs. O gráfico de *grid 29earch* ilustra a variação na precisão dos classificadores LDA, SVM e KNN em relação ao número de componentes principais (PCs) utilizados. O processo ocorre de maneira sequencial: primeiro, é elaborado um modelo que leva em conta apenas a primeira componente principal (PC1), a qual captura a maior parte da variação nos dados. Em seguida, são integradas a PC1 e a PC2, fazendo com que o modelo utilize uma quantidade mais ampla de dados

originais. Depois, é incluída também a PC3, e este procedimento continua até que todas as componentes principais significativas sejam incorporadas na análise. A cada nova adição de um componente, o classificador passa por uma nova avaliação e a precisão é medida, gerando diferentes níveis de desempenho à medida que a quantidade de PCs cresce. Este processo é realizado para cada classificador (LDA, SVM e KNN), e os resultados são comparados de forma gráfica.

Figura 10 – Grid Search para 12 PCs selecionadas pelo *K-means.* Em azul, o classificador LDA. Em vermelho, o classificador SVM. Em verde, o classificador KNN



Fonte: Autora.

Pode-se notar que, de maneira geral, a performance dos classificadores tende a se aprimorar com o aumento do número de componentes até um certo limite, após o qual se estabiliza.

O classificador SVM mostra o melhor desempenho durante quase todo o intervalo, alcançando precisões que se aproximam de 100% a partir de cerca de cinco componentes principais, o que revela sua notável habilidade de generalização e eficácia na distinção das classes. O LDA exibe um comportamento similar, com uma progressão gradual da precisão conforme o número de PCs cresce, atingindo também valores altos próximos aos do SVM nas últimas componentes.

Em contrapartida, o KNN demonstra um desempenho mais irregular e inferior em comparação aos outros, apresentando pouca variação e uma tendência à queda na precisão após um certo ponto, sugerindo uma possível suscetibilidade ao aumento da dimensionalidade e à presença de

ruídos nos dados. De maneira geral, o gráfico ressalta que o SVM é o classificador mais robusto e estável em diferentes dimensões do espaço de variáveis, seguido pelo LDA, enquanto o KNN revelou uma atitude mais sensível e menos eficaz no processo de otimização dos parâmetros.

5. CONCLUSÃO

Os resultados obtidos deixam claro o potencial das abordagens de aprendizado de máquina quando utilizadas na espectroscopia para o processo de determinação do sexo em aves, especialmente na espécie Bicudo. A avaliação dos espectros médios inicialmente mostrou uma grande semelhança entre os sexos, evidenciando a necessidade de técnicas computacionais mais sofisticadas para identificar diferenças sutis e complexas que não podem ser vistas em uma análise visual tradicional.

Dentro desse cenário, a Análise de Componentes Principais (PCA) foi crucial ao simplificar os dados e realçar as áreas espectrais mais relevantes. A primeira componente principal (PC1), que explica 86,66% da variância total, foi fundamental para distinguir os grupos, com uma tendência que concentra as amostras femininas no lado positivo e as masculinas no lado negativo, o que indica que ela capta as diferenças químicas mais significativas entre os sexos.

Além disso, o método *K-means* desempenhou um papel importante no agrupamento das amostras de forma não supervisionada, demonstrando que, na ausência de etiquetas anteriores, os dados revelam padrões claros entre as diferentes categorias. Os resultados do gráfico de *loadings* reforçaram essa constatação, mostrando que as áreas entre 500–1800 cm^{-1} e 2800–3000 cm^{-1} têm um impacto mais significativo na distinção. Essas faixas espectrais costumam estar ligadas a vibrações moleculares de grupos funcionais orgânicos, que podem indicar diferenças bioquímicas sutis relacionadas à composição hormonal, metabólica ou estrutural entre os sexos masculino e feminino.

Na fase supervisionada, os classificadores LDA, KNN e SVM demonstraram desempenhos variados, possibilitando uma comparação direta da eficiência de cada método. O LDA apresentou uma boa relação entre sensibilidade e especificidade, alcançando 73,33% de acurácia, enquanto o KNN conseguiu 82,50%, mas com uma tendência maior a confundir amostras masculinas. Já o SVM se destacou como a abordagem mais sólida, atingindo uma precisão de 99,17%, praticamente sem

erros de classificação. Esse desempenho excepcional é atribuído à sua habilidade de identificar um limite ideal entre as classes no espaço multidimensional das componentes principais, maximizando a distinção entre fêmeas e machos, mesmo frente a sobreposições sutis.

A validação com o *Grid Search* para um máximo de 12 principais componentes mostrou que a eficácia dos classificadores aumenta até um certo limite e depois se estabiliza, com o SVM apresentando maior consistência e capacidade de generalização em várias configurações (100% de acuracia). Esse padrão reafirma a competência do modelo em reconhecer padrões complexos, estabelecendo-o como a opção mais adequada para esse tipo de aplicação.

Conclui-se que a pesquisa mostrou que a utilização conjunta de PCA, *K-means* e classificadores supervisionados foi extremamente eficaz para a determinação do sexo das aves Bicudo. Apesar de atingido resultados satisfatórios com e sem a utilização do *K-means*, o *K-means* escolheu as componentes principais (PCs) mais relevantes que possuem a melhor habilidade para diferenciar machos de fêmeas. Dentre os classificadores, o SVM teve o desempenho mais sobressalente, apresentando elevada precisão e capacidade de generalização. Dessa forma, a aplicação integrada dessas metodologias provou ser uma abordagem eficaz e não invasiva para a identificação do sexo das aves e interpretabilidade, com a possibilidade de ser utilizada em outras espécies.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] SHUKLA, Usha**.** Fourier transform infrared spectroscopy: a power full method for creating fingerprint of molecules of nanomaterials**.** ***Journal of Molecular Structure***, **v. 1322**, p. 140454, 2024.

[2] PASIECZNA-PATKOWSKA, S.; CICHY, M.; FLIEGER, J**.** Application of Fourier Transform Infrared (FTIR) Spectroscopy in characterization of green synthesized nanoparticles. ***Molecules***, **v. 30**, p. 684, 2025.

[3] DAS, S.; BHATI, V.; DEWANGAN, B. P. et al. Combining Fourier-transform infrared spectroscopy and multivariate analysis for chemotyping of cell wall composition in mungbean (*Vigna radiata* (L.) Wizcek). ***Plant Methods*****, v. 20**, p. 135, 2024

[4] SU, K.-Y.; LEE, W.-L. Fourier Transform Infrared Spectroscopy as a cancer screening and diagnostic tool: a review and prospects. ***Cancers*****, v. 12**, n. 1, p. 115, 2020.

[5] BRITO, N. M. R. DE.; LOURENÇO, F. R.. Rapid identification of microbial contaminants in pharmaceutical products using a PCA/LDA-based FTIR-ATR method**. Brazilian Journal of Pharmaceutical Sciences**, v. 57, p. e18899, 2021.

[6] AVCU, F. M**.** Clustering honey samples with unsupervised machine learning methods using FTIR data. ***Anais da Academia Brasileira de Ciências*****, v. 96**, n. 1, e20230409, 2024.

[7] MORINHA, F.; CABRAL, J. A.; BASTOS, E. Molecular sexing of birds: a comparative review of polymerase chain reaction (PCR)-based methods. ***Theriogenology*****, v. 7**8, n. 4, p. 703–714, 2012.

[8] STEINER, Gerald et al**.** Label free molecular sexing of monomorphic birds using infrared spectroscopic imaging. ***Talanta*****, v. 150**, p. 155–161, 2016.

[9] REYNOLDS, S. J.; MARTIN, G. R.; WALLACE, L. L.; WEARN, C. P.; HUGHES, B. J. Sexing sooty terns on Ascension Island from morphometric measurements. **Journal of Zoology**, v. 274, n. 1, p. 2–8, 2008.

[10] VOLODIN, Ilya; KAISER, Martin; MATROSOVA, Vera; VOLODINA, Elena; KLENOVA, Anna; FILATOVA, Olga; KHOLODOVA, Marina. The technique of noninvasive distant sexing for four monomorphic Dendrocygna whistling duck species by their loud whistles.

Bioacoustics: The International Journal of Animal Sound and its Recording, v. 18, p. 277–290, 2009.

[11] RICHNER, Heinz. Avian laparoscopy as a field technique for sexing birds and an assessment of its effects on wild birds. **Journal of Field Ornithology**, v. 60, n. 2, p. 137–142, 1989.

[12] GRAY, Catherine M.; HAMER, Keith C. Food-provisioning behaviour of male and female Manx shearwaters (Puffinus puffinus). **Animal Behaviour**, v. 62, p. 117–121, 2001.

[13] MARON, J. L.; MYERS, J. P. A description and evaluation of two techniques for sexing wintering sanderlings**. Journal of Field Ornithology**, v. 55, n. 3, p. 336–342, 1984.

[14] STEINER, G.; et al. Bird sexing by Fourier Transform Infrared Spectroscopy. **In: Biomedical Vibrational Spectroscopy IV: Advances in Research and Industry**. SPIE, 2010

[15] STEINER, G.; et al. Sexing of turkey poults by Fourier Transform Infrared Spectroscopy. **Analytical and Bioanalytical Chemistry**, v. 396, n. 1, p. 465–470, 2010.

[16] 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.

[17] DIAS, E. A.; OLIVEIRA, C. DE. Psittacine sex determination by radioimmunoassay (RIA) of sex steroids using fecal samples. 2006.

[18] ELLEGREN, H.; FRIDOLFSSON, A.-K. Male–driven evolution of DNA sequences in birds. **Nature Genetics**, v. 17, n. 2, p. 182–184, 1997.

[19] 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.

[20] FERREIRA, M.; et al. Quimiometria I: calibração multivariada, um tutorial**. Química Nova**, v. 22, p. 724–731, 1999.

[21] FERREIRA, M. M. C. Quimiometria: conceitos, métodos e aplicações**. Campinas: Editora da UNICAMP**, 2015.

[22] GEMPERLINE, P. Practical guide to chemometrics. **Boca Raton: CRC Press**, 2006.

[23] GONÇALVES, V. P. Monitoramento da estabilidade oxidativa de biodiesel empregando espectroscopia vibracional associada a ferramentas quimiométricas. 2022.

[24] GRANDO, A. P. Utilização de tomografia por ressonância magnética nuclear para sexagem de aves silvestres sem dimorfismo sexual. 2002.

[25] GRIFFITHS, R.; et al. A DNA test to sex most birds. **Molecular Ecology**, v. 7, n. 8, p. 1071–1075, 1998.

[26] 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.

[27] SKOOG, Douglas A.; HOLLER, F. James; CROUCH, Stanley R**.** ***Principles of instrumental analysis*****.** 6. ed. Boston: Cengage Learning, 2019.

[28] 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.

[29] HOLLAS, J. M. Modern spectroscopy. 4. ed. Chichester: John Wiley & Sons, 2004

[30] LILO, Taha; MORAIS, Camilo L. M.; SHENTON, Catriona; RAY, Arup; GURUSINGHE, Nihal**.** Revising Fourier-transform infrared (FT-IR) and Raman spectroscopy towards brain cancer detection**.** ***Photodiagnosis and Photodynamic Therapy***, v. 38, p. 102785, 2022.

[31] 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.

[32] KELLEHER, J. D.; MAC NAMEE, B.; D'ARCY, A. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. Cambridge: **MIT Press**, 2020.

[33] KNACKFUSS, F. Sexagem de aves da espécie Amazona aestiva (Papagaio verdadeiro) por meio de técnica de PCR. **Pubvet,** v. 14, n. 6, 2020.

[34] 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, e0161788, 2016.

[35] HONGYU, Kuang; SANDANIELO, Vera Lúcia Martins; OLIVEIRA JUNIOR, Gilmar Jorge de. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S – Engineering and Science**, v. 5, n. 1, p. 83–91, 2016.

[36] 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

[37] LEITE, A. A.; FERREIRA, J. L.; DE OLIVEIRA, R. O. R. G. Técnica de sexagem por laparoscopia em arara-canindé (Ara ararauna). **Brazilian Journal of Animal and Environmental Research**, v. 5, n. 4, p. 3641–3643, 2022.

[38] RINNAN, Åsmund; VAN DEN BERG, Frans; ENGELSEN, Søren Balling. Review of the most common pre-processing techniques for near-infrared spectra. **Trends in Analytical Chemistry**, v. 28, n. 10, p. 1201–1222, 2009.

[39] GUO, Q.; WU, W.; MASSART, D. L. The robust normal variate transform for pattern recognition with near-infrared data. **Analytica Chimica Acta**, v. 382, p. 87–103, 1999.

[40] MAROCHI, M. Z.; et al. Dimorfismo sexual em Hepatus pudibundus (Crustacea, Decapoda, Brachyura). **Iheringia. Série Zoologia**, v. 106, 2016.

[41] MARTINS, Aline L. S. et al. Near-Infrared (NIR) spectroscopy and chemometrics applied to the identification of counterfeit and substandard medicines: A systematic review**. Sensors**, v. 22, n. 24, p. 9764, 2022.

[42] NICOLAI, Bart M. et al. Nondestructive measurement of the microstructure of food products**. Trends in Food Science & Technology**, v. 21, n. 12, p. 871–885, 2010.

[43] MÜLLER, A. C.; GUIDO, S. Introduction to machine learning with Python: a guide for data scientists**. Sebastopol: O'Reilly Media**, 2016

[44] ECKHARDT, C. M.; MADJAROVA, S. J.; WILLIAMS, R. J. et al. Unsupervised machine learning methods and emerging applications in healthcare. **Knee Surgery, Sports Traumatology, Arthroscopy**, v. 31, n. 2, p. 376–381, 2023.

[45] MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. p. 281–297.

[46] JAIN, A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, v. 31, n. 8, p. 651–666, 2010.

[47] KHANORKAR, Yash; KANE, P. V. Selective inventory classification using ABC classification, multi-criteria decision making techniques, and machine learning techniques. **Materials Today: Proceedings**, v. 72, p. 1270–1274, 2023.

[48] LLOYD, S. P. Least squares quantization in PCM. IEEE Transactions on Information Theory, v. 28, n. 2, p. 129–137, 1982

[49] ARORA, S.; KAPOOR, R. Clustering algorithms: A comprehensive review. International Journal of Computer Applications, v. 75, n. 13, p. 1–7, 2013.

[50] MIKENGUYEN13. *Supervised Machine Learning — Machine Learning in Python*.

[51] BZDOK, D.; KRZYWINSKI, M.; ALTMAN, N. Aprendizado de máquina: métodos supervisionados. **Nature Methods**, v. 15, p. 5–6, 2018.

[52] IBM. Support Vector Machines (SVMs): What they are and how they work. 2023.

[53] SCIKIT-LEARN. Support Vector Machines and Linear Discriminant Analysis Documentation. 2024.

[54] SEROKELL. Support Vector Machine Algorithm Explained. 2023.

[55] MATHWORKS. Support Vector Machines for Binary Classification. 2024.

[56] KIM, S.; PARK, K.; KANG, M. Human activity recognition using KNN and LDA methods. **Pattern Recognition Letters**, v. 32, n. 11, p. 1443–1450, 2011.

[57] MEDIUM. Human Activity Recognition by Applying LDA, QDA and KNN. 2023.

[58] MATHWORKS. Discriminant Analysis. 2024.

[59] IBM. Linear Discriminant Analysis: Understanding the Basics. 2023.

[60] STEINER, G.; et al. Label free molecular sexing of monomorphic birds using infrared spectroscopic imaging. **Talanta**, v. 150, p. 155–161, 2016.

[61] STUART, B. H. Infrared spectroscopy: fundamentals and applications. **Chichester: John Wiley & Sons**, 2004.

[62] CHAPELLE, O.; VAPNIK, V.; BOUSQUET, O.; MUKHERJEE, S. *Choosing Multiple Parameters for Support Vector Machines*. **Machine Learning**, v. 46, n. 1–3, p. 131–159, 2002.

[63] MCCLACHLAN, G. J. *Discriminant Analysis and Statistical Pattern Recognition*. **Wiley-Interscience**, 2004.

[64] COVER, T.; HART, P. *Nearest Neighbor Pattern Classification*. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, 1967.

[65] VIEIRA, J.; COELHO, E.; OLIVEIRA, D. Sexagem molecular em aves silvestres. **Revista Brasileira de Reprodução Animal**, v. 33, n. 2, p. 66–70, 2009.

[66] UBEY, Ankita; CHOUBEY, Abha. A systematic review on K-means clustering techniques. **International Journal of Scientific Research Engineering & Technology (IJSRET)**, v. 6, n. 6, p. 624–627, 2017.

[67] SUYAL, Manish; SHARMA, Sanjay. A review on analysis of K-means clustering machine learning algorithm based on unsupervised learning. **Journal of Artificial Intelligence and Systems**, v. 6, p. 85–95, 2024.

[68] NAVES, Silvano Dias Pereira; FERNANDES, Victor Fidelis; RIBEIRO, Matheus Cícero; FRANÇA, Thiago; SENESI, Giorgio S.; SANCHES, Simone; GALVÃO, Cleber; MANTOVANI, Cynthia; CENA, Cícero; MARANGONI, Bruno. Early and noninvasive bird gender identification by ATR-FTIR spectra coupled with a Random Forest algorithm. **ACS Omega**, v. 10, p. 49118–49125, 2025.

[69] THOMPSON, J. M. Infrared Spectroscopy. **Singapura: Pan Stanford Publishing Pte. Ltd.**, 2018.

[70] Silverstein, R.M.; Webster, F.X.; Kiemle, D.J.; Bryce, D.L. Spectrometric Identification of Organic Compounds, 8th ed.; **John Wiley & Sons**, Ltd: Hoboken, NJ, USA, 2015.

[71] PROCTOR, N.; LYNCH, P. Manual of Ornithology: Avian Structure and Function. **Yale University Press**, 1993.

[72] LIU, D. et al. Characterization of feather keratin by FTIR and its application in materials science. **Journal of Spectroscopy**, 2014.

[73] 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.

[74] COUTRIM, Mauricio X. Absorção no infravermelho e luminescência molecular. 2016.

[75] BAMPI, L. N. et al. Identification of abnormal peaks in chemical substance FTIR spectra. ***Information Sciences***, v. 624, p. 359–374, 2023.