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

**Curso de FÍSICA - BACHARELADO**

Trabalho de Conclusão de Curso

# Melhoria dos Processos de Classificação Espectral: Um Estudo sobre Redes Neurais, Técnicas de Normalização e Seleção de Variáveis

***MURILO NECO SARAIVA***

***Orientador**: Prof. Dr. Bruno Spolon Marangoni*

Trabalho de Conclusão de Curso apresentado ao curso de Física Bacharelado do Instituto de Física (INFI), da Universidade Federal de Mato Grosso do Sul (UFMS).

Campo Grande – MS

Dezembro/2023

*"Eu sou eu e minha circunstância, e se não salvo a ela, não me salvo a mim."*

*(José Ortega y Gasset)*

## AGRADECIMENTOS

A Deus, à minha família, aos meus amigos, à Fight Sports Pantanal, aos meus professores da graduação e ao meu orientador, Bruno Spolon Marangoni. A todos que contribuíram e foram relevantes nessa jornada, aqui expresso a minha sincera gratidão.

## RESUMO

Nos últimos anos, diversos estudos utilizando a espectroscopia FTIR e a análise multivariada foram realizados. Seus resultados mostram a eficiência, viabilidade, precisão e potencial do método em identificar padrões complexos, classificar e discriminar amostras e automatizar processos. O presente trabalho apresenta uma metodologia que envolve o uso de técnicas de normalização, seleção de variáveis e redes neurais para a identificação de polimorfismos de nucleotídeo único no DNA de bovinos (homozigoto ou heterozigoto). Os dados espectrais foram pré-processados utilizando Variável Normal Padrão e derivadas de primeira e segunda ordem. Análise de Componentes Principais e Boruta foram aplicadas como técnicas de redução de dimensionalidade e seleção de variáveis para alimentar a rede neural. Técnicas de validação interna utilizando validação cruzada leave-one-out foram utilizadas. Otimização dos hiperparâmetros do classificador com base na melhora da acurácia e validação externa foi aplicada para averiguar a robustez do modelo. O melhor resultado obtido demonstrou uma acurácia de validação interna de 95,33% e validação externa de 98%, que é melhor do que a acurácia interna de 90% obtida pela metodologia de RIOS et. al para as mesmas classes. Isso corrobora para o uso do método para a melhoria de processos de classificação.

**Palavras-chave:** Espectroscopia FTIR, Análise Multivariada, Aprendizado de Máquina, Classificação.

## ABSTRACT

In recent years, several studies using FTIR spectroscopy and multivariate analysis have been carried out. Their results show the efficiency, feasibility, accuracy, and potential of the method in identifying complex patterns, classifying and discriminating samples, and automating processes. This paper presents a methodology that involves the use of normalization techniques, variable selection, and neural networks for the identification of single nucleotide polymorphisms in bovine DNA (homozygous or heterozygous). The spectral data were pre-processed using the Standard Normal Variate and first and second order derivatives. Principal Component Analysis (PCA) and Boruta were applied as dimensionality reduction techniques and variable selection to feed the neural network. Internal validation techniques using leave-one-out cross validation were used. Optimization of the classifier's hyperparameters based on improving accuracy and external validation was applied to verify the robustness of the model. The best result demonstrated an internal validation accuracy of 95.33% and external validation of 98%, which is better than the 90% accuracy obtained by the RIOS et. al. methodology for the same classes. This corroborates the use of the method for improving classification processes.

**Keywords:** FTIR Spectroscopy, Multivariate Analysis, Machine Learning, Classification

# LISTA DE FIGURAS

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| FTIR | Espectroscopia no infravermelho por transformada de Fourier |
| TF | Transformada de Fourier |
| RN | Redes neurais |
| VNP | Variável normal padrão |
| MZSA | Valor-Z máximo |
| PCA | Análise de componentes principais |
| PC | Componentes principais |
| VC | Validação cruzada |
| LOO-CV | Validação cruzada leave-one-out |
| CD | Heterozigótico |
| CC | Homozigótico |
| SNP | Polimorfismos de nucleotídeo único |
| RTA | Refletância total atenuada |
| MVS | Máquina de vetores de suporte |

# SUMÁRIO

# 1. INTRODUÇÃO

Com a crescente demanda por métodos de análise química que sejam rápidos, confiáveis e não destrutivos, a espectroscopia no infravermelho por transformada de Fourier, - FTIR, do acrônimo em inglês de Fourier Transform Infrared Spectroscopy -, tem se mostrado promissora e útil na identificação de componentes de um meio cultural, que vai da análise de moléculas de menor complexidade até de sistemas mais elaborados (AGUIAR, 2015; WANG, 2020; LIMA, 2021; FORATO, 2010). O FTIR é uma técnica que, além de contemplar esses aspectos, não exige dissolução ou extração de algumas amostras, o que simplifica a análise e reduz o tempo de teste. Seus espectros podem ser analisados posteriormente por outras ferramentas também (WANG, 2020; FORATO, 2010).

A espectroscopia em geral é importante para identificação de compostos orgânicos e inorgânicos, determinação de grupos funcionais, da composição molecular de superfícies, da conformação molecular e estereoquímica, da orientação molecular, etc (HSU, s.d.). Os primeiros espectrômetros infravermelhos eram dispersivos, - analógicos, menos sensíveis e lentos -, mas com o advento dos microcomputadores os espectrômetros que utilizam a Transformada de Fourier tornaram-se populares devido às suas vantagens em relação aos dispersivos: rápida varredura, melhora da razão sinal/ruído e as frequências de infravermelho fora da faixa de interesse são descartados pelo interferômetro de Michelson que os compõem (FORATO, 2010).

Os algoritmos de aprendizado de máquina, por sua vez, otimizam o desempenho de uma tarefa usando exemplos e/ou experiências passadas. Eles fazem parte de nossa vida cotidiana, sendo utilizados para reconhecimento de imagem e fala, nas pesquisas na web, na detecção de fraudes, filtragem de e-mails/spam, etc. Porém, dentro do contexto de aplicação em pesquisas científicas e industriais, as técnicas de aprendizagem têm se destacado por serem capazes de acelerar o desenvolvimento de pesquisas fundamentais e

aplicadas, pois realizam regressão, classificação, agrupamento e redução de dimensionalidade de conjuntos com uma grande quantidade de dados (SCHMIDT, 2019).

Dado o exposto, quando a espectroscopia no infravermelho e o aprendizado de máquina são combinados, eles são capazes de criar modelos preditivos. A integração dessas técnicas envolve o uso de algoritmos de aprendizado para processar os dados espectrais obtidos pelo espectrômetro e, com isso, realizar previsões, identificar e classificar amostras, detectar anomalias, etc. Essa abordagem é utilizada para analisar dados espectrais em uma grande variedade de contextos, tais como na indústria alimentícia e farmacêutica e na ciência dos materiais (AGUIAR, 2015; WANG, 2020; LIMA, 2021; ALLEGRETTA, 2020).

Neste trabalho, os dados espectrais colhidos por FTIR referentes ao polimorfismo de nucleotídeo único, - SNP, do acrônimo em inglês de Single nucleotide polymorphisms -, em DNA de bovinos serão utilizados para treinar modelos supervisionados de aprendizagem de máquina, empregando redes neurais artificiais. O objetivo é avaliar a capacidade de classificação do algoritmo e propor uma metodologia que envolva a normalização dos dados espectrais, seleção das variáveis relevantes para o problema e definição de uma arquitetura de rede neural capaz de reconhecer padrões e discriminar as amostras.

# 2. REVISÃO BIBLIOGRÁFICA

A espectroscopia por infravermelho é amplamente empregada em várias áreas por ser uma ferramenta sensível capaz de fornecer detalhes sobre a estrutura molecular e composição de amostras (AKKAS, 2007). Várias amostras, - sejam elas sólidas, líquidas ou gasosas -, podem ser analisadas e algumas aplicações comuns são: identificação por assinatura espectral, identificação de grupos funcionais, da orientação molecular, de polímeros, plásticos e resinas, detecção de impurezas, etc. (HSU, s.d.).

Na caracterização de combustíveis, por exemplo, o FTIR foi utilizado em conjunto com redes neurais artificiais e regressão de mínimos quadrados parciais para modelar propriedades do diesel, como cetano index, densidade, viscosidade, temperaturas de destilação e teor total de enxofre. Mínimos Quadrados Parciais e redes neurais artificiais foram usados como classificadores, e os resultados indicam que os modelos de calibração utilizados possuem um desempenho confiável na avaliação das propriedades do diesel consideradas neste estudo (SANTOS, 2005).

Um trabalho em bioquímica usando o FTIR investigou as mudanças moleculares proporcionadas pelo ácido lipóico no tecido cerebral de ratos. O ácido lipóico parece aumentar a quantidade de proteínas na região, porém, como previsto por redes neurais treinadas com regularização Bayesiana, isso induziu alterações não significativas na estrutura secundária das proteínas do sistema. Isso sugere que o ácido lipóico não é tóxico, apoiando seu uso como antioxidante (AKKAS, 2007). Outro estudo identificou polimorfismos de nucleotídeo único no material genético de animais. Fragmentos de DNA foram analisados por análise multivariada e algoritmos de aprendizado de máquina. O melhor dos resultados exibiu 75 – 95% de acurácia na classificação de genótipos bovinos (RIOS, 2021).

No âmbito da saúde, a espectroscopia associada à aprendizagem de máquina tem sido eficiente na discriminação entre tumores de mama malignos e benignos. A melhor métrica de desempenho entre os modelos de rede neural

testados foi 90,48% ± 10,30% para a métrica da Área sob a Curva (AUC), o que indica um bom desempenho na discriminação entre os tipos de tumores (TOMAS, 2022). No diagnóstico da doença de Parkinson, o algoritmo de aprendizagem usando redes neurais apresentou uma acurácia de 96,29%, o que pode fornecer uma base para o desenvolvimento de um método útil na triagem clínica e na detecção rápida da doença de Parkinson (AHMED, 2010).

Na indústria alimentícia, ela tem mostrado grande potencial para o monitoramento do processo de deterioração da carne e predição de carga microbiana. O classificador utilizado neste estudo foi uma rede neural perceptron de múltiplas camadas, que foi capaz de classificar corretamente 22 das 24 amostras frescas (91,7%), 32 das 34 amostras deterioradas (94,1%) e 13 das 16 amostras semi-frescas (81,2%). Nenhuma amostra fresca foi classificada erroneamente como deteriorada e vice-versa (ARGYRI, 2010). A espectroscopia também se mostrou eficaz na avaliação da qualidade de vinho. Os valores previstos pela rede neural alimentada com os dados espectrais mostraram boa correlação com os valores medidos experimentalmente para a concentração de ácido acético, teor de álcool, fenóis totais e acidez total, cuja acurácia média foi de 89,8% - 94,2% (AGATONOVIC-KUSTRIN, 2013).

Estudos mais recentes avaliaram o impacto do lixo plástico no meio ambiente utilizando uma combinação de espectroscopia de fotoluminescência e aprendizado de máquina. Os resultados revelam que a maioria dos modelos alcançaram uma acurácia maior do que 95%. Em particular, os modelos combinados com Dissecção de Sinal por Maximização de Correlação obtiveram o melhor desempenho (LOTTER, 2022). Outro, propôs uma nova abordagem para o diagnóstico precoce de vários tipos de câncer, utilizando espectroscopia SERS e inteligência artificial. Nesse trabalho o método SERS-AICS distinguiu com sucesso cânceres de controle saudáveis com alta acurácia geral (95,81%), sensibilidade (95,40%), e especificidade (95,87%) (SHI, 2023).

Thomsen et al. utilizaram espectroscopia Raman para identificar fenótipos de bactérias e diferenciar as resistentes à meticilina daquelas suscetíveis a ela. Utilizando um modelo de aprendizado de máquina chamado Transformador

Espectral, foi obtido uma acurácia de classificação superior a 96% em um conjunto de dados composto por 15 classes diferentes e uma acurácia de classificação de 95,6% para seis espécies de bactérias MR-MS (THOMSEN, 2022). Missões como Mars 2020 e ExoMars utilizaram os primeiros instrumentos espectroscópicos validados para análise molecular de amostras coletadas no planeta Marte. Nesse caso, dentre as diversas técnicas analisadas, os espectrômetros Raman se mostraram mais eficazes na detecção de compostos orgânicos em amostras geológicas análogas às da Terra (RULL, 2022).

Conclui-se, portanto, que a aplicação conjunta de técnicas de espectroscopia óptica e de aprendizado de máquina abrange uma ampla gama de aplicações científicas e industriais, proporcionando uma ferramenta valiosa para obtenção de informações e condução de análises complexas em diversas áreas do conhecimento.

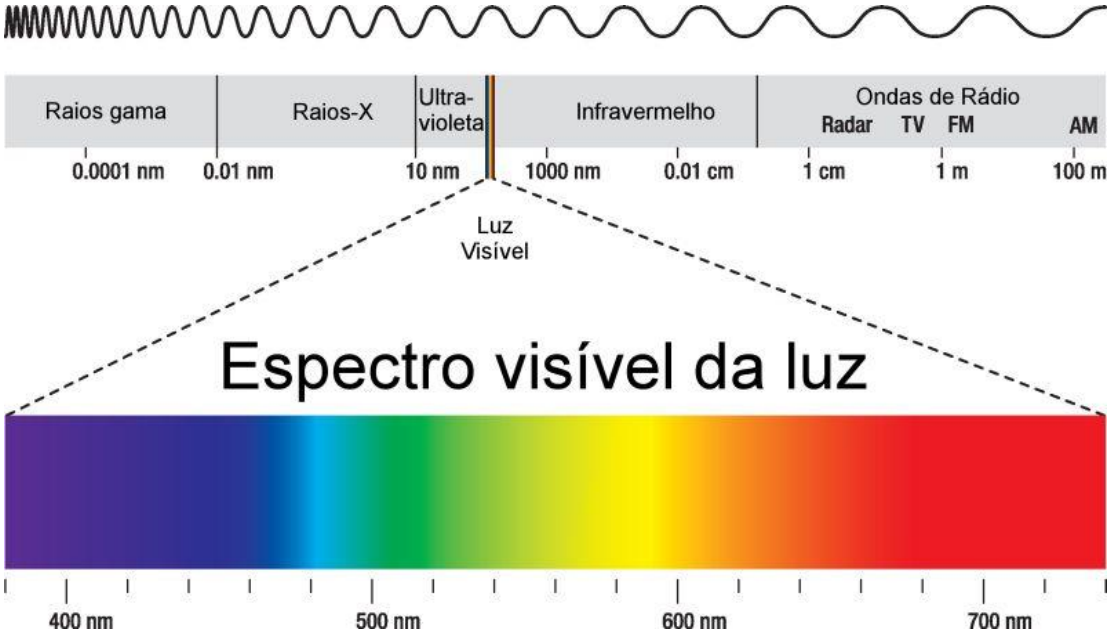
# 3. METODOLOGIA

## *3.1. Espectroscopia no infravermelho por transformada de Fourier*

A espectroscopia no infravermelho por transformada de Fourier é um método de análise cujo princípio de funcionamento é baseado no fato de que ligações e grupamentos químicos vibram em frequências específicas. Sua sensibilidade faz com que o FTIR seja um método eficiente para a determinação da estrutura de moléculas (BARTH, 2007). A energia modulada do infravermelho médio utilizada por esta técnica está na faixa de comprimento de ondas de 4.000 a 400 $cm^{-1}$ (WANG, 2021) e seu uso é justificado por ser um método não invasivo, não destrutivo, rápido e de aplicação universal (PASQUINI, 2003).

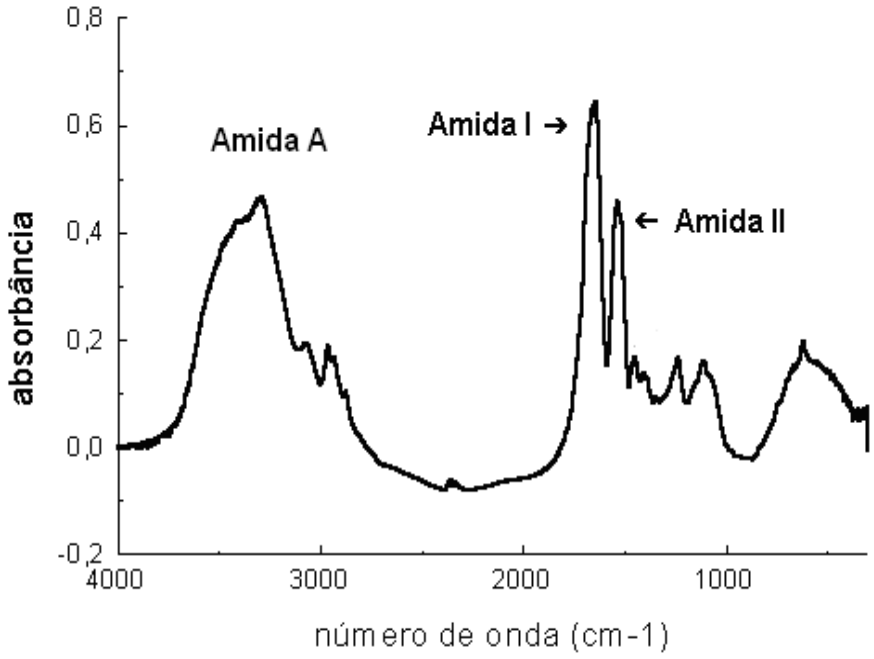
Toda molécula tem níveis de energia discretos associados às suas transições eletrônicas, vibrações e rotações e a absorção da radiação infravermelha excita as transições vibracionais das moléculas. Quando irradiada por luz infravermelha, a substância em análise absorve parte da radiação incidente numa energia específica e sofre uma excitação vibracional do estado fundamental para um estado de energia vibracional superior. O gráfico representa o espectro infravermelho em função do inverso do comprimento de onda, o número de onda, que tem a unidade $cm^{-1}$ (BARTH, 2007; WANG, 2021). A Figura 1 abaixo mostra o espectro eletromagnético, e a Figura 2, um gráfico de FTIR típico.

Figura 1 - Espectro eletromagnético global. A faixa de radiação do infravermelho ocorre na faixa entre o visível e microondas, e a faixa de maior interesse, na região de 4000 a 400 $cm^{-1}$.




Fonte: Referência 22.

Figura 2 - Exemplo de gráfico FTIR. Comportamento da absorbância para a proteína Tripsinogênio em função do número de onda ($cm^{-1}$) (FORATO, 2010).



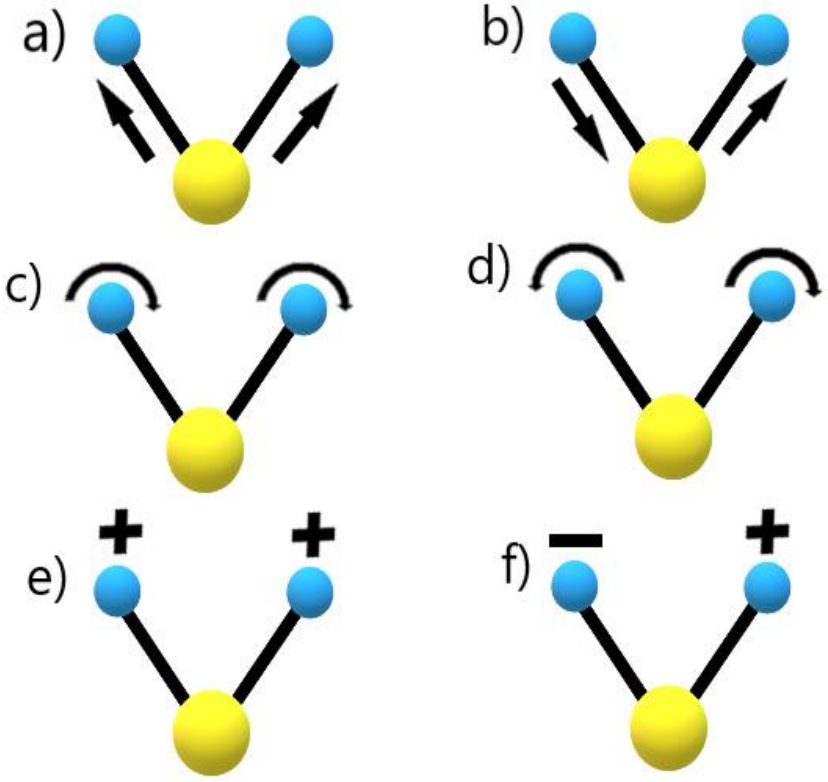

Fonte: Referência 4.

A posição, altura e largura da banda em um gráfico FTIR são determinadas pelas características moleculares e estruturais das substâncias presentes na amostra analisada. A posição é definida pelo tipo de ligação molecular e intermolecular. A altura, pela concentração das moléculas químicas e a intensidade da absorção de energia em uma determinada frequência de vibração. A largura fornece informações sobre a intensidade das interações intermoleculares (WANG, 2021).

A radiação infravermelha interage com a matéria e é absorvida por ela quando o momento de dipolo elétrico permanente das moléculas sofre uma variação durante seu movimento de rotação ou vibração, sendo este um processo quantizado. Somente energias selecionadas de radiação infravermelho são absorvidas pela molécula e é isso o que causa um aumento nos movimentos vibracionais das ligações. Vale ressaltar, no entanto, que somente as ligações que têm um momento de dipolo que muda como função do tempo são capazes de absorver esse tipo de radiação, ao passo que a transferência de energia só ocorre quando um dipolo elétrico muda na mesma frequência da radiação que está sendo introduzida (ESPECTROSCOPIA, s.d.).

Os modos ativos, - ou movimentos de vibração -, no infravermelho que permitem o processo de absorção são os de estiramento e dobramento. Ao sofrer uma alteração no comprimento de suas ligações, a molécula sofre um estiramento, e se a variação ocorre no ângulo entre duas ligações, a molécula sofre um dobramento. O estiramento em fase é chamado de simétrico, e fora, de assimétrico. A deformação, por sua vez, pode ocorrer ou no plano, ou fora dele (KHAN, 2018). Esses fenômenos estão ilustrados na Figura 3.

Figura 3 - Esquema de movimento de moléculas pela interação da radiação no IR. a) estiramento simétrico, b) estiramento assimétrico, c) dobramento simétrico, d) dobramento assimétrico, e) dobramento simétrico fora do plano, f) dobramento assimétrico fora do plano.




Fonte: Autor.

O modelo de um oscilador harmônico simples é a aproximação para baixas energias que melhor descreve as interações da matéria com um feixe infravermelho. Seja uma molécula diatômica heteronuclear. Esse sistema se comporta tal como duas massas m ligadas por um elástico com uma constante de mola de força k e cujo movimento é controlado pela Lei de Hooke (PASQUINI, 2003). Seu Hamiltoniano, nesse caso, é dado pela Equação 1:

$$H = \frac{1}{2m}[p^2 + m^2\omega^2x^2], \qquad [1]$$

onde $\omega = \sqrt{\frac{k}{m}}$ é a frequência natural de vibração do sistema.

É útil reescrever a Equação 1 em termos dos operadores de levantamento e rebaixamento da Mecânica Quântica. O produto desses operadores é fornecido pela Equação 2:

$$a_- a_+ = \frac{1}{2\hbar m\omega}[p^2 + (m\omega x)^2 - im\omega(xp - px)], \quad [2]$$

que, em função do comutador [x,p], se transforma na Equação 3:

$$a_- a_+ = \frac{1}{2\hbar m\omega}[p^2 + (m\omega x)^2] - \frac{i}{2\hbar}[x, p], \quad [3]$$

Com isso, a Equação 1 e 3 podem ser unidas na Equação 4:

$$H = \hbar\omega(a_- a_+ - \frac{1}{2}), \quad [4]$$

Como o Hamiltoniano é igual à energia do sistema, no estado fundamental a Equação de Schrödinger pode ser escrita como:

$$H\psi_0 = E_0\psi_0, \quad [5]$$

Substituindo a Equação 4 na 5 e considerando que o operador rebaixamento não pode ser aplicado ao estado fundamental, - ou seja, $a_-\psi_0 = 0$ -, temos a Equação 6 e a energia do estado fundamental, representada pela Equação 7:

$$\hbar\omega[0 - \frac{1}{2}]\psi_0 = E_0\psi_0, \quad [6]$$

$$E_0 = \frac{1}{2}\hbar\omega, \quad [7]$$

Aplicando o operador de levantamento sucessivamente para gerar os estados excitados a energia aumenta por um fator $\hbar\omega$ a cada passo:

$$E_n = (n + \frac{1}{2})\hbar\omega, \quad [8]$$

Quando uma ligação vibra, a energia de vibração muda de energia cinética para potencial e vice-versa. Essa energia, que corresponde à diferença entre dois níveis energéticos adjacentes, é proporcional à frequência da vibração e é dada pela Equação 9:

$$\triangle E = \hbar\omega, \quad [9]$$

que, para um oscilador harmônico, é determinada pela constante K do elástico e pelas massas dos dois átomos unidos. A frequência natural de vibração de uma ligação é dada pela Equação 10:

$$\upsilon = \frac{1}{2\pi c}\sqrt{\frac{k}{\mu}}, \qquad [10]$$

com K sendo uma constante que varia de uma ligação para outra e $\mu$ sendo a massa reduzida do sistema, fornecida pela Equação 11:

$$\mu = \frac{m_1 m_2}{m_1 + m_2}, \qquad [11]$$

A Equação 10 é usada para calcular a posição aproximada de uma banda no espectro infravermelho. Valores experimentais e calculados, no entanto, dependem também de fatores como a ressonância e a hibridização que operam em moléculas orgânicas (ESPECTROSCOPIA, s.d.).

Além disso, se as forças de restauração atuam ao longo das ligações de valências e a lei da elasticidade de Hooke é obedecida para pequenos deslocamentos, a Equação 12 é uma relação quantitativa entre a energia da vibração e as massas dos átomos das partículas diretamente envolvidas na oscilação (ESPECTROSCOPIA Molecular, s.d.).

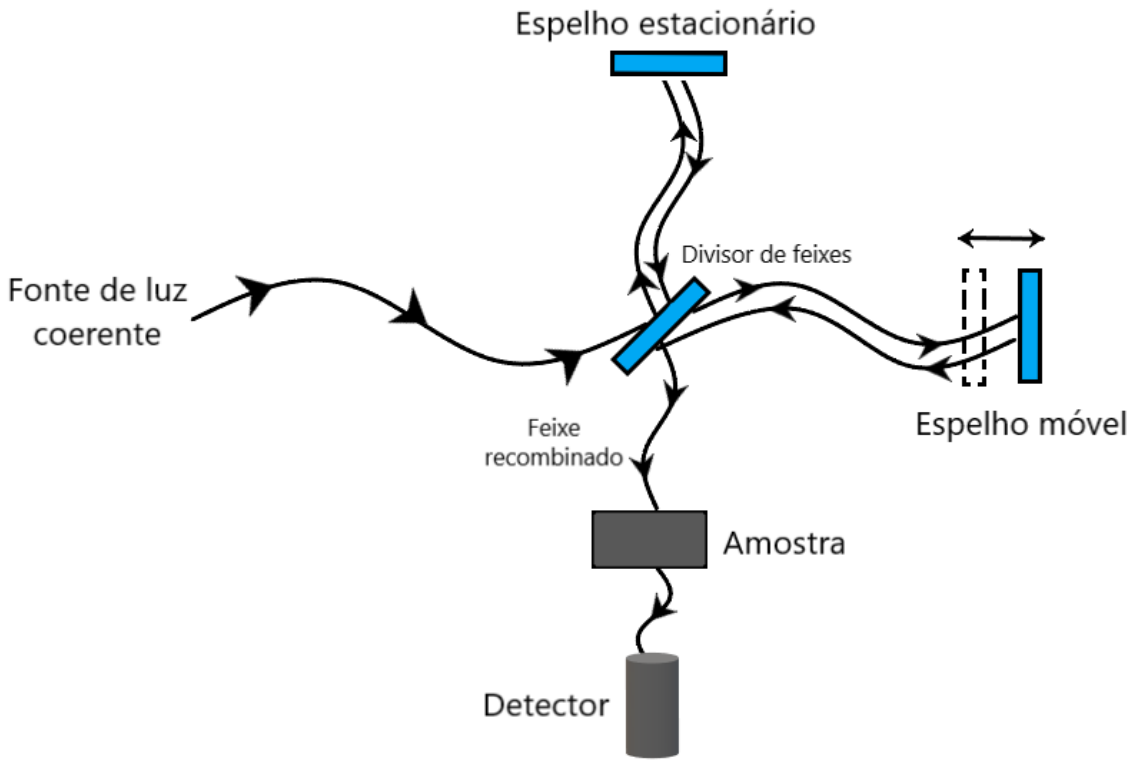
$$\nu = \frac{\pi}{2}\left(\frac{k}{\mu}\right)^{1/2}, \qquad [12]$$

onde $\nu$ é a frequência da oscilação, $k$, a constante de força, e $\mu$, a massa reduzida dada pela Equação 11.

### 3.1.1. Instrumentação

Dos espectrômetros de infravermelho, - dispersivos e de transformada de Fourier (TF) -, o de TF é o mais rápido. Nesses equipamentos, o caminho óptico produz o interferograma, que é um gráfico de intensidade em função do tempo. No entanto, o espectrômetro TF utiliza tratamento matemáticos, - a saber, a Transformada de Fourier -, para separar as frequências das absorções individuais. Esse processo produz um espectro virtualmente idêntico ao obtido com um espectrômetro dispersivo e em um tempo bem menor, pois é capaz de gerar um interferograma em menos de um segundo (ESPECTROSCOPIA, s.d.). A Figura 4 é um diagrama que mostra como esse equipamento funciona.

Figura 4 - Diagrama esquemático de um espectrofotômetro de transformada de Fourier.




Fonte: Autor. Adaptado da Referência 26.

No interferômetro de Michelson da Figura 4 podemos observar que a energia da fonte de alimentação atravessa um divisor de feixes (espelho semi-prateado) que a separa em dois feixes perpendiculares entre si que percorrem dois caminhos distintos. Um deles vai para o espelho fixo e é, então, refletido novamente para o divisor, e o outro, para o espelho móvel, sendo igualmente refletido de volta para o divisor. Ao se encontrarem, os feixes se recombinam com um perfil de batimento e as diferenças de caminho causam as interferências destrutivas e construtivas que formam o interferograma. Esse novo feixe é o que atravessa a amostra em análise.

Uma vez incidido sobre a amostra, o feixe é direcionado para um detector fotossensível responsável por captar o estímulo luminoso e um interferograma final, que contém toda a informação de um sinal de domínio temporal, é gerado. Este interferograma, por fim, passa pela Transformada de Fourier mediante análise computacional, que extrai as frequências individuais absorvidas e as converte em um espectro de absorção óptica infravermelha, para, assim, nos dar as informações relativas à intensidade de absorção ou transmissão em função do número de onda (relacionado à energia vibracional) (FIGUEIREDO, 2009). Essas variáveis são as que caracterizam o desenho típico da Figura 2.

A Transformada de Fourier, - utilizada nesse processo final de conversão do interferograma -, pode ser expressa como um caso limite da série de Fourier. A série de Fourier para uma função $f(x)$ na forma exponencial é dada pela Equação 13, com o coeficiente sendo definido pela Equação 14:

$$f(x) = \sum_{-\infty}^{\infty}[c_n . e^{in\omega_0 x}], \qquad [13]$$

na qual,

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(x)e^{-in\omega_0 x}dx, \qquad [14]$$

onde f(x) é a função que estamos tentando representar, $c_n$ são os coeficientes da série, n é o número de termos na série, T é o período da função e $\omega_0$ é a frequência angular.

Ao fazer T tender ao infinito, diminuindo, desta maneira, a distância entre duas frequências harmônicas, transformamos uma função periódica em aperiódica. Isso nos dá a Transformada de Fourier expressa pelas Equações 15 e 16:

$$h(\nu) = \int_{-\infty}^{\infty} g(\delta) e^{2\pi i \nu \delta} d\delta, \quad [15]$$

$$I(\delta) = \int_{0}^{\infty} h(\nu) e^{-2\pi i \nu \delta} d\nu, \quad [16]$$

Como na espectroscopia $I(\delta)$ é a potência espectral de um número de onda $\nu$, - o qual é dado pela função $h(\nu)$ -, a Equação 16 representa a variação da potência espectral, e a Equação 15, a variação da intensidade em função do número de onda (BATES, 1976).

## 3.2. Aprendizagem de Máquina

Uma máquina, - ou computador -, é capaz de “aprender” quando sua estrutura de programação é alterada a fim de se obter uma melhora no seu desempenho, seja para fins de reconhecimento de padrões, diagnósticos, predições, etc. Dentro desse contexto, a Aprendizagem de Máquina é um termo que está associado às mudanças feitas em um sistema que visa solucionar problemas. Alguns dos motivos que justificam seu uso são: capacidade de lidar com grandes volumes de informações, automação de tarefas, descoberta de padrões, aprendizado contínuo, etc. (NILSSON, 1996).

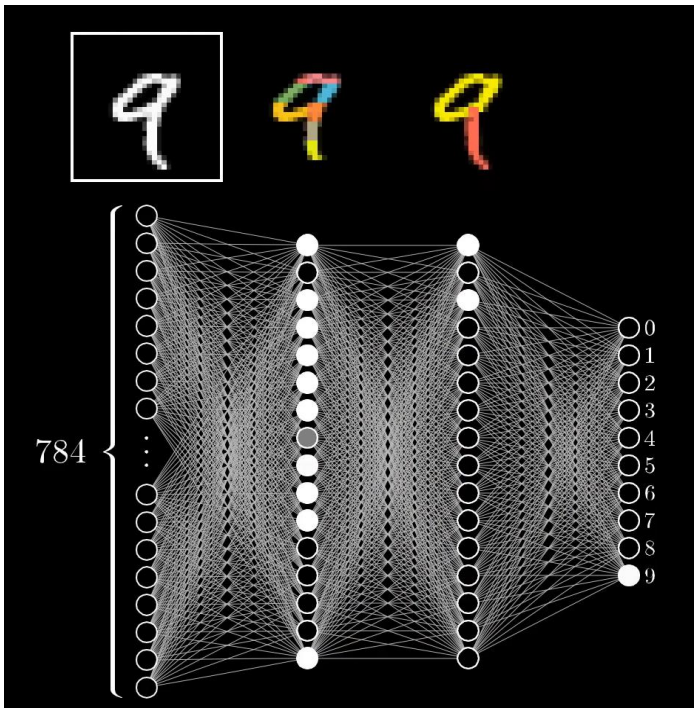
Quase todas as tarefas que podem ser concluídas com um padrão ou conjunto de regras definido por dados podem ser automatizadas. O aprendizado de máquina pode ser supervisionado ou não supervisionado. O método de treinamento supervisionado permite o treino de um modelo usando um conjunto de dados rotulados, cujas respostas são conhecidas. O não supervisionado, por sua vez, descobre padrões desconhecidos nos dados, tendo em vista que o

treinamento é realizado com um conjunto de dados que contém exemplos não rotulados (HOW... 2020).

### 3.2.1. Redes Neurais

Redes Neurais (RN) têm como inspiração biológica o cérebro humano e são um tipo de Aprendizado de Máquina. Formadas por uma rede adaptativa de elementos não-lineares, as Redes Neurais ensinam os computadores a processar e manipular dados para resolver problemas e reconhecer padrões. Sua estrutura é composta por neurônios artificiais interconectados em camadas (NILSSON, 1996). A Figura 5 representa uma RN simples.

Figura 5 - Esquema de neurônios compondo uma rede neural unidirecional (feedfoward) utilizada na identificação de uma imagem cuja resolução é 28 x 28 pixels.




Fonte: Referência 30.

Na RN, os neurônios artificiais são as unidades de processamento que trabalham em conjunto para resolver um determinado problema matemático. Esses neurônios possuem um viés (ou tendência para inatividade) e ligam-se uns aos outros por canais que têm um peso determinado (LEON, s.d.). O que o computador fará é trabalhar com os pesos e tendências a fim de encontrar uma solução para a problemática proposta.

Primeiramente, os sinais, - características dos dados iniciais -, são apresentados à camada de entrada da rede neural e são multiplicados por um peso sináptico. Depois, a soma ponderada dos sinais de entrada multiplicados pelos pesos é calculada, - o que representa a ativação do neurônio -, e um termo de viés é adicionado à soma. Esse termo, que é responsável por controlar o nível de atividade da unidade, é, então, aplicado em uma função de ativação (curva logística, sigmóide). Se o resultado da função de ativação exceder um determinado limite, a unidade produzirá uma resposta de saída específica. O processo é repetido para todas as unidades nas camadas intermediárias e na camada de saída da rede neural (LEON, s.d.).

Considerando as informações de entrada $a_0^{(0)}$, $a_1^{(0)}$, … , $a_n^{(0)}$ , - cada qual com seus respectivos pesos, $W_{0,0}$, $W_{0,1}$, … , $W_{0,n}$ -, tendências $b_0$, $b_1$, … , $b_n$ , limitador t e função de ativação sigmóide $\sigma$, o nível de atividade dos neurônios da camada posterior, $a^{(1)}$, é dado pela Equação 17, que está em notação matricial:

$$a^{(1)} = \sigma(Wa^{(0)} + b), \qquad [17]$$

Com base no conjunto de dados de entrada e suas saídas correspondentes, o objetivo do processo de aprendizagem é minimizar o erro entre as saídas produzidas e as desejadas. A regra de treinamento, - como o Gradiente Descendente, que ajusta o peso em direção ao mínimo local mais próximo da função custo -, arruma os pesos das conexões de forma iterativa, atualizando-os em um processo que é repetido para cada exemplo de

treinamento. Isso faz com que a rede aprenda a mapear os padrões de entrada para as saídas desejadas (GALUSHKIN, 2007).

### 3.2.2. Técnicas de Normalização

Dados iniciais dos espectros FTIR são geralmente influenciados por variações sistemáticas não relacionadas às propriedades químicas do sistema, - tais como espalhamento, tamanho da partícula e diferenças nas intensidades globais dos sinais, flutuações eletrônicas (HUANG, 2010). O pré-processamento de espectros envolve eliminar ou minimizar a variabilidade não relacionada à variável-alvo. Dois métodos muito utilizados no contexto da espectroscopia para o tratamento do conjunto de dados espectral são a Variável Normal Padrão (VNP) e derivadas de primeira e segunda ordem (HUANG, 2010).

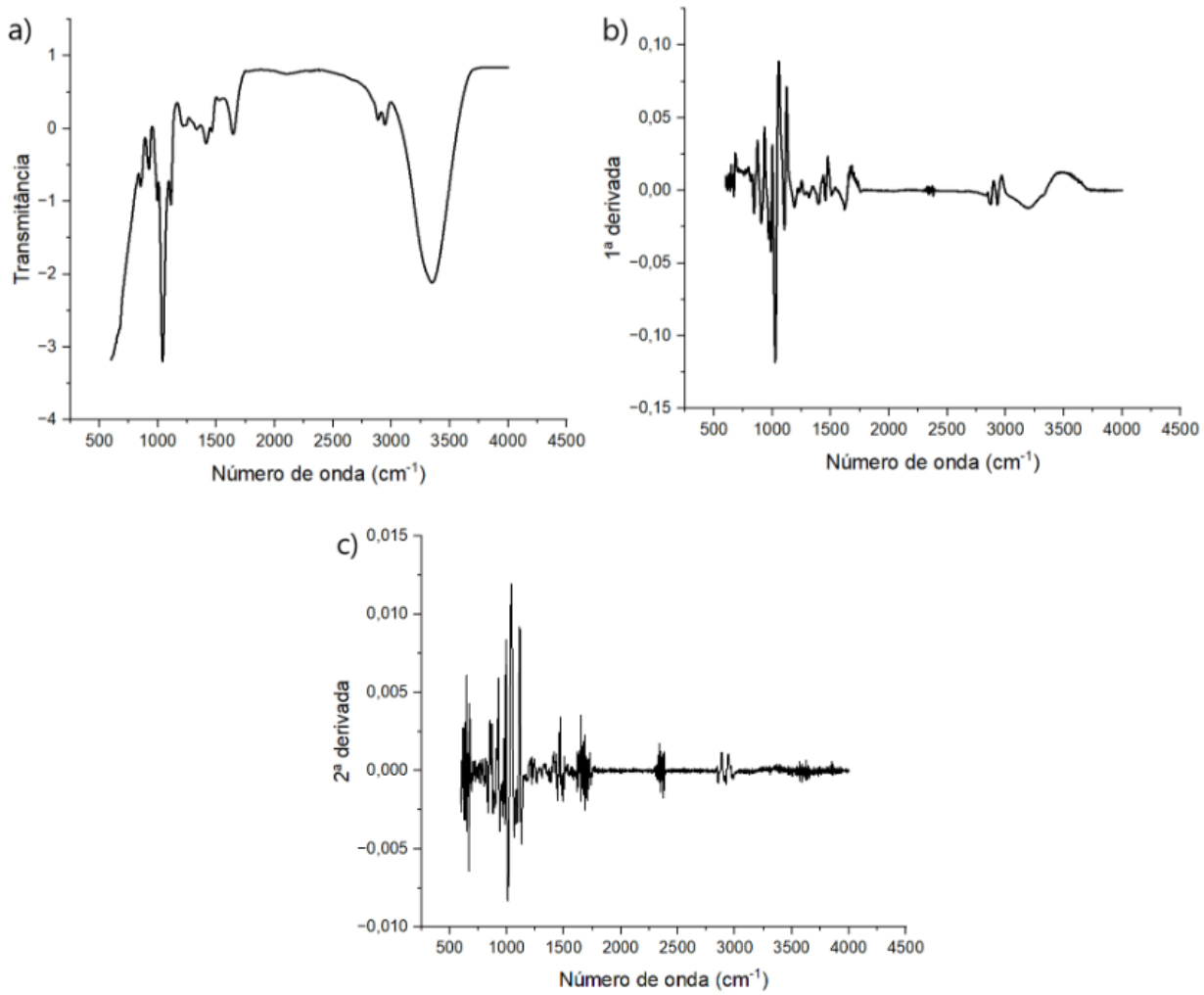
Por meio de um processo de normalização, a VNP torna todos os espectros comparáveis em termos de intensidade. O seu cálculo consiste em subtrair cada espectro por sua própria média e dividi-lo por seu desvio padrão. O resultado disso é que todos os espectros têm uma média de zero e um desvio padrão unitário (FERRÉ, 2009). A VNP é dada pela Equação 18:

$$x_{i,j}^{SNV} = \frac{(x_{i,j} - \overline{x_i})}{\sqrt{\frac{\sum_{j=1}^{p}(x_{i,j} - \overline{x_i})^2}{p-1}}}, \qquad [18]$$

onde $x_{i,j}^{SNV}$ é o elemento do espectro transformado, $x_{i,j}$, o elemento correspondente original do espectro i na variável j, $\overline{x_i}$, o espectro médio i, e $p$, o número de variáveis ou comprimento de onda no espectro.

A derivada representa matematicamente a inclinação da reta tangente de uma função em um ponto qualquer. Na espectroscopia, a primeira derivada é útil por remover as linhas de base de um conjunto de dados. Contudo, a segunda derivada é preferível porque seu gráfico melhora a compreensão e interpretação das informações espectrais por realçar os pontos de inflexão, assim como também elimina aumentos lineares de linha de base que não são tratados pela primeira derivada (COLUMN, 2007). A Figura 6 representa a primeira e segunda derivadas aplicadas a um gráfico de FTIR.

Figura 6 - Comportamento da primeira (b) e segunda (c) derivadas do espectro médio de amostras homozigóticas (a) em função do número de onda ($\mathrm{cm}^{-1}$).



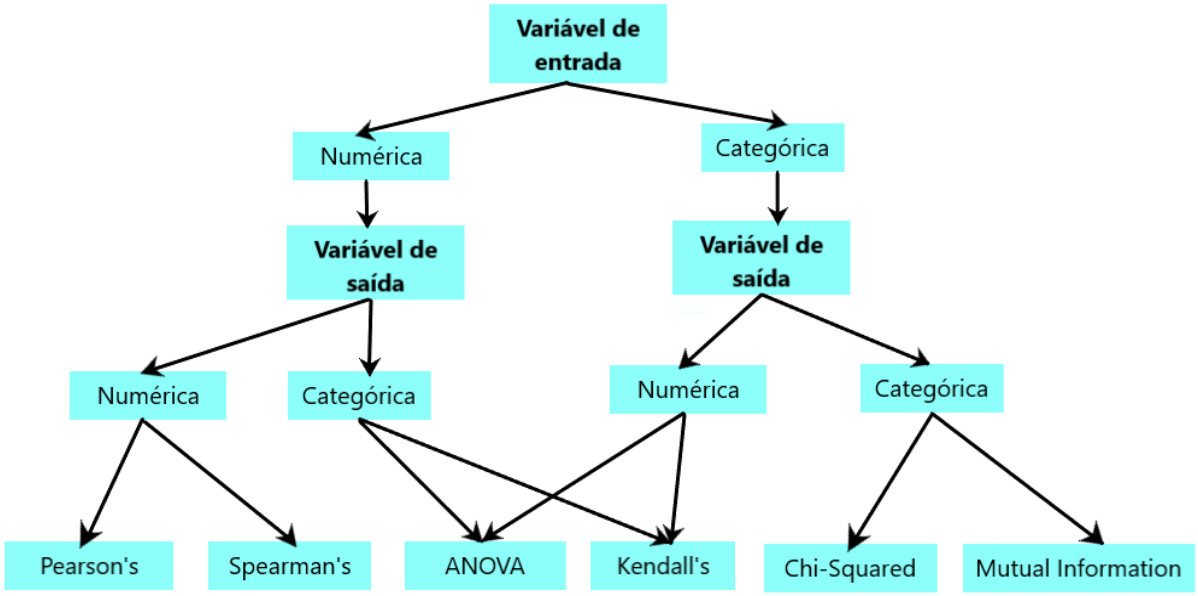

Fonte: Autor.

### 3.2.3. Seleção de Variáveis

Nem sempre todas as variáveis relacionadas ao espectro de um objeto de estudo são úteis para a resolução do problema. Muitas variáveis tornam o modelo lento, e o treinamento feito com base em informações irrelevantes pode produzir resultados imprecisos e indesejados. A Seleção de Variáveis reduz as variáveis de entrada do modelo de aprendizagem de máquina por meio da separação de dados relevantes de não relevantes (BROWNLEE, 2020). Alguns dos motivos que justificam o seu uso são: evitar com que o modelo aprenda com informações irrelevantes (overfitting), aumento da precisão, redução do tempo de treinamento e criação de um modelo mais fácil de entender (HUANG, 2015). Em espectroscopia FTIR, as variáveis são relacionadas às transições moleculares presentes no espectro.

A Seleção de Variáveis pode ser supervisionada ou não supervisionada. A diferença entre elas está relacionada à seleção com base na presença ou ausência de uma variável resposta. Se o alvo é ignorado, a técnica é não supervisionada, e se é considerado, a técnica é supervisionada. Os métodos de seleção supervisionada mais comuns são o Filter, Wrapper e Intrinsic (KUHN, 2013). Adicionalmente, a depender do tipo de variável de entrada e saída, - numérica ou categórica -, podemos escolher nosso modelo da forma que mostra a Figura 7.

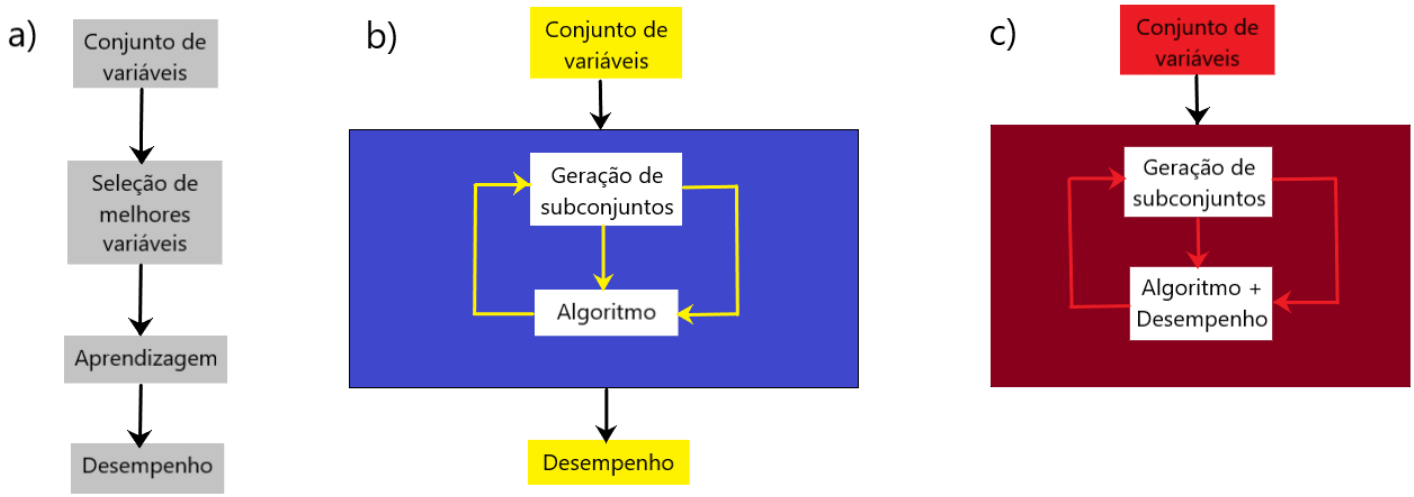
Figura 7 - Esquema de seleção de variáveis segundo o tipo de entrada e saída.




Fonte: Autor. Adaptado da Referência 36.

No caso da seleção supervisionada, o Filter elimina variáveis de acordo com a sua correlação positiva ou negativa com os rótulos de saída. O Wrapper divide os dados em subconjuntos, treina o algoritmo com base neles, usa o resultado para decidir quais variáveis serão descartadas e, então, treina o modelo novamente. O Intrinsic une o Filter e o Wrapper para criar vários subconjuntos e encontrar o melhor dentre eles (MENON, 2023). A Figura 8 representa um fluxograma para cada um desses processos descritos.

Figura 8 - Fluxograma dos métodos de seleção de variáveis supervisionados Filter (a), Wrapper (b) e Intrinsic (c).




Fonte: Autor. Adaptado da Referência 39.

Outro algoritmo utilizado para selecionar variáveis é o Boruta, que usa o método wrapper baseado em um classificador por Random Forest. O Boruta compara o desempenho de variáveis reais com o de variáveis aleatórias, -sombras, geradas a partir dos dados originais -, e remove iterativamente aquelas que são comprovadas por um teste estatístico como menos relevantes (KURSA, 2010). O algoritmo de Boruta funciona segundo esses passos (KURSA, 2010):

1. Criar cópias das variáveis originais (atributos de sombra) e embaralhar as mesmas;
2. Treinar o classificador Random Forest no conjunto de dados estendido e calcular os Valores-Z. Eles indicam o quão longe uma variável está da média;
3. Encontrar o Valor-Z máximo (MZSA) entre os atributos de sombra e atribuir relevância aos atributos que pontuaram melhor que o MZSA;
4. Realização do Teste Bilateral de Igualdade para cada atributo com importância indeterminada;

5. Descartar os atributos com pontuação significativamente inferior ao MZSA (‘sem importância’) e considerar os atributos com pontuação significativamente maior ao MZSA ('importante');
6. Remover os atributos de sombra;
7. Repetir o processo até que a importância seja atribuída para todos os atributos, ou até que o algoritmo atinja o limite previamente definido de execuções.

### 3.2.4. Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA) é um tratamento estatístico que visa encontrar padrões entre múltiplas variáveis mediante a redução da dimensionalidade de um conjunto de dados inicial, sem excluir informações relevantes. Para isso, o algoritmo do PCA gera um conjunto novo de variáveis conhecidas como componentes principais, mais conhecidas como PC`s, proveniente do acrônimo em inglês de Principal Components. Essas novas variáveis são não correlacionadas e são ordenadas de forma que as primeiras componentes principais capturem a maior parte da variância presente em todas as variáveis originais (JOLLIFE, 2016).

Seja um conjunto de dados com observações sobre p variáveis numéricas para cada uma das n entidades presentes no sistema. Isso define a matriz de dados X com dimensões n x p. Desejamos obter a máxima variância para a combinação linear das colunas da matriz X. A Equação 19 nos dá a combinação linear:

$$\sum_{j=1}^{m} a_j x_j = Xa, \qquad [19]$$

com $a$ sendo um vetor de constantes $a_1$, $a_2$,…, $a_n$.

A variância de uma combinação linear é dada pela Equação 20:

$$var(Xa) = a'Sa, \qquad [20]$$

com S sendo a matriz de covariância dos dados originais, e $a'$, a transposta de $a$.

Para maximizar a variância expressa pela Equação 20 é preciso encontrar um vetor $a$ tal que a forma quadrática de $a'Sa$ seja o máximo possível. Uma restrição é necessária para encontrar a solução desse problema, no entanto: devemos trabalhar com vetores unitários ($a'a = 1$). Logo, precisamos maximizar a Equação 21:

$$a'Sa - \lambda(a'a - 1), \qquad [21]$$

onde $\lambda$ é um multiplicador de Lagrange.

Derivando a Equação 21 em relação ao vetor $a$ e igualando-a ao vetor nulo, obtemos a Equação 22:

$$Sa - \lambda a = 0 \Leftrightarrow Sa = \lambda a, \qquad [22]$$

Da Equação 22 vemos que $a$ deve ser um autovetor, e $\lambda$, um autovalor relacionado à matriz $S$. Como queremos o maior autovalor, $\lambda_1$, e seu correspondente autovetor, $a_1$, substituímos a Equação 22 na Equação 20:

$$var(Xa) = a'Sa = \lambda a'a = \lambda, \qquad [23]$$

Qualquer matriz simétrica real p × p tem p autovalores reais, $\lambda_k$ ($k = 1, \ldots, p$), e seus autovetores formam um conjunto de vetores tal que $a'_k a_{k'} = 1$ se $k = k'$ $ou$ $a'_k a_{k'} = 0$ se $k \neq k'$. Portanto, o conjunto completo de autovetores dessa matriz são as soluções que buscamos. A nova combinação linear pode ser escrita como na Equação 24:

$$\sum_{j=1}^{p} a_{jk} x_j = Xa_k \ , \qquad [24]$$

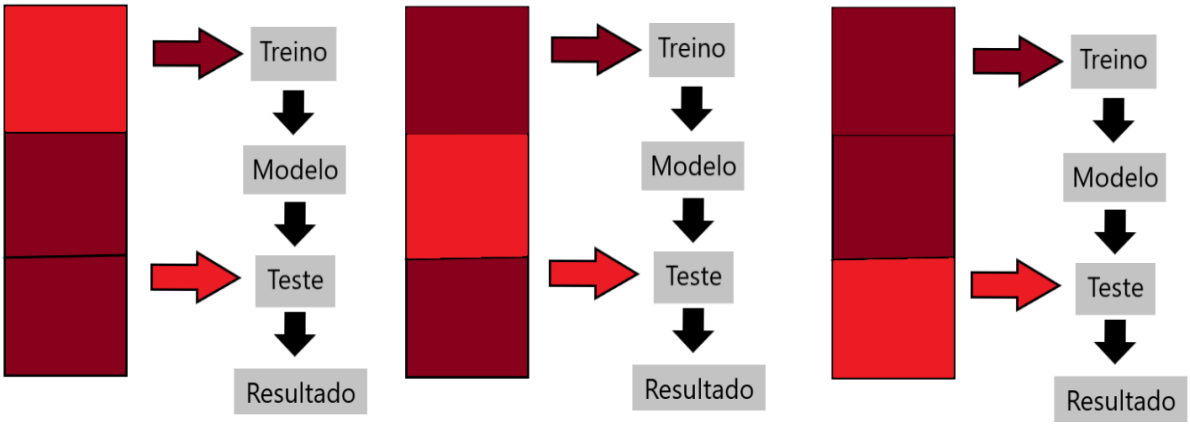
no qual $Xa_k$ são os PCs do conjunto de dados, $a_k$, os loadings, e $Xa_k$, os scores (JOLLIFE, 2016).

### 3.2.4. Validação cruzada

A validação cruzada (VC) é uma técnica utilizada em aprendizado de máquina que divide os dados para estimar o risco de cada algoritmo e avaliar classificadores. Parte dos dados (treino) é usada para treinar o algoritmo, e a parte restante (teste), para estimar o seu risco. Depois, a VC seleciona o algoritmo com o menor risco estimado. Como a amostra de treinamento é independente da amostra de validação, a validação cruzada evita o overfitting, além de ser simples e de aplicação praticamente universal (ARLOT, 2010).

A Figura 9 ilustra o modelo K-fold cross-validation de validação cruzada. Nesse procedimento, os dados são inicialmente divididos em "$k$" agrupamentos. Após isso, k iterações subsequentes de treinamento e validação são realizadas de forma que dentro de cada iteração um agrupamento diferente dos dados é desconsiderado para validação, enquanto os outros k - 1 agrupamentos restantes são usados para aprendizado (REFAEILZADEH, 2020).

Figura 9 - Procedimento de validação cruzada tripla com o método K-fold.




Fonte: Autor. Adaptado da Referência 43.

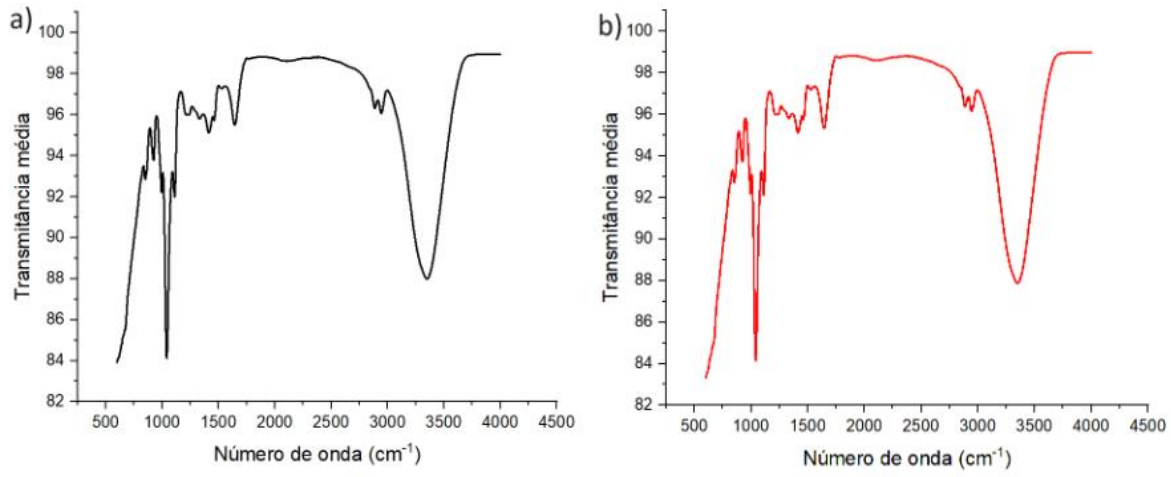
Há também um caso especial de validação cruzada, a validação cruzada leave-one-out, - LOO-CV, do acrônimo em inglês de leave-one-out cross-validation -, em que a cada iteração praticamente todos os dados, - exceto uma única amostra -, são usados para treinamento. O modelo é, então, testado nessa única amostra deixada de fora. O LOO-CV é um tipo de validação cruzada muito conhecido por ser quase imparcial, mas tem uma alta variância associada (ARLOT, 2010).

# 4. RESULTADOS E DISCUSSÃO

Utilizando o espectrômetro FTIR (Spectrum 100, Perkin Elmer) com resolução 4 $cm^{-1}$ e acessório de refletância total atenuada (RTA), - disponível no Laboratório de Óptica e Fotônica do INFI-UFMS -, RIOS et. al. obtiveram os espectros correspondentes aos polimorfismos de nucleotídeo único (SNP) no DNA de bovinos na forma heterozigótica (CD) e homozigótica (CC). Um total de 200 amostras foram coletadas, - 100 amostras heterozigóticas “CD” e 100 amostras homozigóticas “CC”. O objetivo foi empregá-las na construção de um modelo de classificação capaz de identificar novas amostras.

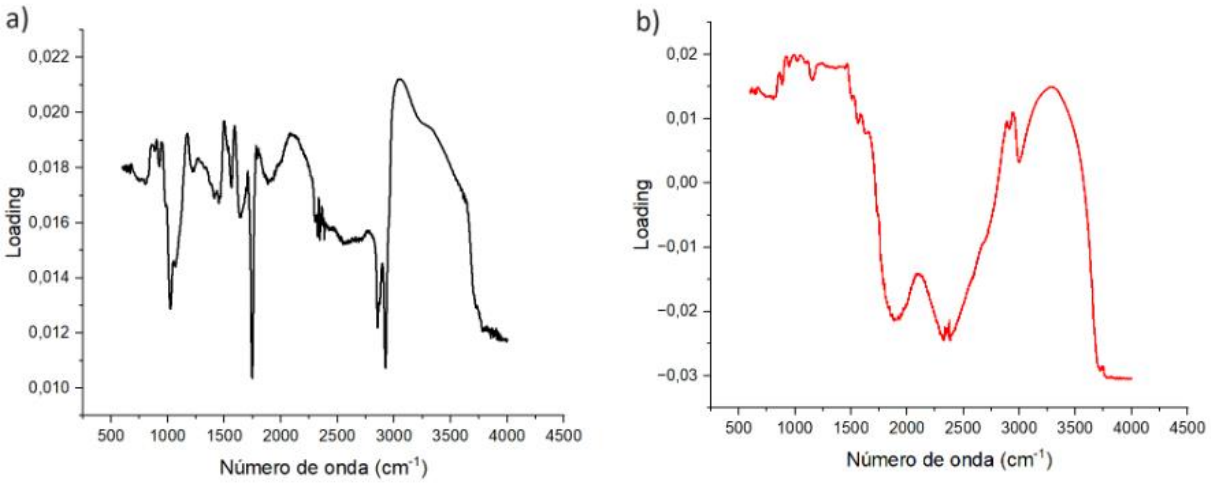
Os espectros FTIR de DNA bovino foram normalizados utilizando a Variável Normal Padrão e derivadas e em seguida foram tratados por uma Análise de Componentes Principais a fim de reduzir a sua dimensionalidade para duas componentes principais, PCs, que não são correlacionadas entre si. Com isso, construímos os gráficos das Figuras 10, 11 e 12.

Figura 10 - Comportamento da transmitância média, normalizada, para as amostras de DNA bovino CC (a) e CD (b) em função do número de onda ($cm^{-1}$).



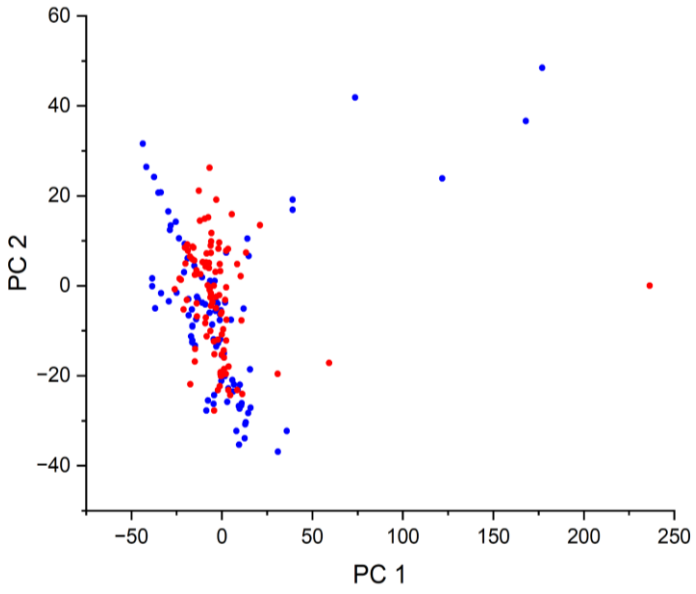

Fonte: Autor.

Figura 11 - Comportamento do loading para o PC1 (a) e PC2 (b) em função do número de onda ($cm^{-1}$).




Fonte: Autor.

Figura 12 - Comportamento do PC-2 em função do PC-1 para as classes CC (azul) e CD (vermelho).




Fonte: Autor.

A Figura 10 nos mostra que as amostras de DNA bovino possuem espectros FTIR semelhantes entre si. Observamos que as principais bandas

vibracionais estão em três regiões: 800–1250 $cm^{-1}$, 1250–1500 $cm^{-1}$ e 1500–1800 $cm^{-1}$. A região entre 800–1250 $cm^{-1}$ associa-se ao $PO_4^{-2}$ e o estiramento da desoxirribose, a que está entre 1250-1500 $cm^{-1}$, aos modos vibracionais de anéis aromáticos na base nitrogenada, - pirimidina e purina -, e a banda de absorção entre 1500–1800 $cm^{-1}$, aos modos vibracionais de C = N, C = O e C = C (RIOS, 2021).

Os gráficos de loadings (Figura 11) mostram a relação entre as variáveis originais e os componentes principais (PCs). Já o gráfico dos scores dos primeiros PCs (Figura 12) revela a tendência de agrupamento de amostras semelhantes, possibilitando a avaliação do potencial preditivo para a classificação de futuras amostras. Após a aplicação do PCA e a análise exploratória proveniente deste, foi realizado o treinamento de algoritmos de aprendizado de máquina supervisionados, com o intuito de extrair padrões relevantes para a classificação de amostras.

Utilizamos a linguagem de programação Python e várias bibliotecas para construção de um modelo de aprendizado de máquina e análise de dados. *pandas* e *numpy* foram usadas para manipulação de dados. *sklearn* para pré-processamento, treinamento de classificadores e cálculo de métricas. *BorutaPy* auxiliou na seleção de recursos, *matplotlib* na visualização de dados e a biblioteca *time* para medição do tempo de execução do código.

Realizamos o treinamento do algoritmo com base em 30 componentes principais. A quantidade de Componentes Principais (PCs) foi selecionada após a construção de uma grade de otimização na etapa de pré-análise. Os dados reduzidos foram divididos em conjuntos de treinamento e teste usando a função 'train_test_split' do *sklearn*, alocando 25% dos dados para teste e 75% para treinamento.

Desenvolvemos um classificador Random Forest com 50 árvores de decisão como estimadores, o que foi determinado por meio de um processo de pesquisa em grade (GridSearchCV) da biblioteca *sklearn*. A seleção das componentes principais a serem utilizadas no modelo de classificação foi

realizada pelo *BorutaPy*, que identificou os PCs 1, 2, 4, 6, 8, 9, 11, 13, 14, 15, 16, 18, 19, 27 e 29 como os mais relevantes. A redução da dimensionalidade dos dados foi realizada com o objetivo de aprimorar o desempenho do modelo de aprendizado, reduzindo o overfitting e acelerando o treinamento.

Para diferenciar as classes CC e CD, o *sklearn* do Python foi utilizado para treinar uma rede neural usando o algoritmo de classificação perceptron multicamadas. A função LeaveOneOut do módulo *model_selection* foi usada para realizar a validação cruzada Leave-One-Out (LOO-CV). Inicialmente, os hiperparâmetros padrões foram empregados, que foram obtidos durante o processo de otimização por GridSearch, também do módulo *model_selection*. Esses hiperparâmetros incluem ‘solver’ como ‘sgd’, ‘learning_rate_init’ como 0.1, ‘activation’ como ‘logistic’, ‘max_iter’ como 1500 e ‘random_state’ como 1.

A seguir, realizamos uma otimização de hiperparâmetros para determinar a configuração ideal de neurônios e camadas da rede neural no cenário em estudo. O código utilizou um loop *for* e a função *product* do módulo *itertools* para criar uma lista de todas as combinações possíveis dos parâmetros de camadas e neurônios, cujo tempo de processamento foi de 8 horas.

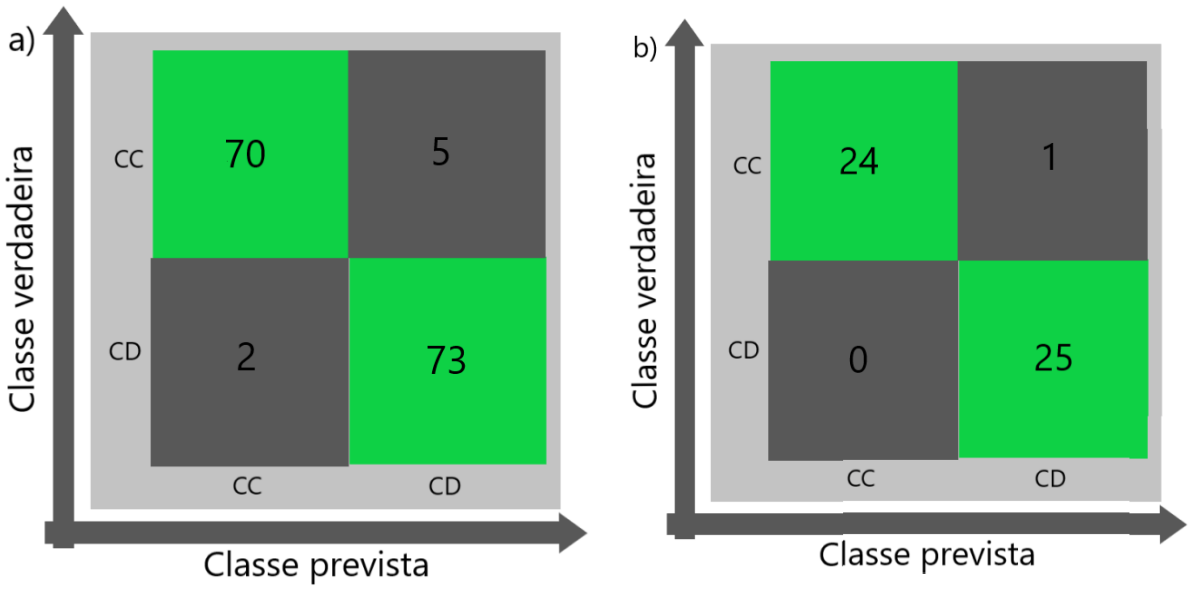
Esta otimização resultou na escolha de uma rede neural com 2 camadas: a primeira camada com 21 neurônios e a segunda com 11 neurônios. Essa configuração alcançou uma acurácia interna de 95,33% e externa de 98,00%. Isso mostra que o algoritmo pôde se ajustar aos dados de treinamento usados para treiná-lo e conseguiu prever bem resultados em dados desconhecidos.

Para a validação interna, 70 amostras da classe CC foram corretamente classificadas como CC, 73 amostras da classe CD foram corretamente classificadas como CD, 2 amostras da classe CD foram incorretamente classificadas como CC e 5 amostras da classe CC foram incorretamente classificadas como CD. Nesse cenário, o modelo obteve um bom desempenho e demonstrou a capacidade de distinguir efetivamente entre as duas classes. Ele cometeu apenas 7 erros, classificando erroneamente 2 amostras da classe CD como CC e 5 amostras da classe CC como CD.

Para a validação externa, por sua vez, 24 amostras da classe CC foram corretamente classificadas como CC, 25 amostras da classe CD foram corretamente classificadas como CD, 1 amostra da classe CD foi incorretamente classificada como CC e nenhuma amostra da classe CC foi incorretamente classificada como CD. Como apenas uma amostra da classe CD foi incorretamente classificada como CC, o modelo possui alta capacidade de generalização para novos dados.

A Figura 13 apresenta duas matrizes de confusão, fornecendo uma representação visual do desempenho desse modelo de classificação obtido a partir da validação cruzada Leave-One-Out e validação externa.

Figura 13 - Matriz de confusão para a validação interna (a) e externa (b).




Fonte: Autor.

No estudo de RIOS et. al., - no qual foram empregados algoritmos como análise discriminante linear e quadrática, máquina de vetores de suporte (MVS) linear, quadrática e cúbica, e K – vizinhos mais próximos -, foi obtida uma acurácia geral de 75% para os primeiros 9 PCs e MVS quadrático na faixa

espectral de 1800 – 800 $cm^{-1}$. Estratégias adicionais resultaram em maior acurácia, atingindo 95% na classificação CD x DD com os primeiros 22 PCs e MVS quadrático.

Os testes envolvendo CC x CD forneceram a segunda maior acurácia (90%) usando os primeiros 11 PCs e MVS linear, o que é menor do que o resultado obtido no presente trabalho usando 15 PCs selecionadas pelo Boruta. Ambos utilizaram FTIR, análise multivariada e aprendizado de máquina com validação cruzada para identificar polimorfismos de DNA bovino. Porém, RIOS et. al. não utilizaram validação externa.

# 5. CONCLUSÃO

A espectroscopia no infravermelho por transformada de Fourier em conjunto com análise multivariada e aprendizado de máquina demonstrou ser uma ferramenta eficaz para detectar e classificar amostras de DNA bovinos. A Variável Normal Padrão corrigiu o espalhamento multiplicativo nos dados, enquanto as derivadas primeira e segunda ajudaram a eliminar o efeito de fundo e a realçar as mudanças na curvatura do espectro. Além disso, a seleção de variáveis com o Boruta capturou todos os recursos importantes com relação às variáveis resultado, melhorando a eficácia e precisão do modelo.

Os resultados obtidos neste estudo para acurácia interna (95,33%) e externa (98%) do algoritmo validam o grande potencial das técnicas de normalização e seleção de variáveis em aprendizado de máquina para fornecer resultados mais precisos e confiáveis em processos de classificação espectral. Essas constatações mostram que o método pode não só contribuir de maneira significativa com estudos que visam identificar marcadores genéticos associados à produção animal e à produção de alimentos, mas também para o melhor tratamento e análise de dados espectroscópicos FTIR.

**REFERÊNCIAS**

1. AGUIAR, Josafá C.; MITTMANN, Josane; FERREIRA, Isabelle; FERREIRA-STRIXINO, Juliana; RANIERO, Leandro. Differentiation of Leishmania species by FT-IR spectroscopy. **Spectrochimica Acta Part A**: Molecular and Biomolecular Spectroscopy, [S.L.], v. 142, p. 80-85, maio 2015. Elsevier BV. http://dx.doi.org/10.1016/j.saa.2015.01.008.
2. WANG, Yu-Tang; LI, Bin; XU, Xiao-Juan; REN, Hai-Bin; YIN, Jia-Yi; ZHU, Hao; ZHANG, Ying-Hua. FTIR spectroscopy coupled with machine learning approaches as a rapid tool for identification and quantification of artificial sweeteners. **Food Chemistry**, [S.L.], v. 303, p. 1-4, jan. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.foodchem.2019.125404.
3. LIMA, Juliana Soares. **ESPECTROFOTOMETRIA FTIR (Fourier Transform Infrared) E TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA A DETECÇÃO DE FRAUDE POR ADIÇÃO DE SORO DE QUEIJO AO LEITE CRU**. 2021. 124 f. Tese (Doutorado) - Curso de Ciência Animal, Universidade Federal de Minas Gerais, Belo Horizonte, 2021.
4. FORATO, Lucimara Aparecida *et al*. **A Espectroscopia na região do Infravermelho e algumas aplicações.** São Carlos: Embrapa Instrumentação, 2010. 16 p.
5. HSU, Sherman. **Infrared Spectroscopy**. [S. L.]: Separation Sciences Research And Product Development, [20--?]. 38 p. Disponível em: https://mmrc.caltech.edu/FTIR/Literature/General/IR%20spectroscopy%20Hsu.pdf. Acesso em: 22 set. 2023.
6. Schmidt, J., Marques, M.R.G., Botti, S. *et al.* Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater* 5, 83 (2019). https://doi.org/10.1038/s41524-019-0221-0.
7. ALLEGRETTA, Ignazio; MARANGONI, Bruno; MANZARI, Paola; PORFIDO, Carlo; TERZANO, Roberto; PASCALE, Olga de; SENESI, Giorgio S.. Macro-

classification of meteorites by portable energy dispersive X-ray fluorescence spectroscopy (pED-XRF), principal component analysis (PCA) and machine learning algorithms. **Talanta**, [S.L.], v. 212, p. 1-4, maio 2020. Elsevier BV. http://dx.doi.org/10.1016/j.talanta.2020.120785.

8. AKKAS, S; SEVERCAN, M; YILMAZ, O; SEVERCAN, F. Effects of lipoic acid supplementation on rat brain tissue: an ftir spectroscopic and neural network study. **Food Chemistry**, [S.L.], v. 105, n. 3, p. 1-2, 2007. Elsevier BV. http://dx.doi.org/10.1016/j.foodchem.2007.03.015.
9. SANTOS, Vianney O.; OLIVEIRA, Flavia C.C.; LIMA, Daniella G.; PETRY, Andrea C.; GARCIA, Edgardo; SUAREZ, Paulo A.Z.; RUBIM, Joel C.. A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis. **Analytica Chimica Acta**, [S.L.], v. 547, n. 2, p. 1-2, ago. 2005. Elsevier BV. http://dx.doi.org/10.1016/j.aca.2005.05.042.
10. RIOS, Thaynádia Gomes; LARIOS, Gustavo; MARANGONI, Bruno; OLIVEIRA, Samuel L.; CENA, Cícero; RAMOS, Carlos Alberto do Nascimento. FTIR spectroscopy with machine learning: a new approach to animal dna polymorphism screening. **Spectrochimica Acta Part A**: Molecular and Biomolecular Spectroscopy, [S.L.], v. 261, p. 1-2, nov. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.saa.2021.120036.
11. TOMAS, Rock Christian; SAYAT, Anthony Jay; ATIENZA, Andrea Nicole; DANGANAN, Jannah Lianne; RAMOS, Ma. Rollene; FELLIZAR, Allan; NOTARTE, Kin Israel; ANGELES, Lara Mae; BANGAOIL, Ruth; SANTILLAN, Abegail. Detection of breast cancer by ATR-FTIR spectroscopy using artificial neural networks. **Plos One**, [S.L.], v. 17, n. 1, p. 1-2, 26 jan. 2022. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0262489.
12. AHMED, Shiek S.s.J.; SANTOSH, Winkins; KUMAR, Suresh; CHRISTLET, T. Hema Thanka. Neural network algorithm for the early detection of Parkinson's disease from blood plasma by FTIR micro-spectroscopy.

**Vibrational Spectroscopy**, [S.L.], v. 53, n. 2, p. 1-2, jul. 2010. Elsevier BV. http://dx.doi.org/10.1016/j.vibspec.2010.01.019.

13. ARGYRI, A.A.; PANAGOU, E.Z.; TARANTILIS, P.A.; POLYSIOU, M.; NYCHAS, G.-J.e.. Rapid qualitative and quantitative detection of beef fillets spoilage based on Fourier transform infrared spectroscopy data and artificial neural networks. **Sensors And Actuators B**: Chemical, [S.L.], v. 145, n. 1, p. 1-2, 4 mar. 2010. Elsevier BV. http://dx.doi.org/10.1016/j.snb.2009.11.052.
14. AGATONOVIC-KUSTRIN, Snezana. The Use of Fourier Transform Infrared (FTIR) Spectroscopy and Artificial Neural Networks (ANNs) to Assess Wine Quality. **Modern Chemistry & Applications**, [S.L.], v. 01, n. 04, p. 1-2, 2013. OMICS Publishing Group. http://dx.doi.org/10.4172/2329-6798.1000110.
15. Lotter, B., Konde, S., Nguyen, J. *et al.* Identifying plastics with photoluminescence spectroscopy and machine learning. *Sci Rep* 12, 18840 (2022). https://doi.org/10.1038/s41598-022-23414-3.
16. Shi, L., Li, Y. & Li, Z. Early cancer detection by SERS spectroscopy and machine learning. *Light Sci Appl* 12, 234 (2023). https://doi.org/10.1038/s41377-023-01271-7.
17. Thomsen, B.L., Christensen, J.B., Rodenko, O. *et al.* Accurate and fast identification of minimally prepared bacteria phenotypes using Raman spectroscopy assisted by machine learning. *Sci Rep* 12, 16436 (2022). https://doi.org/10.1038/s41598-022-20850-z.
18. RULL, Fernando; VENERANDA, Marco; MANRIQUE-MARTINEZ, Jose Antonio; SANZ-ARRANZ, Aurelio; SAIZ, Jesus; MEDINA, Jesús; MORAL, Andoni; PEREZ, Carlos; SEOANE, Laura; LALLA, Emmanuel. Spectroscopic study of terrestrial analogues to support rover missions to Mars – A Raman-centred review. **Analytica Chimica Acta**, [S.L.], v. 1209, p. 339003, maio 2022. Elsevier BV. http://dx.doi.org/10.1016/j.aca.2021.339003.
19. BARTH, Andreas. Infrared spectroscopy of proteins. **Biochimica Et Biophysica Acta (Bba) - Bioenergetics**, [S.L.], v. 1767, n. 9, p. 1073-

1101, set. 2007. Elsevier BV. http://dx.doi.org/10.1016/j.bbabio.2007.06.004.

20. WANG, Rong; WANG, Yong. Fourier Transform Infrared Spectroscopy in Oral Cancer Diagnosis. **International Journal Of Molecular Sciences**, [S.L.], v. 22, n. 3, p. 1206, 26 jan. 2021. MDPI AG. http://dx.doi.org/10.3390/ijms22031206.
21. PASQUINI, Celio. Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications. **Journal Of The Brazilian Chemical Society**, [S.L.], v. 14, n. 2, p. 198-219, abr. 2003. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0103-50532003000200006.
22. CARVALHO, Thomas. **Espectro Eletromagnético**. [20--?]. Disponível em: https://www.infoescola.com/fisica/espectro-eletromagnetico/. Acesso em: 10 out. 2023.
23. **ESPECTROSCOPIA - EADQUI051** . Juiz de Fora: Universidade Federal de Juiz de Fora, 2013. Disponível em: https://www2.ufjf.br/quimicaead/wp-content/uploads/sites/224/2013/05/ESPECTROSCOPIA-NO-INFRAVERMELHO-PARTE1.pdf. Acesso em: 11 jul. 2023.
24. KHAN, S. A. et al. **Fourier transform infrared spectroscopy: Fundamentals and application in functional groups and nanomaterials characterization.** Handbook of Materials Characterization, p. 317–344, 2018.
25. **ESPECTROSCOPIA Molecular**. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro. Disponível em: https://www.maxwell.vrac.puc-rio.br/4333/4333_4.PDF. Acesso em: 18 set. 2023.
26. GRANT, Jacob. **How-To: Fourier Transform Infrared Spectroscopy (FTIR)**. Iowa: Professor Scott Shaw Research Group, [20--?]. 10 slides, color. Disponível em: https://chem.uiowa.edu/sites/chem.uiowa.edu/files/people/shaw/JSG%20-%20How%20To%20FTIR%20-%20141027.pdf. Acesso em: 12 jul. 2023.
27. FIGUEIREDO, M.S. **Estudo das Propriedades Ópticas e Termo-Ópticas do Biodiesel e suas Misturas. Dissertação (Programa de**

**Pós-Graduação em Física Aplicada)** - Universidade Federal de Mato Grosso do Sul. Campo Grande-MS. 2009.

28. BATES, J. B. **Fourier Transform Infrared Spectroscopy**. Published by : American Association for the Advancement of Science. Science, v. 191, n. 4222, p. 31–37, 1976.
29. NILSSON, Nils. **INTRODUCTION TO MACHINE LEARNING**: an early draft of a proposed textbook. Palo Alto: Stanford University, 1996. 208 p.
30. MAS o que *é* uma Rede Neural? | Deep learning, capítulo 1. [S.I.]: 3Blue1Brown, 2017. (18 min.), son., color. Disponível em: https://www.youtube.com/watch?v=aircAruvnKk. Acesso em: 13 out. 2023.
31. LEON, André Ponce de. **Redes Neurais Artificiais**. [20--?]. Disponível em: https://sites.icmc.usp.br/andre/research/neural/. Acesso em: 13 jul. 2023.
32. GALUSHKIN, Alexander. **Neural Networks Theory**. Moscow: Springer, 2007. 402 p.
33. HUANG, Jun; ROMERO-TORRES, Saly; MOSHGBAR, Mojgan. **Practical Considerations in Data Pre-treatment for NIR and Raman Spectroscopy**. 2010. Disponível em: https://www.americanpharmaceuticalreview.com/Featured-Articles/116330-Practical-Considerations-in-Data-Pre-treatment-for-NIR-and-Raman-Spectroscopy/. Acesso em: 14 jul. 2023.
34. FERRÉ, Romà Tauler I; WALCZAK, Beata; BROWN, Steven. **Comprehensive Chemometrics**: chemical and biochemical data analysis. [S. L.]: Elsevier, 2009. 2896 p.
35. COLUMN, Tony Davies. **Back to basics: spectral pre-treatments, derivatives**. 2007. Disponível em: https://www.spectroscopyeurope.com/td-column/back-basics-spectral-pre-treatments-derivatives. Acesso em: 14 jul. 2023.
36. BROWNLEE, Jason. **How to Choose a Feature Selection Method For Machine Learning**. 2020. Disponível em: https://machinelearningmastery.com/feature-selection-with-real-and-

categorical-data/#:~:text=Feature%20selection%20is%20the%20process,the%20performance%20of%20the%20model.. Acesso em: 15 jul. 2023.

37. HUANG, Samuel H.. Supervised feature selection: a tutorial. **Artificial Intelligence Research**, [S.L.], v. 4, n. 2, 9 abr. 2015. Sciedu Press. http://dx.doi.org/10.5430/air.v4n2p22.
38. KUHN, Max; JOHNSON, Kjell. **Applied Predictive Modeling**. [S. L.]: Springer, 2013. 613 p.
39. MENON, Kartik. **Feature Selection In Machine Learning: All You Need to Know**. 2023. Disponível em: https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning. Acesso em: 15 jul. 2023.
40. Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11.
41. JOLLIFE, I. T.; CADIMA, J. **Principal component analysis: A review and recent developments.** Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, v. 374, n. 2065, 2016. ISSN 1364503X.
42. ARLOT, Sylvain; CELISSE, Alain. A survey of cross-validation procedures for model selection. **Statistics Surveys**, [S.L.], v. 4, n. , p. 3-3, 1 jan. 2010. Institute of Mathematical Statistics. http://dx.doi.org/10.1214/09-ss054.
43. REFAEILZADEH, P. et al. Encyclopedia of Database Systems. Encyclopedia of Database Systems, 2020.
44. HOW machine learning works. 2020. Elaborado por DataRobot. Disponível em: www.datarobot.com/blog/how-machine-learning-works/. Acesso em: 20 out. 2023.