DEYVIS CABRINI TEIXEIRA DELFINO

ESTIMAÇÃO E CLASSIFICAÇÃO DO POTENCIAL HÍDRICO DE CAFEEIROS UTILIZANDO REFLECTÂNCIA ESPECTRAL E TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL

LAVRAS – MG

2024

DEYVIS CABRINI TEIXEIRA DELFINO

ESTIMAÇÃO E CLASSIFICAÇÃO DO POTENCIAL HÍDRICO DE CAFEEIROS UTILIZANDO REFLECTÂNCIA ESPECTRAL E TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

Prof. Dr. Danton Diego Ferreira
Orientador

Dra. Vânia Aparecida Silva
Coorientador
Dra. Margarete Marin Lordelo Volpato
Coorientadora

LAVRAS – MG
2024

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Delfino, Deyvis Cabrini Teixeira.
Estimação e Classificação do Potencial Hídrico de Cafeeiros Utilizando Reflectância Espectral e Técnicas de Inteligência Computacional / Deyvis Cabrini Teixeira Delfino. - 2023.
86 p. : il.

Orientador(a): Danton Diego Ferreira.
Coorientador(a): Margarete Marin Lordelo Volpato, Vânia Aparecida Silva.
Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2023.
Bibliografia.

1. Inteligência Computacional. 2. Potencial Hídrico. 3. Reflectância Espectral. I. Ferreira, Danton Diego. II. Volpato, Margarete Marin Lordelo. III. Silva, Vânia Aparecida. IV. Título.

DEYVIS CABRINI TEIXEIRA DELFINO

ESTIMAÇÃO E CLASSIFICAÇÃO DO POTENCIAL HÍDRICO DE CAFEEIROS UTILIZANDO REFLECTÂNCIA ESPECTRAL E TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

APROVADA em 21 de Dezembro de 2023.

Prof. Dr. Danton Diego Ferreira	UFLA
Dra. Margarete Marin Lordelo Volpato	EPAMIG
Dra. Vânia Aparecida Silva	EPAMIG
Prof. Dr. Augusto Santiago Cerqueira	UFJF
Prof. Dr. Wilian Soares Lacerda	UFLA

gov.br
Documento assinado digitalmente
DANTON DIEGO FERREIRA
Data: 27/02/2024 12:13:23-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Danton Diego Ferreira
Orientador

Dra. Vânia Aparecida Silva
Co-Orientador

Dra. Margarete Marin Lordelo Volpato
Co-Orientadora

LAVRAS – MG
2024

Dedico este trabalho aos meus pais e irmão que sem eles isso seria impossível.

AGRADECIMENTOS

Grato à minha família, que me apoiou e me inspirou para a realização deste trabalho. Grato ao meu orientador Danton, por todo apoio e atenção. As coorientadoras Margarete e Vânia que me ajudaram neste processo. Agradeço aos pesquisadores do Programa de Pós-Graduação da Universidade Federal de Lavras e aos pesquisadores da EPAMIG. Pelos apoios do Consórcio Pesquisa Café, CNPq, INCT-Café, Fapemig e Capes. Uma vez que, em equipe, foram essenciais para a realização desse projeto, com o suporte em conhecimento e desenvolvimento.

"A primeira regra é manter o espírito tranquilo. A segunda é enfrentar as coisas de frente e tomá-las pelo que realmente são."
(Marcus Aurelius)

RESUMO

O potencial hídrico é um dos principais indicadores das condições hídricas das plantas e por isso é muito utilizado nos estudos relacionados a produção agrícola. É medido de forma direta por um equipamento denominado bomba de Scholander, que requer um procedimento complexo e demorado. Entretanto, existe na literatura diversos estudos que relacionam o potencial hídrico e as curvas espectrais das folhas e dossel das plantas. Este trabalho tem como objetivo estudar as curvas espectrais de folhas de cafeeiros com diferentes potenciais hídricos através do uso de ferramentas de inteligência computacional e reconhecimento de padrões, utilizando amostras espectrais de lavouras de café cultivados em região com deficit hídrico climatológico. Foram utilizados dois bancos de dados, sendo o primeiro composto pelo espectro de frequência e energia refletida dos cafeeiros, e o segundo com os potenciais hídricos de cada planta. As amostras coletadas correspondem a cafeeiros da cidade de Diamantina, Minas Gerais, Brasil. Utilizou-se técnicas de pré-processamento para estruturar os dados, foram implementadas quatro técnicas de *Machine Learning*: Rede Neural Artificial tipo MLP (*Multi-Layer Perceptron*), Árvore de Decisão, *Random Forest* e KNN (*K-Nearest Neighbor*). Estas técnicas foram aplicadas na tarefa de regressão e na tarefa de classificação. Com o intuito de melhorar o desempenho das técnicas de *Machine Learning*, posteriormente, o algoritmo SMOTE foi executado gerando amostras sintéticas. Os resultados expõem que a Árvore de Decisão foi superior para a o método de regressão, com raiz do erro quadrático médio (*RMSE*) de 0,4342 e o coeficiente de determinação (R^2) de 0,6993 e para classificação a Rede Neural Artificial obteve acurácia balanceada 62,05%. Os resultados obtidos com a aplicação das metodologias foram positivos, já que as técnicas abordadas foram capazes de realizar as atividades propostas, estimar o potencial hídrico por meio de curvas espectrais e valores espectrais.

Palavras-chave: Cafeeiros, Aprendizado de Máquina, Potencial hídrico, Análise de Dados.

ABSTRACT

The water potential stands as one of the principal indicators of plant water conditions, and therefore, it is extensively used in agricultural production studies. Its direct measurement involves the use of a device called a Scholander pressure bomb, which necessitates a complex and time-consuming procedure. However, literature showcases several studies that correlate water potential with spectral curves of leaves and plant canopies. This work aims to investigate the spectral curves of coffee plant leaves exhibiting different water potentials by utilizing computational intelligence tools and pattern recognition, leveraging spectral samples from coffee plantations cultivated in regions experiencing climatological water deficits. Two databases were employed: the first comprised frequency spectrum and reflected energy data of the coffee plants, while the second contained the water potentials of each plant. The collected samples correspond to coffee plants from Diamantina, Minas Gerais, Brazil, subjected to pre-processing techniques to structure the data. Four machine learning techniques were developed: Multilayer Perceptron Artificial Neural Network (MLP), Decision Tree, Random Forest, and K-Nearest Neighbor (KNN). Two distinct methods—regression and classification—were implemented for the four techniques. To enhance the performance of the machine learning methods, the SMOTE algorithm was executed, generating synthetic samples. The results indicate that Decision Tree outperformed in the regression method, achieving a Root Mean Squared Error (*RMSE*) of 0.4342 and a Coefficient of Determination (R^2) of 0.6993. For classification, the artificial neural networks attained an overall accuracy of 62,05%. The outcomes derived from these methodologies were positive, given that the techniques employed effectively estimated water potential through spectral curves and spectral values..

Keywords: Coffee Plants, Machine Learning, Water Potential, Data Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O café é um dos principais produtos do agronegócio brasileiro, desde sua chegada ao país, em 1727, o café foi um dos maiores geradores de riquezas e um dos produtos mais importantes da história nacional. A produção de café está diretamente relacionada com as condições de hidratação das plantas, a forma mais confiável de se medir as condições hídricas da planta é com medidas no campo, porém, tradicionalmente são demoradas e complexas. É necessário desenvolver modelos que demonstrem a relação das condições hídricas das plantas com as curvas espectrais. As técnicas de análise de dados e inteligência artificial podem contribuir para que o monitoramento das condições hídricas de cafeeiros seja realizado com sensores ópticos de maneira rápida e segura.

Tradicionalmente, a condição hídrica da planta é mensurada via bomba de pressão, também conhecida como Câmara de Scholander, em que o valor do potencial hídrico (Ψ_W) é determinado através de amostras de folhas recolhidas das plantas que são submetidas a diferentes níveis de pressão. Contudo, este modo de mensuração implica em um demorado tempo de execução, além de ocasionar um risco ao operador. Considerando estas contrariedades é necessário idealizar maneiras de determinar as condições hídricas que sejam menos adversas. Estão presentes na literatura trabalhos que propõem mensurar as condições hídricas da planta de maneira indireta, uma dessas formas é via reflectância eletromagnética.

Um objeto quando atingido pela radiação solar tem parte da energia refletida para o espaço, fenômeno este denominado reflectância. A radiação que compõe o espectro eletromagnético, quando refletida, torna-se possível medir e determinar sua reflectância, que pode ser diferente para cada tipo de radiação que atinge o objeto. A curva que mostra como varia a reflectância de um objeto para cada comprimento de onda é denominada assinatura espectral e depende das propriedades do objeto. A assinatura espectral é capaz de fornecer diferentes informações sobre diversos aspectos relacionados à saúde da planta. Tais aspectos são estudados por especialistas da área, a fim de garantir a relevância das informações. Dessa forma, determinadas reflectâncias da assinatura espectral apresenta uma relação com o status hídrico da planta, relação esta que pode ser linear ou não linear e em diferentes graus, dependendo ainda do comprimento de onda da curva espectral. No que tange a inteligência artificial, suas diversas técnicas são passíveis de serem implantadas na tentativa de estimar características da planta de maneira indireta.

No trabalho de (GENC et al., 2013), foi proposto determinar o efeito do estresse hídrico no milho (*Zea mays L.*) usando índices espectrais e leituras de clorofila para avaliar os espectros de reflectância usando o método da árvore de decisão para distinguir níveis/severidade de estresse hídrico. Com a mesma premissa,

(NUNES et al., 2020) utilizou, além da técnica árvore de decisão, redes neurais artificiais para determinar as condições hídricas de cafeeiros através de amostras espectrais de duas lavouras. No entanto, assumi-se que as técnicas utilizadas também podem ser empregadas em diferentes conjuntos de dados de diferentes locais, de forma a explorar de uma perspectiva distinta o conceito de estresse hídrico da planta. Além do mais, é plausível implementar técnicas de pré-processamento e seleção de características distintas das encontradas na literatura visando melhorar o desempenho do estimador.

Com o intuito de explorar uma abordagem distinta dos trabalhos de (GENC et al., 2013) e (NUNES et al., 2020), o vigente estudo não aborda índices espectrais como PRI (*Photochemical Reflectance Index*), PSRI (*Plant Senescence Reflectance Index*), NDVI (*Normalized Difference Vegetation Index*), WBU (*Water Band Index*), ARI1 (*Anthocyanin Reflectance Index*), CRI1 (*Carotenoid Reflectance Index*), SIPI (*Structure Insensitive Pigment Index*), FRI (*Flavonol Reflectance Index*) e, sim, uma perspectiva de alcançar o potencial hídrico diretamente pela curva de reflectância e investigar qual é o comprimento de onda ou a faixa de comprimento de onda mais adequado para inferir o potencial hídrico do cafeeiro. A busca por essa abordagem via curva de reflectância é uma tentativa de tornar o processo de estimativa do potencial hídrico mais simples, com a perspectiva de implementá-lo em hardware acoplado a um sensor de baixo custo e dimensão reduzida.

O presente trabalho aborda a implementação de quatro métodos baseados, em redes neurais artificiais (BRAGA; CARVALHO; LUDERMIR, 2007), árvores de decisão (WITTEN; FRANK; HALL, 2011), *Random Forest* (GIANNINI et al., 2012) e *k-nearest neighbors* (*KNN*) (COVER; HART, 1967). São exploradas as técnicas de classificação e regressão de cada um dos quatro métodos desenvolvidos. Por fim, apresenta-se um estudo comparativo entre as metodologias abordadas.

O objetivo de explorar uma perspectiva de alcançar qual é o comprimento de onda ou a faixa de comprimentos de onda mais adequada para inferir o potencial hídrico do cafeeiro, utilizando técnicas de análise de dados e inteligência artificial. Iniciando pela execução do pré-processamento dos dados de potencial hídrico e curvas de reflectância, buscando a configuração que proporcione os melhores indicadores de desempenho e eliminando variáveis de valores indesejados. Em seguida, realizar uma seleção de características, determinando assim os melhores índices de refletância foliar nas diferentes condições de plantio, definir e organizar classes de acordo com o conhecimento especialista e com trabalhos encontrados na literatura, implementar diferentes arquiteturas de inteligência computacional para o estudo do condições hídricas de cafeeiros e identificar a faixa de comprimento de onda mais adequada para a estimativa do potencial hídrico de cafeeiros.

2 REFERENCIAL TEÓRICO

O presente capítulo busca aprofundar a compreensão dos conceitos e técnicas, situando-o dentro das teorias relevantes. Explorar as contribuições acadêmicas que informam a estrutura teórica desta pesquisa, delineando as correntes teóricas essenciais para a análise crítica. Ao mapear as perspectivas teóricas chave, pretende-se fornecer um alicerce sólido para a investigação, destacando a relevância e o papel destes referenciais no desenvolvimento do conhecimento sobre o tema em estudo.

2.1 Cafeicultura no Brasil

Localizado na quinta posição dos produtos de origem vegetal mais exportados pelo Brasil, o café é uma das commodities que mais contribuíram para a expansão das vendas externas do agronegócio em 2022 e atualmente, é um cultivo que gera 8 milhões de empregos diretos e indiretos, segundo relatório Ministério da Agricultura, Pecuária e Abastecimento (MAPA).

Conforme *United States Department of Agriculture* (*USDA*), o Brasil é o maior produtor de café tipo Arábica (*Coffea Arabica*) e o segundo no ranking na produção do café da espécie Robusta (*Coffea Canephora*), popular pela variedade Conilon, estando atrás apenas do Vietnã. No ranking global, o Brasil ocupa a primeira posição, considerando ambos os tipos (*Arabica* e *Canephora*), tornando assim o maior produtor de café no mundo.

De acordo com o Terceiro Boletim da Safra de Café do ano de 2023 publicado pela Conab (Companhia Nacional de Abastecimento) a exportação de café atingiu recorde de US$ 9,2 bilhões em 2022. A produção mundial de café, na safra 2023/24, está prevista em 174,3 milhões de sacas de 60 quilos, o que representa uma alta de 2,5%, na comparação com a temporada anterior. No acumulado dos oito primeiros meses de 2023, o Brasil exportou café para 143 países, sendo Estados Unidos e Alemanha os principais destinos, com respectivas participações de 17% e 13,2%, em quantidade, seguidos por Itália, com 7,9%, Bélgica, com 6,2% e Japão, com 6,1%.

No entanto, apesar de ser um dos principais rendimentos da área agro, ainda há grandes desafios para tornar o seu cultivo eficiente. Especificamente, existem adversidades que comprometem a saúde do ciclo de vida da safra. Uma das principais dificuldades é o estresse hídrico que pode ser considerado um dos principais fatores limitantes do crescimento do cafeeiro, uma vez que áreas cultivadas podem estar localizadas em regiões que apresentam restrições hídricas (DAMATTA; RAMALHO, 2006).

Em situações de deficiência hídrica no cafeeiro associada a elevadas temperaturas, ocorre o fechamento estomático, resultando na redução da taxa fotossintética devido à ausência de entrada de CO_2. Adicionalmente, o estresse hídrico acentuado pode ocasionar a mortalidade das raízes do café, especialmente nas camadas superficiais do solo. Observa-se também uma maior senescência foliar devido a alterações hormonais desencadeadas por essas condições. Portanto, torna-se imperativo desenvolver estratégias que visem minimizar os efeitos dos períodos de estiagem, conhecidos como veranicos no cafeeiro (GOMES; LIMA; CUSTóDIO, 2007; CARVALHO et al., 2013; GARCIA et al., 2011).

Uma maneira de assegurar o cultivo eficiente e a alta qualidade do produto é o conhecimento das relações hídricas da planta, com o intuito de mantê-la sempre hidratada, garantindo assim uma maior produtividade e um produto final de alto valor.

Uma forma de mensurar a condição hídrica da planta é via bomba de pressão, também conhecida como Bomba de Scholander, em que o valor do potencial hídrico (Ψ_W) é determinado através de amostras de folhas recolhidas das plantas que são submetidas a diferentes níveis de pressão.

2.2 Potencial Hídrico

A água assim como as demais substâncias apresenta a busca constante pelo equilíbrio termodinâmico, sujeito à tendência geral de se mover de locais de maior energia para locais de menores níveis de energia. A energia mencionada aqui é a capacidade de realizar trabalho (SALISBURY; ROSS, 1992; TAIZ; ZEIGER, 2010). Em seu trabalho (SLATYER; TAYLOR, 1960), sugeriu que o potencial químico da água poderia servir de base para propriedades importantes da água, introduzindo assim o conceito de potencial hídrico.

O potencial da água é a energia potencial da água em um sistema em comparação com a água pura quando a temperatura e a pressão são mantidas iguais. É também uma medida de quão livremente as moléculas de água podem se mover em um determinado ambiente ou sistema. O potencial da água é denotado pelo símbolo grego Ψ_W (psi) junto com a letra *W* de *Water* e medido em Pascal (Pa). À temperatura padrão, o potencial hídrico da água pura é zero. A adição de soluto à água pura diminui a energia cinética, diminuindo assim o potencial da água. Comparativamente, uma solução sempre tem baixo potencial de água do que a água pura. Em um grupo de células com diferentes potenciais de água, um gradiente de potencial de água é gerado. A água se moverá de um potencial hídrico mais alto para um potencial hídrico mais baixo (LACERDA; FILHO; PINHEIRO, 2007; HOPKINS; HüNER, 2008; TAIZ; ZEIGER, 2010; VIEIRA, 2008;

SILVA et al., 2021). O potencial hídrico da planta (Ψ_W) é determinado pela soma de vários componentes, que são:

- Potencial Osmótico representado por Ψ_S;
- Potencial de Pressão denotado por Ψ_P;
- Potencial Gravitacional indicado por Ψ_g;
- Potencial Matricial evidenciado por Ψ_m.

Ao correlacionar os fatores, o potencial da água é escrito como mostrado na Equação 2.1.

$$\psi_W = \psi_S + \psi_P + \psi_g + \psi_m \tag{2.1}$$

O potencial do osmótico (Ψ_S), também conhecido como potencial de soluto, denota o efeito do soluto dissolvido no potencial da água. Na água pura, a adição de soluto reduz sua energia livre e diminui o valor do potencial hídrico de zero para negativo. Assim, o valor do potencial do soluto é sempre negativo. O potencial de pressão (Ψ_P) é uma força mecânica trabalhando contra o efeito do potencial de soluto. O aumento do potencial de pressão aumentará o potencial da água e a água entrará na célula e as células ficarão túrgidas. Essa pressão hidrostática positiva dentro da célula é chamada de pressão de Turgor. Da mesma forma, a retirada de água da célula diminui o potencial hídrico e a célula torna-se flácida.

O potencial gravitacional (Ψ_g) expressa a ação do campo gravitacional sobre a energia livre da água. Ele é definido como o trabalho necessário para manter a água suspensa em determinado ponto em relação a atração da gravidade. O potencial matricial (Ψ_m) representa a atração entre a água e o coloide hidratante ou moléculas orgânicas semelhantes a gel na parede celular, representativo em situações particulares de solos secos e sementes em embebição (LACERDA; FILHO; PINHEIRO, 2007; HOPKINS; HüNER, 2008; SALISBURY; ROSS, 1992; TAIZ; ZEIGER, 2010). O estudo do potencial hídrico é altamente relevante para o entendimento das relações hídricas nas plantas e o meio exterior (solo e atmosfera).

Conforme mencionado anteriormente o valor do potencial hídrico (Ψ_W) é determinado através de amostras de folhas recolhidas das plantas que são submetidas a diferentes níveis de pressão. Contudo, este modo de mensuração implica em um demorado tempo de execução, deve ser estimado em um horário específico (entre 4:00 e 5:00 horas), necessita de mão de obra especializada, além de ser um ensaio destrutivo e pode ocasionar um risco ao operador. Considerando estas contrariedades é necessário idealizar maneiras de determinar as condições hídricas que sejam menos adversas. Estão presentes na literatura trabalhos que

propõem mensurar as condições hídricas da planta de maneira indireta, uma dessas formas é via assinatura espectral.

A assinatura espectral é capaz de fornecer diferentes informações sobre diversos aspectos relacionados à saúde da planta. Tais aspectos são estudados por especialistas da área, a fim de garantir a relevância das informações. Dessa forma, determinadas reflectâncias da assinatura espectral possuem uma relação com o status hídrico da planta, relação esta que pode ser linear ou não linear e em diferentes graus, dependendo ainda do comprimento de onda da assinatura espectral.

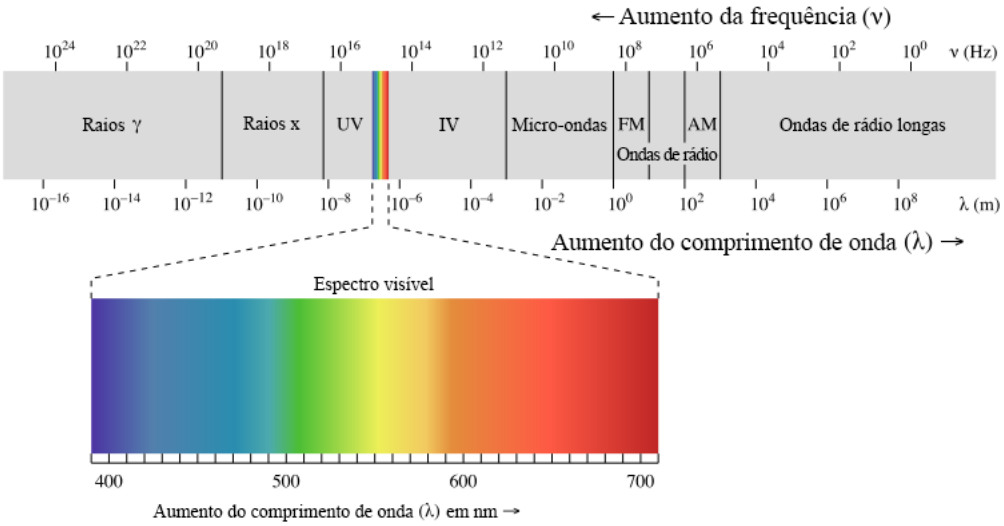
2.3 Espectro Eletromagnético

A principal fonte geradora de energia do sistema solar é a estrela Sol, a energia gerada é irradiada para todo o espaço. Propagando-se pelo vácuo a energia radiante, também chamada radiação solar, atinge a Terra onde é em parte refletida e parte absorvida. A radiação solar é fundamental para a vida na Terra. Essa energia aponta ao funcionamento dos processos atmosféricos e climatológicos. Além disso, é direta ou indiretamente responsável por determinadas circunstâncias cotidianas, como a fotossíntese das plantas, a manutenção de uma temperatura compatível com a vida (MORAES, 1993).

Sempre que um objeto é atingido por radiação solar, a mesma é absorvida, refletida e transmitida. A capacidade de absorver a energia radiante é denominada absortância, a característica de refletir é nomeada de reflectância e por fim a transmitância é a propriedade de transmitir a energia radiante. Considerando um objeto escuro e opaco, é possível afirmar que o próprio possui valor quase nulo de transmitância, baixo valor de reflectância e alto para absortância (MORAES, 1993).

A reflectância de um objeto pode ser mensurada para cada tipo de radiação que compõe o espectro eletromagnético e é possível observar que a reflectância de um mesmo objeto pode ser diferente para cada tipo de radiação que o atinge. A medida que o comprimento de onda de uma determinada onda aumenta, a frequência diminui e, à medida que o comprimento de onda diminui, a frequência aumenta. Assim a radiação eletromagnética é então agrupada em categorias com base em seu comprimento de onda ou frequência no espectro eletromagnético. Os diferentes tipos de radiação eletromagnética mostrados no espectro consistem em ondas de rádio, micro-ondas, ondas infravermelhas, luz visível, radiação ultravioleta, raios X e raios gama. A parte do espectro eletromagnético que podemos ver é o espectro de luz visível (ATKINS; PAULA, 2006; MORAES, 1993). A Figura 2.1 apresenta o espectro eletromagnético e os diferentes tipos de radiação que o compõe.

Figura 2.1 – Espectro eletromagnético.



Fonte: Adaptado (VO; HERNANDEZ; PATEL, 2022)

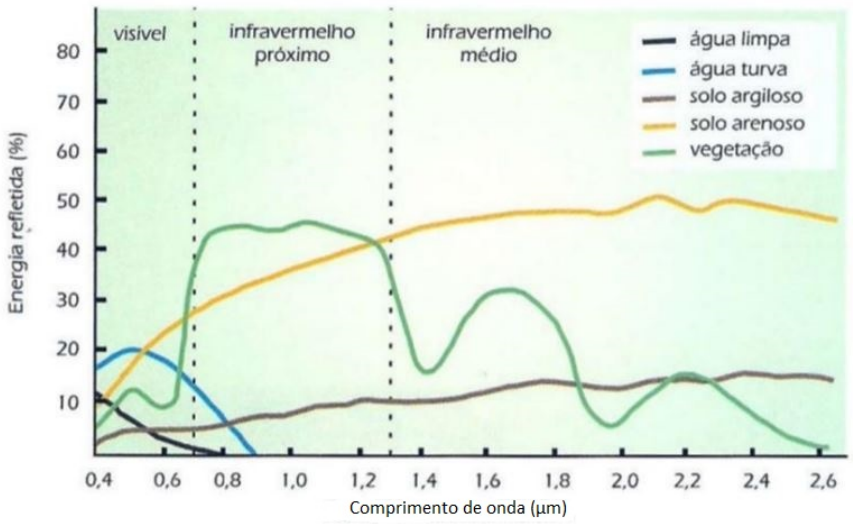
Pela Figura 2.1 a luz visível é a única parte do espectro eletromagnético que os humanos podem ver a olho nu. Esta parte do espectro inclui uma gama de cores diferentes que representam um determinado comprimento de onda. A luz passa através da matéria na qual é absorvida ou refletida com base em seu comprimento de onda e dependendo da composição do objeto/alvo algumas cores são refletidas mais do que outras. Além da parte visível do espectro eletromagnético outra faixa importante é o infravermelho que pode ser liberado como calor ou energia térmica. O infravermelho também pode ser devolvido, o que é chamado de infravermelho próximo por causa de suas semelhanças com a energia da luz visível (ATKINS; PAULA, 2006; CHANG, 2005).

2.4 Assinatura espectral

Como mencionado anteriormente, superfícies distintas, como a água, o solo descoberto ou a vegetação, refletem a radiação de forma diferente em vários canais do espectro. A radiação refletida em função do comprimento de onda é denominada assinatura espectral da superfície. Partindo dessa premissa é possível utilizar as reflectâncias distintas dos objetos/alvos para identificar a presença de determinados elementos em sua composição, por exemplo, a quantidade de água contida em plantas (PONZONI, 2002). É possível descrever um padrão na reflectância de um objeto/alvo para diferentes regiões do espectro eletromagnético, ou seja, objetos que são semelhantes apresentam o mesmo padrão de reflectância e absortância, isso para os

vários comprimentos de onda da radiação incidente, esse padrão é definido por um gráfico e denominado assinatura espectral (SANTOS, 2013; PONZONI, 2002). A Figura 2.2 apresenta o comportamento padrão da assinatura espectral de alguns alvos naturais.

Figura 2.2 – Assinatura espectral alvos naturais.



Fonte: (FLORENZANO, 2002)

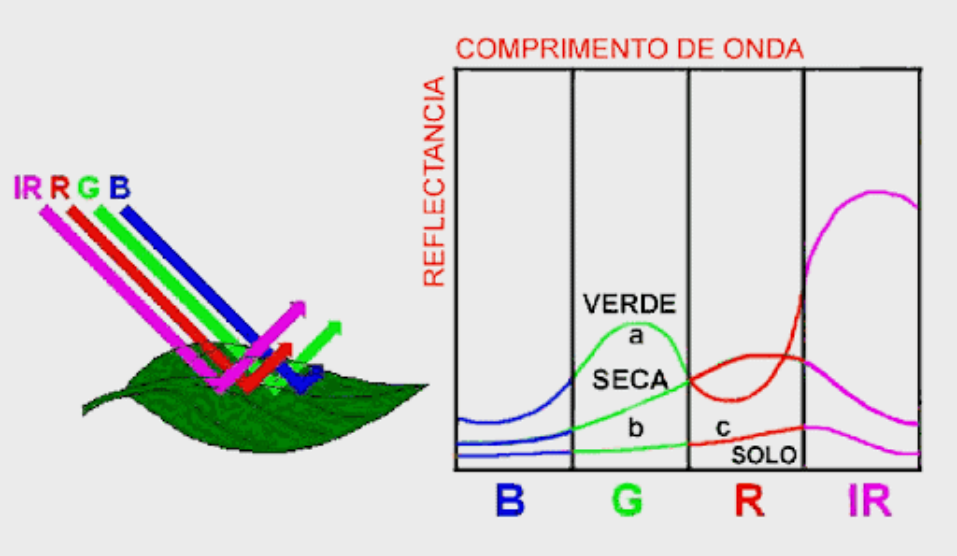
Através da Figura 2.2 a resposta padrão de diferentes curvas para alvos naturais é exibido, o eixo horizontal do gráfico apresenta os comprimento de onda que diferenciam o tipo de energia eletromagnética, já no eixo vertical mostra a reflectância, na qual, valor é expressado geralmente em porcentagem. Quanto mais alta a curva do gráfico maior a reflectância do alvo para aquele comprimento de onda, para curvas mais baixas há pouca reflectância e muita absortância. A curva de cor verde da Figura 2.2 é mostrado a assinatura espectral da vegetação, em que é possível conhecer quais energias do visível e infravermelho serão mais refletidas e em qual proporção. Qualquer vegetação similar a essa, ainda que não seja o mesmo tipo vegetal, apresentará uma assinatura muito parecida com a resposta padrão mostrada (PONZONI, 2002).

Apesar do fato de que as folhas das plantas muitas vezes parecem semelhantes, elas variam amplamente em forma e composição química, tanto quanto a concentração de água nos espaços intercelulares das folhas. Isso resulta em refletância continuamente variada da planta. As folhas das plantas absorvem a maior parte da radiação na faixa visível por pigmentos vegetais como clorofila e xantofilas, mas refletem principalmente no infravermelho próximo (KATSOULAS et al., 2016). Vários autores como (JACQUEMOUDD; FREDERIC, 1990; KACIRA et al., 2005; DELALIEUX et al., 2007; DATT, 1999; CECCATO et al., 2001) relataram que as diferentes características químicas e físicas, como, o teor de água, teor de carbono e com-

postos como hidrogênio, fosforo e potássio em diferentes formas influenciam nas características de sua curva espectral.

A radiação com comprimento de onda superior a 950 nm é geralmente absorvida pelo líquido foliar, enquanto a radiação em aproximadamente 1000 nm é absorvido pela matéria seca da folha (compostos de carbono e nutrientes). A refletância dos comprimentos de onda entre 680-750 nm são influenciados pela concentração de água e nutrientes, enquanto o espectro de refletância entre 750-800 nm sofre as principais variações pela concentração de água na folha (KATSOULAS et al., 2016; SOUDANI et al., 2012). A Figura 2.3 exemplifica a diferença entre uma folha seca e hidratada.

Figura 2.3 – Assinaturas espectrais.



Fonte: Adaptado (STEFFEN, 2016)

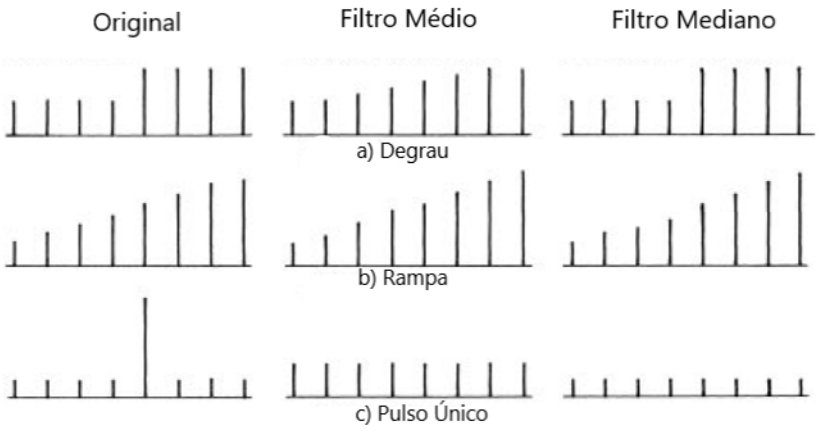
A Figura 2.3 apresenta três assinaturas espectrais, sendo, curva ”a” uma folha verde, ”b” de uma folha seca e a curva ”c” do solo. Além das assinaturas espectrais é representado no gráfico as bandas do visível (B, G e R), em que, ”B” do inglês *blue* representa a banda de cor azul do espectro, o ”G” do inglês *green* retrata a banda verde e a letra ”R” do inglês *red* que caracteriza a banda vermelha. Já as letras ”IR” do inglês *infrared* simboliza o infravermelho do espectro. Considerando a banda do visível, o alto percentual de absorção foliar da radiação solar causa uma saturação rápida do sinal refletido. A aparência verde da vegetação está relacionada com a sua maior reflectância na banda (G) e é produzida pela clorofila. A alta reflectância na banda infravermelha (IR) resulta da interação da radiação com a estrutura celular superficial da folha, com os aspectos fisiológicos e varia com o seu conteúdo de água na estrutura celular superficial

(SOUDANI et al., 2012; STEFFEN, 2016). Desse modo é possível estabelecer o estresse hídrico de uma planta considerando seu percentual de reflectância (GUTIERREZ; REYNOLDS; KLATT, 2010).

2.5 Filtro Mediano

O Filtro Mediano foi desenvolvida por (TUKEY, 1977), que consiste em suavizar o ruído do tipo impulsivo em sinais e imagens digitais. Uma imagem ou um sinal digital pode ser representada por uma matriz. Dada uma matriz A de inteiros positivos, com m linhas e n colunas, e dados dois inteiros positivos e ímpares p e q, o filtro da mediana produz uma matriz transformada M, com as mesmas dimensões que A, definida da seguinte maneira: para cada par de índices (i, j), o elemento $M(i, j)$ da matriz transformada é a mediana dos elementos de A_{ij} da vizinhança $p \times q$ em torno de (i, j) (PRATT, 2001). A Figura 2.4 apresenta exemplos de diferentes sinais e sua forma após passar pelo filtro de média e filtro mediano.

Figura 2.4 – Filtragem mediana em sinais de teste unidimensionais.



Fonte: Adaptado (PRATT, 2001)

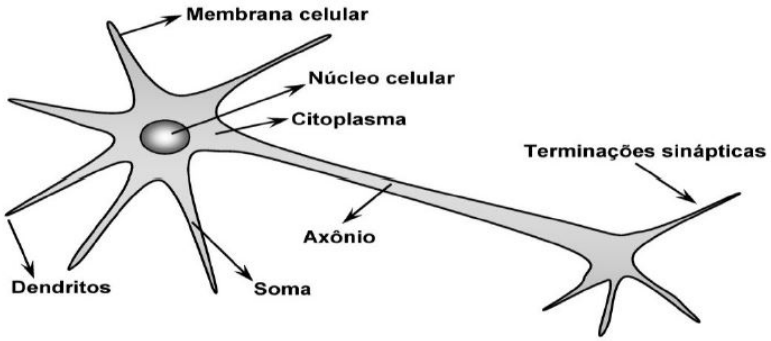
2.6 Redes Neurais Artificiais

A inteligência biológica apresenta como unidade básica os neurônios, que por sua vez conectados em rede promovem uma perfeita e complexa estrutura neurológica. A ciência que estuda as redes neurais biológicas se propõe a replicar artificialmente a estrutura e a função do cérebro humano por meio de mecanismos conhecidos como redes neurais artificiais (RNAs). Uma rede neural é um sistema paralelo distribuído que consiste em unidades de processamento simples (nós) que calculam certas funções matemáticas. Essas

unidades são organizadas em camadas e conectadas entre si por um grande número de conexões, que por sua vez estão associadas a diferentes valores de peso projetados para ponderar a entrada recebida por cada neurônio da rede. A principal característica da RNA é a capacidade de aprender e generalizar informações a partir de exemplos, além de absorver conhecimento de seu ambiente (BRAGA; CARVALHO; LUDERMIR, 2007; HAYKIN, 2001).

A Figura 2.5 apresenta o corpo da célula, os dendritos e o axônio, que são as principais partes de um neurônio biológico. Impulsos nervosos provenientes de outros neurônios são transmitidos pelos dendritos até o corpo celular, este por sua vez é responsável pelo processamento de todas as informações recebidas, resultando em um novo impulso nervoso. O presente impulso move-se pelo axônio do neurônio até os dendritos de outros neurônios. O ponto de contato entre a terminação do axônio de um neurônio e o dendrito de outro neurônio é denominado de sinapse. Os neurônios biológicos são modelados matematicamente de manira a obter um neurônio artificial, ao passo que as redes neurais artificiais representam muitos desses neurônios artificiais interconectados (BRAGA; CARVALHO; LUDERMIR, 2007; HAYKIN, 2001; SILVA; SPATTI; FLAUZINO, 2016).

Figura 2.5 – Neurônio Biológico Simplificado.

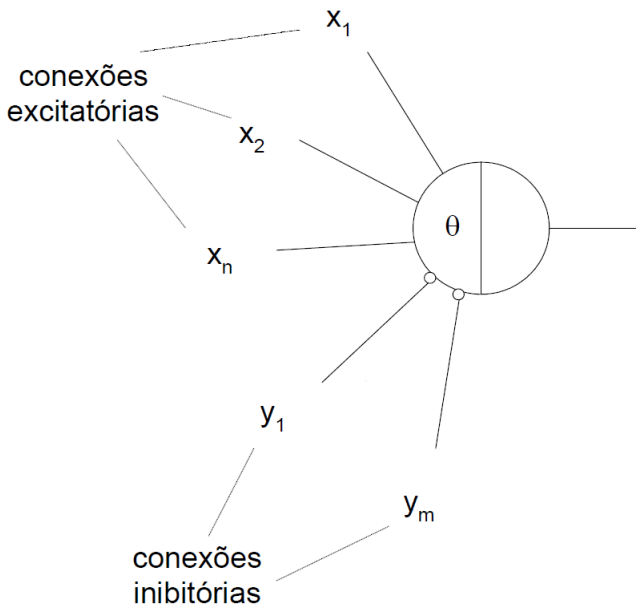


Fonte: (SILVA; SPATTI; FLAUZINO, 2016)

Baseado nas características do neurônio biológico McCulloch e Pitts em seu trabalho (MCCULLOCH; PITTS, 1943) propuseram o primeiro que se tem notícia, neurônio artificial, nomeado MCP. A Figura 2.6 exibe a representação do modelo MCP.

De maneira análoga ao neurônio biológico o modelo artificial MPC mostrado na Figura 2.6 consiste basicamente de um neurônio que executa uma função lógica. Os nós produzem somente resultados binários e são compostas de conexões sem peso. O modelo MLC apresenta três elementos básicos. O primeiro, representado pelos valores de x_1, x_2, ..., x_n são conexões do tipo excitatórios, quanto mais sinais excitatórios

Figura 2.6 – Modelo do neurônio MCP.



Fonte: (BISHOP, 1995)

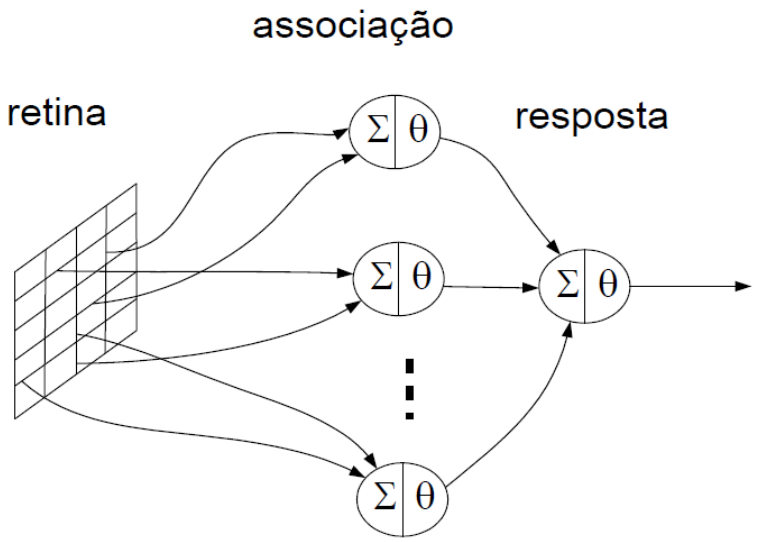
um neurônio receber, mais próximo o total estará do limiar do neurônio e, portanto, mais próximo o neurônio estará de enviar seu sinal de saída. Já os valores y_1, y_2, ..., y_m são conexões do tipos inibitórios, possuem o efeito de inibir o neurônio de enviar um sinal. Quando um neurônio recebe um sinal inibitório, ele fica menos excitado e, portanto, leva mais sinais excitatórios para atingir o limiar do neurônio. Por fim, cada unidade é caracterizada por um certo limiar (threshold) θ (BISHOP, 1995; AZEVEDO; BRASIL; OLIVEIRA, 2000; REZENDE, 2003).

2.6.1 Perceptron

Em 1958 Frank Rosemblatt mostrou em seu livro “Principles of Neurodynamics” o modelo do perceptron, ilustrado na Figura 2.7. Nele, os neurônios (perceptrons) eram organizados em camadas, com o intuito de se atingir a eficiência sináptica, que foi testada no reconhecimento de caracteres.

Pela Figura 2.7, em que, a topologia original descrita por Rosenblatt é exibida, o peceptron é composto por unidades de entrada (retina), que consiste basicamente em elementos sensores, por um nível intermediário formado pelas unidades de associação, que é constituído por nodos MCP de pesos fixos, definidos antes do período de treinamento e por um nível de saída formado pelas unidades de resposta, no qual é a única camada que possui propriedades adaptativas. Por essa característica de apenas o nível de saída pos-

Figura 2.7 – Modelo do neurônio Perceptron.



Fonte: Adaptado (BISHOP, 1995)

suir qualidades adaptativas o modelo é conhecido como perceptron de única camada. O perceptron quando treinado de maneira correta, sempre chega a uma solução para o problema de separação de duas classes linearmente separáveis em um tempo finito (KROSE; SMAGT, 1996; FURTADO; MACAU; VELHO, 2011; BRAGA; CARVALHO; LUDERMIR, 2007).

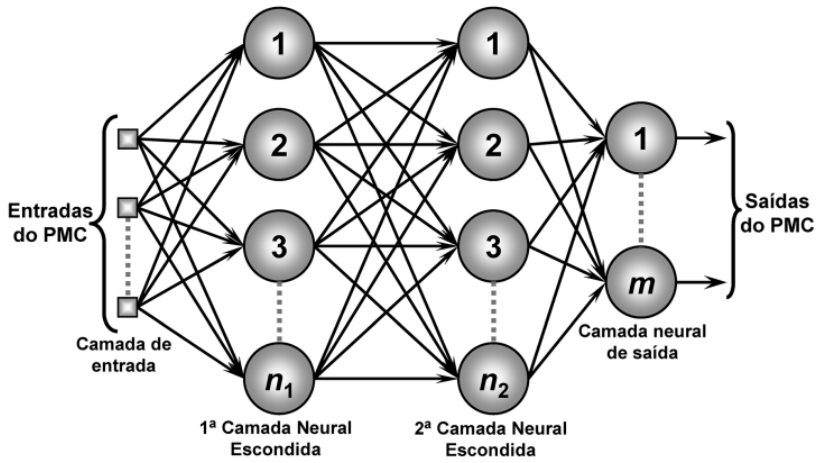
Após duras críticas de Minsky e Papert (MINSKY; PAPERT, 1969) sobre capacidade computacional do perceptron, as pesquisas em rede neurais artificiais sofreu um grande impacto negativo, levando a um grande desinteresse pela área durante os anos 70 e início dos anos 80. Somente com o trabalho de Hopfield em 1982 (HOPFIELD, 1982) e do algoritmo *back-propagation* em 1986, que a área de RNAs ganhou novo impulso, ocorrendo a partir do final dos anos 80, uma forte expansão no número de trabalhos de aplicação e teóricos envolvendo RNAs e técnicas correlatas (BRAGA; CARVALHO; LUDERMIR, 2007).

2.6.2 Redes *Multilayer Perceptron* - Redes MLP

O perceptron simples resolve apenas problemas linearmente separáveis. A solução de problemas não linearmente separáveis passa pelo uso de redes com uma ou mais camadas intermediárias/escondidas, as Redes *Multilayer Perceptron* (MLP). Segundo (CYBENKO, 1988) a utilização de duas camadas intermediárias permite a aproximação de qualquer função continua.

As Redes MLP consistem em diferentes nós interligados, de modo que cada nó é um neurônio. Cada neurônio possui um peso de entrada. A capacidade de adaptação em diferentes ambientes consiste no ajuste dos pesos em cada conexão. A Figura 2.8 ilustra o esquema de uma Rede MLP.

Figura 2.8 – Ilustração de rede Perceptron multicamadas.



Fonte: Adaptado (SILVA; SPATTI; FLAUZINO, 2016)

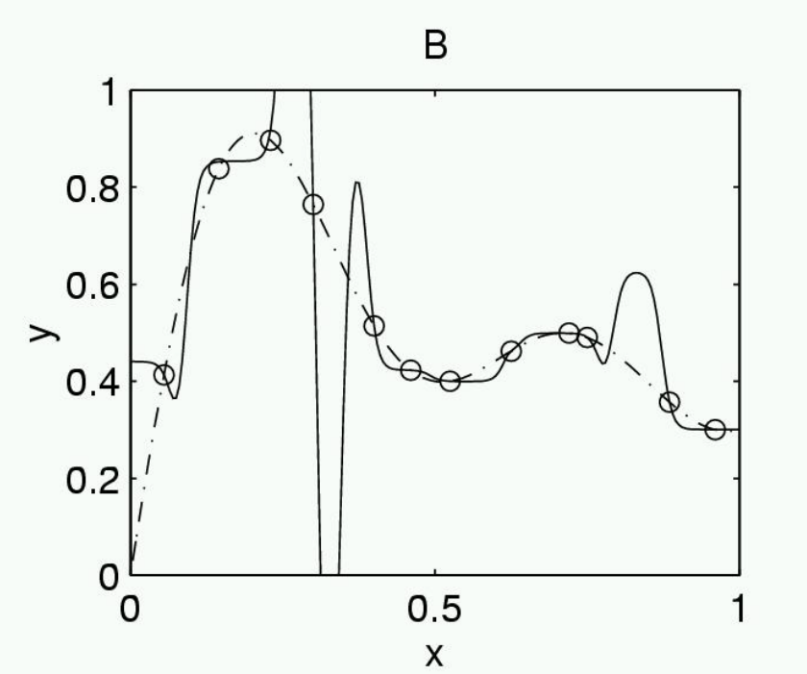
Redes MLP demonstram um poder computacional muito maior do que aquele apresentado pelas redes sem camadas intermediárias. Segundo (CYBENKO, 1988), redes com duas camadas intermediárias podem implementar qualquer função, seja uma função continua ou descontinua. De acordo com (BRAGA; CARVALHO; LUDERMIR, 2007), alguns fatores devem ser levados em consideração para a consolidação de uma Rede MLP. A estrutura é divida em três camadas: a primeira camada, recebe os valores de entrada, não havendo nenhum tipo de processamento; na segunda camada está a maior parte do processamento; e na terceira camada se encontra a resposta da rede, onde a mesma recebe os resultado das camadas intermediárias e realiza o processamento. É importante ressaltar que a segunda camada da estrutura pode ser composta por inúmeras camadas intermediárias;porém quanto maior esse número, mais elevada a complexidade do projeto.

Outro fator a ser considerado no projeto de redes neurais é o número de neurônios em cada camada. Seguindo a mesma ideia de camadas intermediárias, quanto mais neurônios houver, mais complexa a rede se torna. O número apropriado de neurônios depende de vários fatores, que vão desde a complexidade da função que a rede deve aprender e treinar até a distribuição estatística dos dados. Portanto, determinar o valor ótimo do número de neurônios é um dos problemas fundamentais quando se trata de redes neurais. Embora não exista uma regra ou fórmula para determinar o número de neurônios necessários para uma rede neural resolver um determinado problema, existem vários trabalhos na literatura que podem orientar

a estimativa do tamanho da rede (CYBENKO, 1988; BRAGA; CARVALHO; LUDERMIR, 2007; SILVA; SPATTI; FLAUZINO, 2016).

Alguns autores como (BRAGA; CARVALHO; LUDERMIR, 2007; HAYKIN, 2001; SILVA; SPATTI; FLAUZINO, 2016) apontam que o número de neurônios não é infinito. Assim como muitos neurônios podem causar problemas relacionados a erros, números pequenos em uma camada de rede dificultam a convergência da rede para a resposta correta. Portanto, aumentando o número de neurônios nas camadas intermediárias, a capacidade de mapeamento não linear da rede também é aumentada. No entanto, se esse número for grande, o modelo pode se tornar suscetível ao ruído das amostras de treinamento e sobre-ajustar os dados. Nessa situação, diz-se que a rede está sujeita ao sobre-treinamento *overfitting*. A Figura 2.9 apresenta uma situação de *overfitting*.

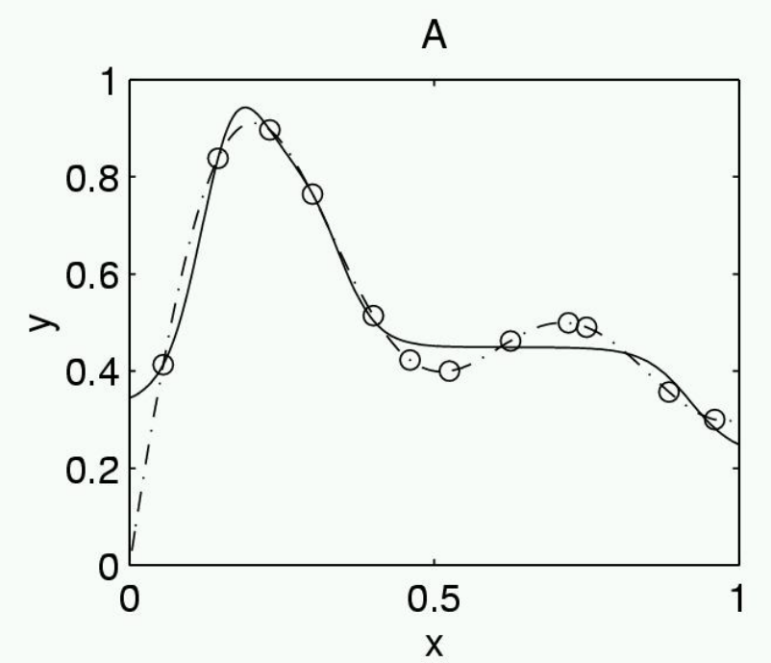
Figura 2.9 – *Overfitting*.



Fonte: Adaptado (KROSE; SMAGT, 1996)

Por outro lado, um número muito pequeno de neurônios obriga a rede a consumir muito tempo tentando encontrar a melhor representação, impedindo que a rede realize o mapeamento desejado. Essa situação é chamada de *underfitting*. Figura 2.10 representa o caso de *underfitting*. Vale ressaltar que para a Figura 2.9 e Figura 2.10, a linha tracejada representa a resposta real da função *f(x)*. A linha contínua mostra a resposta *f(x)* estimada pelo modelo para os valores correspondentes de *x*.

Figura 2.10 – *Underfitting*.



Fonte: Adaptado (KROSE; SMAGT, 1996)

2.6.3 Funções de Ativação

Pelo trabalho de (HAYKIN, 2001), na grande maioria dos projetos, o modelo de cada unidade de uma rede neural possui não linearidades em sua saída. A função de ativação representa a influência das entradas internas e do estado de ativação atual na definição do próximo estado de ativação da rede. Portanto, existem funções de ativação lineares e não lineares. Este último é responsável por fornecer recursos de mapeamento não linear para redes neurais do tipo MLP. Os dois tipos de funções de ativação não lineares mais comumente usados são: a função sigmodal padrão e a função tangente hiperbólica, representadas nas Equações 2.2 e 2.3.

$$f(x) = \frac{1}{(1+e^{-x})} \tag{2.2}$$

$$f(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{(1+e^{-x})} \tag{2.3}$$

A função sigmóide mapeia qualquer valor de x para um intervalo entre 0 e 1. Isso a torna particularmente útil em problemas onde a saída desejada precisa ser interpretada como uma probabilidade (HAYKIN, 2001).

Como mencionado anteriormente, como a função sigmóide padrão apresenta apenas valores de ativação na faixa (0,1), ela é substituída pela função tangente hiperbólica em alguns casos. Dessa forma, a forma sigmóide da função logística é preservada, mas com capacidade de assumir valores positivos e negativos. Segundo (FLECK et al., 2016), devido à necessidade de reduzir o tempo de processamento, a função tangente hiperbólica pode ser substituída por uma função matematicamente equivalente chamada tangente sigmóide, que é mais simples de calcular e, portanto, mais rápida de calcular. Apesar das diferenças numéricas associadas à tangente hiperbólica, ela é uma boa alternativa às redes neurais, onde a forma exata da função é menos importante uma vez que suas propriedades matemáticas são determinadas. A Equação 2.4 apresenta a formulação para a Tangente Hiperbólica Sigmoidal. Para problemas de regressão ou estimação, a função de ativação da última camada (a camada de saída) pode ser linear, permitindo que a rede retorne qualquer valor real em sua saída. Em geral utiliza-se a função puramente linear, dada pela Equação 2.5.

$$f(x) = \frac{2}{(1 - e^{-2x})} - 1 \tag{2.4}$$

$$f(x) = x \tag{2.5}$$

2.6.4 Aprendizado de Redes Neurais Artificiais

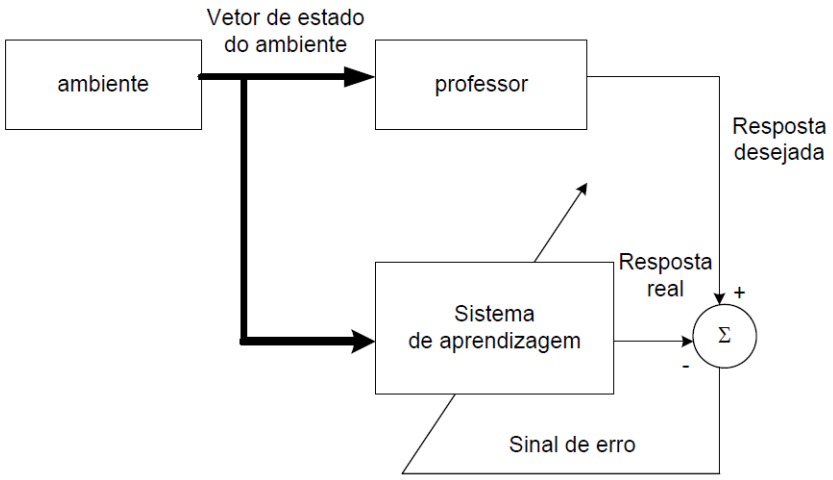
A propriedade mais importante das redes neurais é a habilidade de aprender com seu ambiente e assim melhorar seu desempenho. Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos, o treinamento. O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas. Denomina-se algoritmo de aprendizado a um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos de aprendizado específicos para determinados modelos de redes neurais. Estes algoritmos diferem entre si principalmente pelo modo como os pesos são modificados (KROSE; SMAGT, 1996; HAYKIN, 2001).

De acordo com o (MIRANDA; FREITAS; FAGGION, 2009), o treinamento de uma rede neural envolve um processo iterativo de ajuste aplicado aos pesos. Portanto, acredita-se que o aprendizado ocorre apenas quando a rede neural atinge uma solução generalizada para um determinado problema. Assim, o

termo “treinar uma rede neural” inclui ajustar os pesos sinápticos para que os vetores de saída coincidam com os valores esperados para cada vetor de entrada.

Pelos trabalhos de (BISHOP, 1995; HAYKIN, 2001; BRAGA; CARVALHO; LUDERMIR, 2007; HAYKIN, 2008), se destaca três tipos de aprendizado em redes neurais artificiais: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, uma RNA recebe a saída desejada em relação ao padrão de entrada. Desta forma, a saída da rede neural é comparada com a saída esperada, obtendo-se um erro referente à resposta atual. Em seguida, os pesos são ajustados para minimizar o erro. A Figura 2.11 demonstra um diagrama de blocos do aprendizado supervisionado.

Figura 2.11 – Diagrama de blocos aprendizado supervisionado.



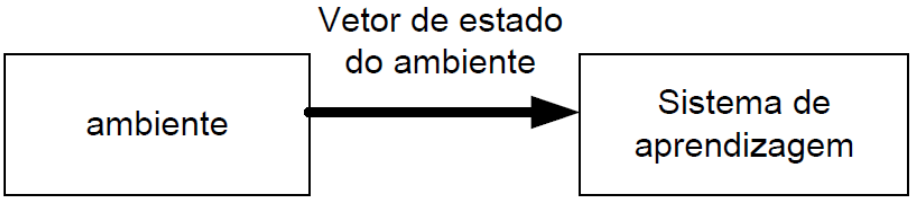
Fonte: Adaptado (HAYKIN, 2008)

Também conhecida com aprendizagem com professor, o aprendizado supervisionado mostrado no diagrama de blocos da Figura 2.11 consiste em que o professor tenha o conhecimento do ambiente, e fornece o conjunto de exemplos de entrada-resposta desejada. Com esse conjunto, o treinamento é feito usando a regra de aprendizagem por correção de erro (KROSE; SMAGT, 1996; HAYKIN, 2001).

No aprendizado não supervisionado, não há supervisor/professor acompanhado pelo processo de aprendizagem. Por esse motivo, as RNAs devem procurar algum tipo de correlação ou redundância nos dados de entrada (HAYKIN, 2008; KROSE; SMAGT, 1996). O digrama de blocos exibido na Figura 2.12 ilustra o funcionamento do aprendizado não supervisionado.

Nesse modelo não supervisionado, também conhecido como auto-organizado, são dadas condições para realizar uma medida da representação que a rede deve aprender, e os parâmetros livres da rede são

Figura 2.12 – Diagrama de blocos aprendizado não supervisionado.

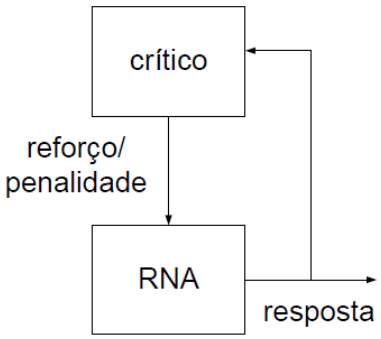


Fonte: Adaptado (HAYKIN, 2008)

otimizados em relação a essa medida. Para a realização da aprendizagem não-supervisionada pode-se utilizar a regra de aprendizagem competitiva (HAYKIN, 2008; KROSE; SMAGT, 1996; HAYKIN, 2001).

Por fim, aprendizagem por reforço pode ser visto como caso particular de aprendizagem supervisionada. A principal diferença entre o aprendizado supervisionado e o aprendizado por reforço é a medida de desempenho usada em cada um deles. A Figura 2.13 ilustra um diagrama de blocos do aprendizado por reforço.

Figura 2.13 – Diagrama de blocos aprendizado por reforço.



Fonte: Adaptado (HAYKIN, 2008)

No aprendizado supervisionado, a medida de desempenho é baseada no conjunto de respostas desejadas usando um critério de erro conhecido, enquanto que no aprendizado por reforço a única informação fornecida à rede é se uma determinada saída está correta ou não. A ideia básica tem origem em estudos experimentais sobre aprendizado dos animais. Quanto maior a satisfação obtida com uma certa experiência em um animal, maiores as chances dele aprender (HAYKIN, 2008; **??**).

2.7 Árvores de Decisão

De acordo com (WITTEN; FRANK; HALL, 2011), uma árvore de decisão é um modelo estatístico de aprendizado supervisionado de padrões em um conjunto de dados. No trabalho de (HALMENSCHLAGER, 2002) esta ferramenta é definida como uma representação em árvore formado por um conjunto de nós interconectados, que é útil para a classificação e previsão de amostras desconhecidas. Desta forma, nós internos testam os atributos de entrada com constantes de decisão e determinam qual será o próximo nó descendente. Os nós folha, por sua vez, classificam as instâncias que os alcançam com base nos rótulos associados a eles. Portanto, o conhecimento nessa estrutura é representado por cada nó, e na hora do teste, a busca é direcionada aos nós descendentes até chegar a um nó folha a Figura 2.14 exibe um exemplo estrutural de uma árvore de decisão.

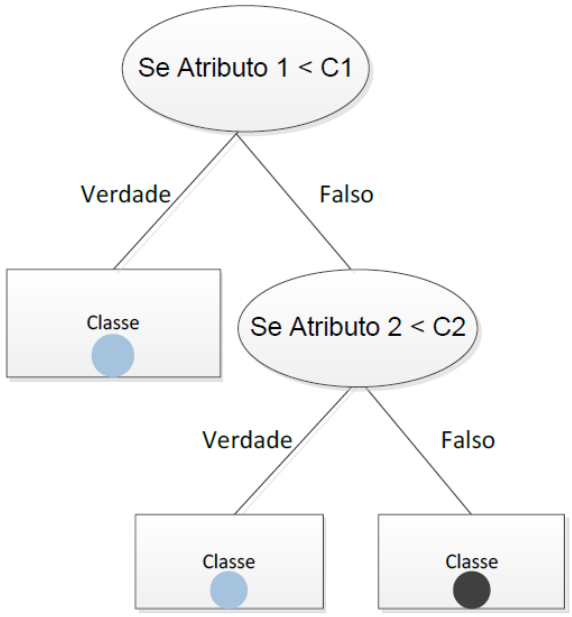
Segundo (BORGES, 2013), as árvores de decisão possuem características relevantes que viabilizam o método tais como:

- Acurácia: sendo o modelo capaz de avaliar ou prever corretamente classes, agrupamentos e regras;
- Robustez: apresentando resultados satisfatórios mesmo quando se utiliza dados com ruído ou valores faltantes;
- Interpretabilidade: com alto nível de compreensão proporcionado pelo modelo;
- Flexibilidade: devido ao fato de o espaço da instância ser compartimentalizado em subespaços, e ajustável a diferentes modelos.

Segundo (QUINLAN, 1983), a árvore de decisão é utilizada para resolver um problema complexo dividindo-o em problemas mais simples, aos quais se aplica recursivamente a mesma estratégia; as soluções para o subproblema podem então ser combinadas na forma de uma árvore para gerar uma solução para o problema complexo.

No trabalho de (WITTEN; FRANK; HALL, 2011), a justificativa para o funcionamento de uma árvore de decisão é fundamentada na estratégia de “dividir para conquistar”. Essa abordagem envolve a subdivisão do espaço definido pelos atributos em subespaços, sendo que cada subespaço pode ser subdividido novamente e associado a uma classe específica. A Figura 2.14 exemplifica o mecanismo de classificação para duas classes, onde inicialmente se compara o valor de um atributo com uma constante. O gráfico, separado por dois espaços, representa de maneira mais apropriada as duas classes, como mostrado na Figura 2.15 (a).

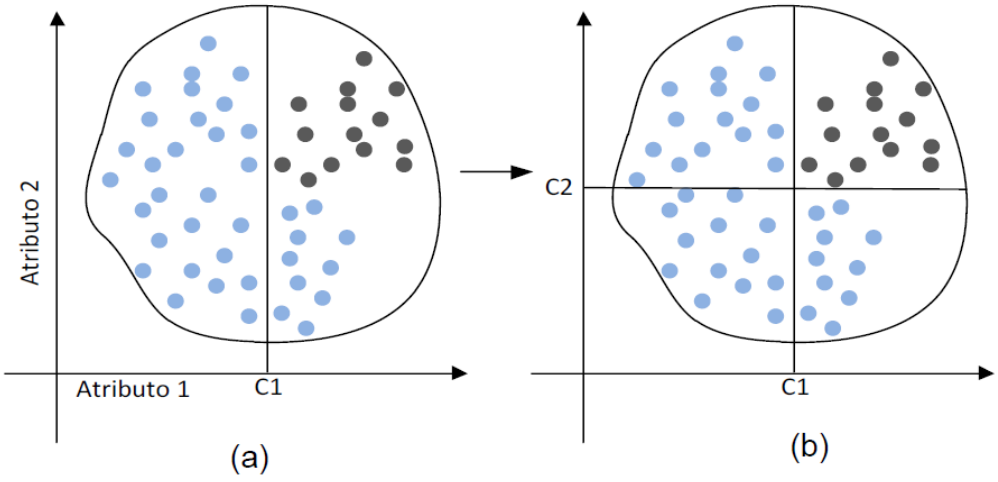
Figura 2.14 – Estrutura de uma árvore de decisão.



Fonte: (BORGES, 2013)

É perceptível que a fronteira inicial não separou satisfatoriamente as duas classes, levando à possibilidade de estabelecer uma nova fronteira que melhor define os espaços das classes, conforme ilustrado na Figura 2.15 (b).

Figura 2.15 – Processo de indução de uma árvore de decisão. (a) Obtenção da Primeira fronteira de decisão. (b) Obtenção da segunda fronteira de decisão.



Fonte: (BORGES, 2013)

Um algoritmo que descreve o processo de indução da raiz até folha de uma árvore de decisão foi proposto por (MURTHY, 1998; RAMOS, 2014), de modo que uma condição de parada seja atendida. As etapas em questão são:

1- Se todos os objetos de treinamento no nó corrente pertencem a mesma categoria, crie um nó folha;

2- Avalie cada um dos possíveis testes condicionais usando uma função heurística;

3- Escolha o melhor teste condicional como teste do nó corrente;

4- Crie um nó sucessor para cada resultado distinto e particione os dados de treinamento entre os nós sucessores utilizando o teste definido no passo 3;

5- Verifique se o nó é puro, ou seja, se todos os objetos de treinamento no nó corrente são da mesma categoria. Repita os passos anteriores (1,2,3 e 4) em todos os nós impuros.

Segundo (FACELI et al., 2011), as condições de parada do crescimento de uma árvore mais comumente usadas são:

a) Todos os objetos de treinamento pertencerem a uma única classe.

b) O tamanho máximo da árvore foi atingido.

c) O número de objetos em um ou mais nós é menor do que um limiar previamente definido.

Pelo trabalho de (RAMOS, 2014), após a construção da árvore, o modelo deve ser validado por classificação ou regressão em amostras não apresentadas anteriormente na fase de treinamento. Este processo funciona da seguinte forma: Os objetos são apresentados ao nó raiz da árvore e submetidos a testes condicionais. O objeto então segue o caminho e é apresentado ao novo nó de subdivisão. Isso determinará o novo caminho. Este processo é repetido até que o objeto encontre um nó folha com um rótulo representando sua classificação.

2.7.1 Escolha de Atributos

A pesquisa de (BORGES, 2013) mostra que, ao construir uma árvore de decisão, deve-se selecionar em cada nó o atributo que melhor divide o conjunto de treinamento. Portanto, o objetivo do algoritmo de derivação é obter o melhor atributo a ser usado no nó, usando um método que valida cada atributo candidato

e seleciona a melhor classe discriminante. De acordo com (MONARD; BARANAUSKAS, 2003), as opções de seleção de atributos mais utilizadas são:

a) Aleatória: o atributo é selecionado aleatoriamente;

b) Menos Valores: seleciona o atributo com a menor quantidade de valores possíveis;

c) Mais Valores: seleciona o atributo com a maior quantidade de valores possíveis;

d) Ganho Máximo: seleciona o atributo que possui o maior ganho de informação esperado.

A seleção aleatória leva a situações complicadas, como árvores muito grandes e, como as propriedades do mundo real geralmente contêm ruído, leva muito tempo para selecionar propriedades que podem ou não importar. Para resolver esse problema, a inspeção determina o modo de seleção de cada recurso e avalia sua qualidade e/ou representatividade para melhorar o desempenho da árvore de decisão. As funções de avaliação mais utilizadas são a função entropia (SHANNON, 1948) e a função Gini (WEISS; KULIKOWSKI, 1990). O objetivo geral é minimizar a contaminação e os danos nos nós atuais e futuros.

2.7.2 Poda de Árvores de Decisão

O critério de parada ou técnica de poda é um passo importante para tomar a decisão certa em uma árvore de decisão. Isso ocorre porque o processo de construção da árvore é propenso ao *overfitting*, que ocorre quando a árvore cresce demais e ocorre quando novos nós dividem pequenas partições dos dados em partições menores (BORGES, 2013). Em seu trabalho (RAMOS, 2014) explicou que o problema do crescimento excessivo de árvores, torna a árvore muito extensas e complexas. Uma técnica chamada poda é usada para resolver esse problema.

Em seus estudos (FACELI et al., 2011) aponta que a poda de árvores consiste na substituição de nós por folhas e visa aumentar a capacidade de generalização, pois nós mais profundos representam mais amostras de treinamento e também reduzem o tamanho da árvore resultante. Porém, nesse sentido, (QUINLAN, 1983) definiu que existem basicamente dois tipos de poda: métodos que interrompem a construção da árvore quando determinados critérios são atendidos, conhecidos como pré-poda, e métodos que fazem a poda em um momento posterior após uma árvore completa ser estabelecida, conhecido como *pós-trimming*.

O método de pré-poda serve para determinar os critérios de parada do algoritmo de construção da árvore. Assim, quando todos os intervalos possíveis usando a mesma propriedade produzirem um ganho

menor que um valor pré-definido, o nó corrente pode ser criado para transformar em um nó com ganho informacional. Construir uma árvore de decisão substituindo subárvores por nós de folha representando as classes mais frequentes no ramo e, em seguida, executar uma técnica de pós-poda. A técnica *post-can* calcula a taxa de erro quando as subárvores são excluídas. Se esta taxa de erro for menor que um valor predeterminado, a árvore é podada. Caso contrário, a poda não ocorrerá.

2.8 *Random Forest*

Segundo Giannini (GIANNINI et al., 2012), *Random Forest* é um importante algoritmo de aprendizado de máquina que usa dados de presença e ausência (binários) para desenvolver padrões de distribuição preditivos, e o algoritmo é expresso como uma sequência finita de instruções executadas em uma linguagem computacional. *Random Forest* é um dos métodos *ensemble* disponíveis na literatura juntamente com o *Boosting* e *Bagging*. Os Métodos *Ensemble* são algoritmos formados por uma coleção de classificadores que realizam o processo de votação das classes mais apontadas pelos próprios classificadores (OSHIRO, 2013).

Muitas vezes, as árvores de decisão não podem aumentar a complexidade sem reduzir sua capacidade de generalização para novos dados. No trabalho de (HO, 1995) foi proposto um método chamado *Random Decision Forests* para remover a limitação de complexidade das árvores de treinamento. O método proposto de construção do modelo permite aumentar arbitrariamente a complexidade da árvore sem incorrer em penalidade de desempenho em dados ainda não visualizados. O método propõe a construção de múltiplas árvores em subespaços escolhidos do espaço preditor para que generalizem os resultados de forma complementar. Por fim, a construção de resultados agregados dessas árvores sem restrições de profundidade pode melhorar o desempenho do modelo sem perda da capacidade de generalização.

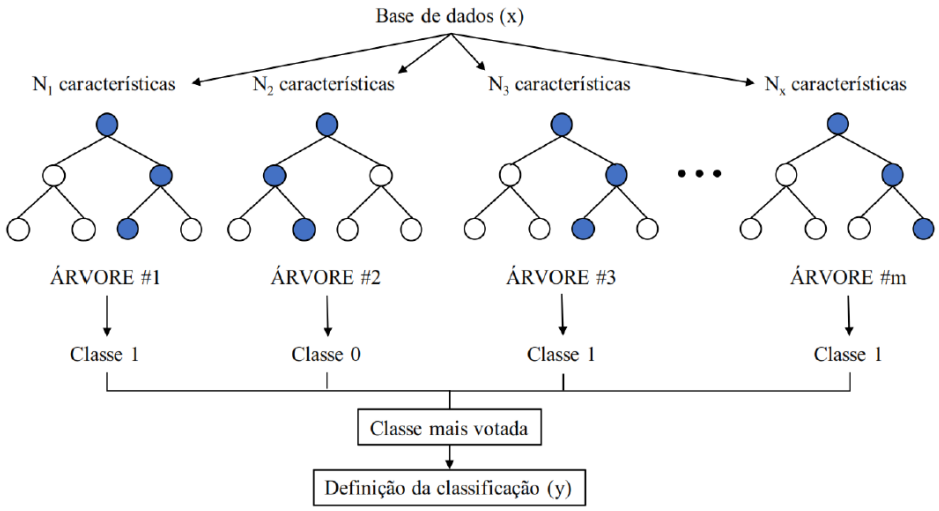
2.8.1 Construção de uma *Random Forest*

Uma *Random Forest* constrói muitas árvores de decisão que serão usadas para classificar um novo exemplo por meio do voto majoritário. Cada árvore de decisão usa um subconjunto de atributos selecionados aleatoriamente a partir do conjunto original, contendo todos os atributos (HO, 1995; BREIMAN, 2001).

Random Forest é um classificador formado por uma coleção de árvores $\{h(\bar{x}, \Theta_k), k = 1, ...\}$ onde os Θ_k são vetores aleatórios independentes e de distribuição idêntica. Cada árvore prevê uma classe da entrada $\bar{x}$ por um voto unitário, e a classe mais popular entre as arvores é eleita para a mesma na entrada (BREIMAN, 2001; DUBATH et al., 2011; ZHAO; ZHANG, 2007).

Portanto, a *Random Forest* gera uma amostra aleatória do conjunto de dados de treinamento para cada árvore. A seleção do subconjunto de variáveis para reduzir a correlação das árvores é similar ao proposto em (HO, 1995), se diferenciando apenas por gerar subconjuntos no nível dos nós e não para toda a árvore. A Figura 2.16 exibe um exemplo de funcionamento de *Random Forest*.

Figura 2.16 – Funcionamento do método *Random Forest*.



Fonte: Adaptado (KIRASICH; SMITH; SADLER, 2018)

De acordo com a Figura 2.16 cada árvore da *Random Forest* apresenta seu voto unitário para a classe e por fim a mais cotada é definida.

A classificação incorreta de uma floresta depende da força das árvores individuais, ou seja, a força de cada árvore na floresta pode ser interpretada como uma medida do desempenho individual. Uma árvore com uma taxa de erro baixa é um classificador forte. Portanto, aumentando a força de árvores individuais, a taxa de erro da floresta pode ser reduzida. A seleção aleatória de atributos tornam as árvores diferentes e, portanto, menos correlacionadas entre si. Correlações baixas tendem a reduzir as taxas de erro de classificação (BREIMAN, 2001; BREIMAN, 2003)).

A construção de modelos *Random Forest* é relativamente simples e não requer tratamento especial de variáveis qualitativas. Um pequeno número de hiperparâmetros, que representam parâmetros previamente definidos para controlar o treino, torna mais fácil encontrar valores que proporcionem um bom desempenho. Etapas mais complexas estão relacionadas ao pré-processamento nas fases de tratamento, imputação e seleção de variáveis. É necessário definir os valores para o número de árvores, o número de variáveis aleatoriamente selecionadas em cada particionamento e o tamanho mínimo do nó. Pelo trabalho de (PROBST;

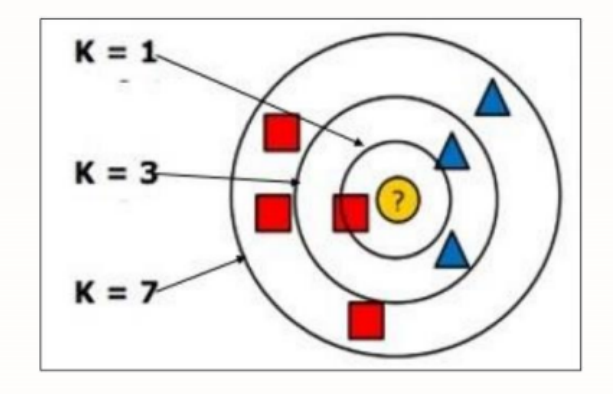
WRIGHT; BOULESTEIX, 2019; WRIGHT; KöNIG, 2019), variáveis qualitativas ordinais podem ser tratadas da mesma forma que preditores numéricos, gerando os mesmos resultados que tratar esses valores como nominais, mas com menor complexidade computacional.

2.9 kNN (*k-nearest neighbors*)

O algoritmo *K-Nearest Neighbor* (KNN), do português K-Vizinho Mais Próximo, pertence à família de algoritmos *Instance-based Learning* (IBL), ou seja, aprendizagem baseada em instâncias (KOTSIANTIS, 2007). Algoritmos de aprendizado baseados em instâncias são algoritmos de *lazy-learning* (MITCHELL, 1997), pois atrasam o processo de indução ou generalização até que a classificação seja realizada. Os algoritmos armazenam exemplos de treinamento e, quando é necessário fazer uma previsão para um novo exemplo (instância), a generalização é adiada até que a decisão específica seja tomada.

Para o *k-Nearest Neighbor (kNN)* uma instância dentro de um conjunto de dados geralmente são encontradas próximas a outras instâncias que possuem propriedades semelhantes (COVER; HART, 1967). Se as instâncias forem marcadas com um rótulo de classificação, o valor do rótulo de uma instância não classificada pode ser determinado observando a classe de seus vizinhos mais próximos. O *kNN* localiza as *k* instâncias mais próximas da instância de consulta e determina sua classe identificando o único rótulo de classe mais frequente (KOTSIANTIS, 2007). A Figura 2.17 exibe o funcionamento do *k-Nearest Neighbor (kNN)*.

Figura 2.17 – Representação *kNN*.



Fonte: (MEDEIROS et al., 2020)

A Figura 2.17 possui uma instância a ser classificada, representada pela interrogação, e instâncias já classificadas representadas por triângulos e quadrados. Nesse contexto, para valores de *k=1*, a operação do

algoritmo *KNN* classifica a nova instância como pertencente à classe quadrado. Isso ocorre porque a classe vizinha mais próxima é uma classe quadrada. Para o valor $k = 3$, se duas instâncias dos três vizinhos mais próximos pertencem a classe triangulo e apenas uma a classe quadrado, então a nova instância terá a classe triangulo. Para $k = 7$, a classe da nova instância é uma classe quadrada. Como mencionado anteriormente, a maior frequência de classe dos k vizinhos mais próximos da instância que está sendo classificada é o determinante da classificação.

Algoritmos de *lazy-learning* requerem menos esforço computacional durante a fase de treinamento, porém é necessário mais esforço computacional durante o processo de classificação. Pois requer a comparação da amostra de teste com todas as amostras presentes no conjunto de treinamento, considerando banco de dados muito grandes é uma fragilidade do método. A escolha do número de vizinhos também afetará a precisão dos resultados do algoritmo e precisa ser escolhida com mais cuidado (KOTSIANTIS, 2007).

Além disso, o algoritmo é mais adequado para problemas com menos parâmetros devido à maldição da dimensionalidade. Quanto mais dimensões no conjunto de dados, mais as métricas de distâncias perdem relevância, fazendo com que a distância de amostras muito próximas a um ponto não seja muito diferente da distância de amostras distantes dele (KOUIROUKIDIS; EVANGELIDIS, 2011; CONSONNI et al., 2021; MAATEN; HINTON, 2009).

2.9.1 Número de Amostras

Em seu trabalho (KUSNER et al., 2014) aponta que um dos problemas do método *kNN* é o alto custo computacional para realizar os testes. Para efetuar a classificação de uma instância é preciso calcular a distância de cada instância já anteriormente classifica com a nova, dependendo do número de dimensões essa classificação pode requerer um tempo elevado. Além disso, à medida que o banco de dados cresce, o tamanho necessário para armazenar cada amostra em teste pode causar problemas de escalabilidade.

Isso torna o *kNN* um método pouco adequado para previsão em tempo real em bancos de dados muito grandes. Existem várias maneiras de reduzir o tempo e o espaço necessários para cálculos de previsão algorítmica, algumas das quais reduzem o número de cálculos da distância, a dimensionalidade e reduzem o número de amostras (KUSNER et al., 2014) .

Com o intuito de reduzir o problema de escalabilidade sem perder as informações que o conjunto original oferece, vários autores como (KUSNER et al., 2014; ANGIULLI; PIZZUTI, 2005) desenvolveram métodos que reduzem o número de instâncias de diferentes maneiras. Um desse métodos é o de sintetizar

informações, que reduz o número de amostras selecionando subconjuntos e descartando amostras redundantes.

2.9.2 Número de Vizinhos

A fim de obter resultados precisos, um critério importante no algoritmo *kNN* é o número de vizinhos (k). Considerando dados esparsos e ruído, pequenas vizinhanças podem produzir resultados ruins, enquanto vizinhanças muito grandes podem levar a resultados bastante abrangentes, em que amostras de outras classes poderão ser incluídas (IMANDOUST; BOLANDRAFTAR, 2013; MARIZ, 2017).

O número de vizinhos deve ser escolhido com cuidado, e foi demonstrado que escolher um número apropriado de vizinhos no início do algoritmo torna-se difícil se a distribuição das instâncias não for uniforme (IMANDOUST; BOLANDRAFTAR, 2013).

2.9.3 Cálculo de Distância

Com a finalidade de identificar os vizinhos mais próximos, é necessário primeiramente calcular a distância entre eles. Para isso, funções de distância são utilizadas, em que a distância Euclidiana é a técnica mais empregada. A distância Euclidiana é normalmente definida como a distância mais curta entre dois pontos (TANG; HE, 2015). A Equação 2.6 ilustra a manira de calcular a distância Euclidiana.

$$dist_{Euc}(E_i, E_j) = \sqrt{\sum_{l=1}^{M} (x_{il} - x_{jl})^2} \tag{2.6}$$

Considerando que existe dois pontos E_i e E_j pertencentes a um espaço de m-dimensões, cada ponto contempla um dado e sua notação é definida por $E_i = (x_{i1}, x_{i2}, ..., x_{iM})$ e $E_j = (x_{j1}, x_{j2}, ..., x_{jM})$. A Equação 2.6 calcula o comprimento do segmento de reta que conecta os pontos E_i e E_j (MEDEIROS et al., 2020).

Além da distância Euclidiana existem outras técnicas que calculam a distância entre duas instâncias. Dois métodos muito utilizados são *Manhattan* e *Minkowsky*. A distância *Manhattan* entre dois pontos $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ e $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ no espaço n-dimensional é a soma das distâncias em cada dimensão. Uma característica da *Manhattan* é que instâncias muito distantes não possuem influencia considerável na determinação de uma nova (SAMMUT; WEBB, 2011). A Equação 2.7 apresenta como calcular distância *Manhattan*. Já a distância *Minkowsky* é uma métrica em um espaço vetorial normado, a qual pode ser considerada como uma generalização de ambas as distâncias Euclidiana e *Manhattan* (SAMMUT; WEBB, 2011). A Equação 2.8 mostra como calcular a distância *Minkowsky*.

$$dist_{Man}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_j| \tag{2.7}$$

$$dist_{Min}(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{n} |x_i - y_j|^r)^{1/r} \tag{2.8}$$

2.9.4 Peso e Dimensionalidade

Em modelos tradicionais, cada vizinho possui a mesma relevância para definir qual é a classe predominante na vizinhança, independentemente de sua distância ao ponto considerado. Dependendo do objetivo da previsão, escolher uma fórmula que dê mais peso a vizinhos específicos pode aumentar ou diminuir a eficiência do algoritmo, considerando uma amostra que possua dados esparsos ou ruído, que podem interferir na classificação. É possível atribuir pesos usando diferentes grupos de cálculos de distância, pois conforme a técnica escolhida instâncias mais próximas ou mais distantes serão beneficiadas (IMANDOUST; BOLANDRAFTAR, 2013; MARIZ, 2017).

Os efeitos da maldição da dimensionalidade vêm sendo tratados com estratégias que buscam a redução da dimensão por meio de extração e de seleção de atributos (WANG H.AND ZHOU; YAN, 2021; DING; PENG, 2005; DARBELLAY; VAJDA, 1999). Outra maneira explorada, é através da utilização de descritores, considerados estados da arte para imagens, como por exemplo, as *CNNs (Convolutional Neu-ral Networks)*(ALZUBAIDI et al., 2021; WANG et al., 2019). Esse descritor de características possibilita eleger camadas distintas da rede para formar os vetores de características. Vetores com menor dimensionalidade são encontrados geralmente em camadas mais externas (LIMA; FARIA; BARIONI, 2022). Enquanto os vetores gerados por camadas mais internas têm maior dimensão. (ROMASZEWSKI; GŁOMB; CHOLEWA, 2018; WU et al., 2020) em seus trabalhos, têm empregado de maneira eficaz o conceito denominado *hubness* para dados de alta dimensão. O aspecto *hubness* consiste na tendência de algumas instâncias de dados, chamadas *hubs*, ocorrerem com maior frequência nas listas dos K-vizinhos mais próximos de outras instâncias (MANI P.AND VAZQUEZ et al., 2019).

2.10 Métricas de Desempenho

As métricas de desempenho representam um papel fundamental na avaliação e refinamento de modelos de *machine learning*, fornecendo uma medida quantitativa do quão bem um modelo realiza em uma tarefa específica. As métricas de desempenho são ferramentas essenciais para avaliar, comparar e melho-

rar modelos, desempenhando um papel crucial em todas as fases do ciclo de vida de desenvolvimento de modelos, desde a concepção até a implementação e manutenção contínua.

2.10.1 Acurácia Balanceada

Um desafio recorrente na avaliação de modelos de classificação é a obtenção de medidas que capturem de forma precisa o grau de acurácia na identificação de exemplos não vistos. A prática comum envolve a utilização da média das precisões obtidas em dobras de validação cruzada individuais. Entretanto, esta abordagem apresenta limitações substanciais. Uma dessas limitações é que na presença de um conjunto de dados desequilibrado resulta em estimativas otimistas, particularmente quando um classificador enviesado é empregado (BRODERSEN et al., 2010; KELLEHER; NAMEE; D'ARCY, 2015).

A acurácia balanceada, uma métrica de avaliação de desempenho de classificadores multi-classes, revela-se particularmente valiosa em cenários de desequilíbrio de classes, frequentemente encontrados em aplicações práticas (BRODERSEN et al., 2010).

A métrica de acurácia balanceada baseia-se nas abordagens amplamente reconhecidas de sensibilidade (taxa de verdadeiro positivo) e especificidade (taxa de falso positivo). A sensibilidade aborda a questão crucial de quantos casos positivos são corretamente identificados, enquanto a especificidade responde à mesma pergunta no contexto dos casos negativos. Este enfoque meticuloso na análise da acurácia balanceada visa proporcionar uma avaliação mais sólida e confiável do desempenho de algoritmos de classificação, especialmente em ambientes desafiadores e complexos (KELLEHER; NAMEE; D'ARCY, 2015). A Equação 2.9 exibe a maneira de calcular a acurácia balanceada.

$$AB = \frac{1}{N}\left(\sum_{i=1}^{N} \frac{VP_i}{VP_i + FN_i}\right) \tag{2.9}$$

onde,

N é o número de classes no banco de dados;

VP_{i} são as previsões corretas para a classe i;

FN_{i} são as previsões incorretas para a classe i.

A acurácia balanceada busca mitigar o viés introduzido pela distribuição desigual das classes ao calcular uma média das acurácias individuais de cada classe, conferindo peso igual a todas, independentemente do tamanho.

2.10.2 *Root Mean Squared Error* (RMSE)

A Raiz Quadrada do Erro Médio (RMSE) uma métrica de avaliação amplamente empregada no domínio de aprendizado de máquina e análise estatística. Esta métrica desempenha um papel instrumental na quantificação da acurácia e precisão de modelos preditivos, proporcionando uma medida robusta da discrepância entre os valores previstos e os observados (SHEKHAR; XIONG, 2017; REN et al., 2017).

A RMSE é uma métrica de desempenho particularmente relevante em problemas de regressão, onde a predição de valores contínuos é essencial. Seu cálculo, conforme estabelecido na literatura, é a raiz quadrada da média dos quadrados dos erros residuais entre as previsões do modelo e os valores reais. Matematicamente, a RMSE é expressa segundo a Equação 2.10 (CHAIN; DRAXLER, 2014; SHEKHAR; XIONG, 2017).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.10}$$

em que $\hat{y}_i$ é o valor estimado de y_i (valor observado).

A relevância da RMSE reside na sua capacidade de capturar tanto erros positivos quanto negativos, fornecendo uma visão abrangente da qualidade das previsões. Além disso, a natureza da raiz quadrada atenua a sensibilidade a valores extremos, conferindo uma robustez adicional à métrica (CHAIN; DRAXLER, 2014).

Ao explorar as nuances da RMSE, é imperativo contextualizar sua aplicação em diferentes domínios. Em análise estatística, a RMSE é frequentemente utilizada para avaliar a eficácia de modelos de previsão, enquanto em aprendizado de máquina, ela serve como uma ferramenta essencial na seleção e ajuste de modelos. Sua interpretação está intrinsecamente ligada à minimização do erro preditivo, refletindo a busca incessante por modelos que otimizem a concordância entre previsões e observações.

2.10.3 Coeficiente de Determinação (R^2)

O coeficiente de determinação (R^2), representa uma medida-chave que avalia a proporção da variabilidade na variável dependente que é explicada pelo modelo, oferecendo insights críticos sobre a adequação e eficácia do ajuste do modelo aos dados observados (GUJARATI; PORTER, 2008; CHICCO; WARRENS; JURMAN, 2021).

Matematicamente, o R^2 é calculado como a razão entre a variação explicada pelo modelo e a variação total na variável dependente. Esta métrica é expressa na Equação 2.11 (GUJARATI; PORTER, 2008).

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \tag{2.11}$$

em que $\hat{y}_i$ é o valor estimado de y_i é o valor real da variável dependente, $\hat{y}_i$ é o valor previsto pelo modelo e $\bar{y}$ é a média dos valores reais.

A relevância do R^2 reside na sua capacidade de quantificar a proporção da variabilidade total que é explicada pelo modelo, indicando assim a qualidade do ajuste. Um R^2 próximo de 1 sugere que o modelo explica a maioria da variação, enquanto um valor próximo de 0 indica que o modelo não oferece uma explicação significativa (GUJARATI; PORTER, 2008; CHICCO; WARRENS; JURMAN, 2021).

Explorar o R^2 implica contextualizá-lo em diferentes domínios. Em contextos estatísticos, o R^2 é uma ferramenta fundamental para aferir a adequação do modelo de regressão aos dados. No campo do aprendizado de máquina, o R^2 é crucial na avaliação do desempenho de modelos de previsão e no refinamento de técnicas de modelagem.

2.10.4 Teste Estatístico (ANOVA)

A Análise de Variância (ANOVA) é uma ferramenta estatística amplamente empregada na comparação de médias entre três ou mais grupos distintos. Sua aplicação abrange diversos domínios, como indústria, ciências sociais, medicina e engenharia, sendo fundamental na investigação científica. A análise estatística dessas variações é crucial para compreender o impacto de diferentes variáveis operacionais (MONTGOMERY, 2017; MICKEY; DUNN; CLARK, 2004).

A metodologia da ANOVA parte da decomposição da variação total dos dados em componentes atribuíveis às discrepâncias entre os grupos e aquelas atribuíveis às variações dentro dos grupos. Posteriormente, o teste estatístico verifica se as diferenças entre as médias dos grupos superam as esperadas devido ao acaso (MICKEY; DUNN; CLARK, 2004; SEBER; LEE, 2003).

Para avaliar a significância estatística das diferenças entre as médias, é necessário comparar o valor-p com um nível de significância predefinido, a fim de testar a hipótese nula, a qual afirma a igualdade das médias populacionais. Geralmente, um nível de significância (representado por α ou alfa) de 0,05 é adotado, indicando um risco de 5% de concluir erroneamente a existência de uma diferença real quando não há (MONTGOMERY, 2017; MICKEY; DUNN; CLARK, 2004; SEBER; LEE, 2003).

Se o valor-p for menor ou igual ao nível de significância, a hipótese nula é rejeitada, sugerindo que pelo menos uma média populacional difere das demais. Caso contrário, se o valor-p for maior que o

nível de significância, não há evidências suficientes para rejeitar a hipótese nula, indicando que as médias da população podem ser consideradas iguais (MONTGOMERY, 2017; MICKEY; DUNN; CLARK, 2004; SEBER; LEE, 2003).

2.11 Estado da Arte

Está presente na literatura algumas formas de avaliar as propriedades das plantas, dentre essas maneiras, a que mais se destaca é via índices de reflectância foliar. A correlação direta de propriedades das plantas com reflectância é pouco explorada.

No âmbito de avaliar as propriedades de plantas via índices o trabalho de (NUNES et al., 2023) explora variáveis espectrais para estimar o potencial hídrico de plantas de café utilizando abordagens de inteligência computacional. Mostrando que os índices PRI, NDVI, CRI1 e SIPI foram os mais relevantes para estimar e classificar o potencial hídrico do café. Foram implantados redes neurais (Multi-Layer Perceptron) e também árvore de decisão.

Por sua vez, o trabalho de (CHEMURA; MUTANGA; DUBE, 2017) investigou a capacidade de bandas de ondas selecionadas na faixa VIS/NIR para prever o conteúdo de água da planta no café, utilizando o algoritmo de floresta aleatória. A experimentação expôs as plantas de café a diferentes níveis de estresse hídrico e de refletância, com a medição do teor de água da planta. A análise identificou três bandas de ondas (485nm, 670nm e 885nm) como significativas, sendo treinadas e testadas em dados independentes para prever o conteúdo de água da planta de café. Os resultados indicaram que as bandas selecionadas de sensibilidade de refletância apresentaram o melhor desempenho na detecção de estresse hídrico. A robustez dessas descobertas sugere a viabilidade de prever o conteúdo de água da planta usando bandas de ondas VIS/NIR por meio de algoritmos de floresta aleatória, embora sejam necessárias mais pesquisas em escala de campo e paisagem para operacionalizar essas conclusões.

Visando ainda explorar as condições hídricas em uma plantação de café o trabalho de (SANTOS et al., 2022) emprega técnicas de agricultura de precisão (AP) em conjunto com geoestatística e imagens de alta resolução. O potencial hídrico foliar, obtido via bomba Scholander, foi espacializado e interpolado por meio de análise geoestatística. Os índices de vegetação derivados das imagens da RPA foram calculados para análise de regressão e correlação em conjunto com os dados de potencial hídrico. O grau de dependência espacial (GDS) dos dados geoestatísticos revelou forte dependência espacial em ambos os períodos. Na análise de correlação e regressão linear, apenas a faixa vermelha demonstrou correlação significativa. A

geoestatística se mostrou crucial para a espacialização do potencial hídrico, enquanto os índices de vegetação provenientes da RPA exibiram eficácia limitada na avaliação das condições hídricas dos cafeeiros.

Tradicionalmente estimar o potencial hídrico foliar em áreas cafeeiras ocorre de maneira direta. O que proporciona o surgimento de uma demanda por tecnologias capazes de estimar grandes áreas ou regiões de forma indireta. O trabalho (MACIEL et al., 2020) utilizando valores de refletância superficial e índices de vegetação provenientes do sensor de imagens *Landsat-8/OLI* em Minas Gerais, Brasil, em que diferentes algoritmos foram avaliados, destacando-se um modelo quadrático com o Índice de Vegetação por Diferença Normalizada (NDVI), implementado juntamente com validação cruzada. A aplicação deste modelo às imagens *Landsat-8/OLI* permitiu a estimativa do (Ψ_W) em uma área representativa do cultivo de café, contribuindo para o monitoramento eficaz da seca e redução de custos para os produtores.

No trabalho de (TOSIN et al., 2020), foi explorado dados hiperespectrais coletados por um espectrorradiômetro portátil para avaliar o potencial hídrico foliar da videira em condições de verão quente e seco. Três cultivares de videira foram examinados em vinhas de sequeiro e de irrigado. Utilizou-se um modelo de regressão logística ordinal com base em índices de vegetação e variáveis estruturais, em que foi identificado quatro preditores significativos: tratamento de irrigação, local de teste, índice de refletância de antocianina otimizado e índice de razão normalizada.

Na pesquisa de (ZAKALUK; RANJAN, 2008), a eficácia de uma câmera digital RGB de 5 megapixels, montada em um poste telescópico de 2,5 m, para a determinação do potencial hídrico foliar (Ψ_W) de plantas de batata no campo, por meio da modelagem de uma Rede Neural Artificial (RNA), foi explorada. Um desenho amostral sistemático em uma grade aleatória de 45 x 45m foi empregado, com a medição aleatória de parcelas amostrais para a obtenção de informações sobre o potencial hídrico foliar, teor de nitrato, teor volumétrico de água e a captura de imagens digitais. As imagens foram radiometricamente calibradas e classificadas para isolar a folhagem verde, distinguindo-a de solo, flores, sombras e folhas senescentes. Além das imagens RGB, seis transformações de imagem e nove índices de vegetação foram avaliados como candidatos a neurônios de entrada para a modelagem de RNA. Neurônios de entrada com colinearidade significativa foram submetidos à análise de componentes principais (PCA). Uma relação linear entre o nitrato do solo e a componente verde da folhagem, através da banda verde da imagem, foi considerada significativa.

Com base nas referências mencionadas, é evidente que a pesquisa em reflectância espectral aplicada à agricultura está atualmente em destaque. Isso abre caminho para o desenvolvimento de diversas investigações nesse contexto, especialmente no âmbito da cafeicultura, que constitui o foco central deste estudo.

3 MATERIAIS E MÉTODOS

No presente capitulo será exibido as técnicas, os equipamentos necessários para a coleta dos dados e insumos importantes para todo desenvolvimento do projeto.

3.1 Metodologia

O estudo adotou uma abordagem metódica que integrou tanto técnicas de regressão quanto de classificação, visando uma análise holística dos dados. Para realizar tal análise, os dados foram inicialmente organizados manualmente, padronizando-os matricialmente e normalizando-os para mitigar possíveis distorções causadas por valores discrepantes. Em seguida, uma seleção criteriosa de características foi executada para identificar os atributos mais relevantes a serem utilizados nos sistemas. A validação dos estimadores foi realizada por meio de duas técnicas distintas: validação cruzada *K-folds* e *Holdout*, garantindo a robustez dos resultados. Para a tarefa de regressão e classificação, uma variedade de estimadores foi aplicada e comparada. Uma arquitetura de redes neurais artificiais foi projetada e refinada para otimização, enquanto estimadores baseados em árvores de decisão, *Random Forest* e KNN *(k-nearest neighbors)* foram desenvolvidos e testados. O estudo culminou com uma análise comparativa criteriosa para validar e identificar os melhores resultados alcançados por cada técnica empregada. Para a execução do trabalho, seguindo-se as etapas:

1 - Organização manual dos dados de forma a padronizá-los matricialmente, possibilitando seu uso na plataforma de desenvolvimento de algoritmos *MatLab* e normalizá-los para evitar influência marcante de variáveis com valores discrepantes;

2 - Execução de uma seleção de características, visando determinar os melhores atributos a serem aplicados nas entradas dos sistemas;

3 - Implementação dos métodos de validação cruzada *K-folds* e *Holdout*;

4 - Projeto de uma arquitetura de redes neurais artificiais para a regressão e classificação, testando diferentes cenários e selecionando a que obteve os melhores resultados;

5 - Projeto de um estimador baseado em árvores de decisão para a regressão e classificação, com variação de parâmetros;

6 - Projeto de um estimador baseado em *Random Forest* para a regressão e classificação, com variação de parâmetros;

7 - Projeto de um estimador baseado em KNN *(k-nearest neighbors)* para a regressão e classificação, com variação de parâmetros;

8 - Validação dos resultados por meio de um estudo comparativo via teste estatístico ANOVA.

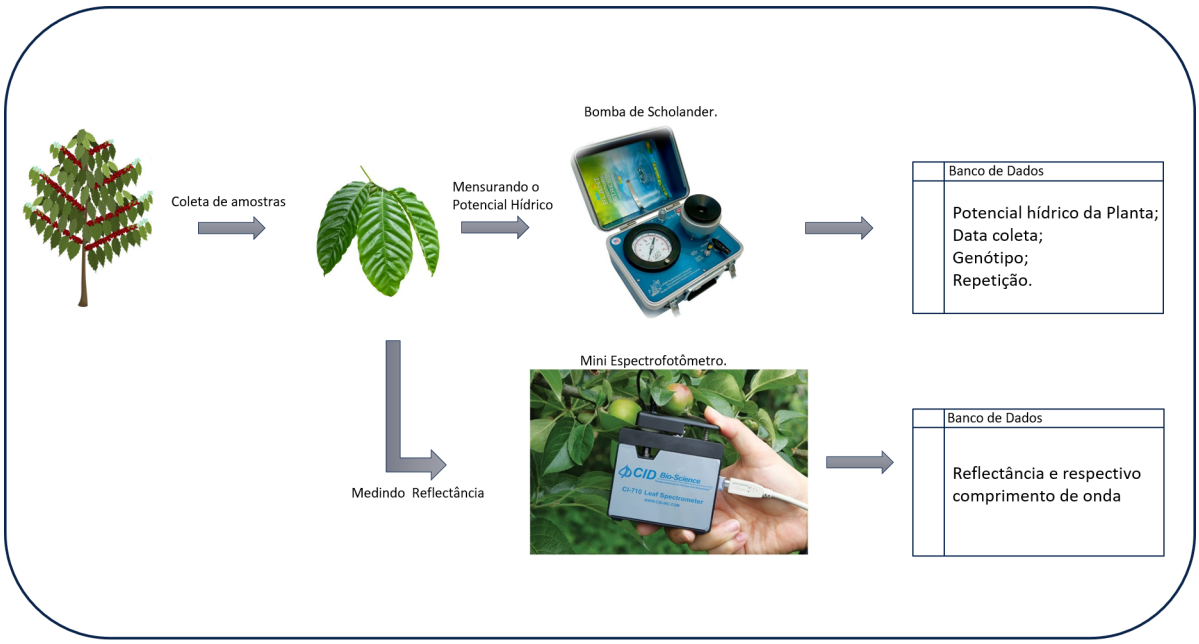
3.2 Base de Dados

Os dados foram coletados em diferentes datas (2014, 2015 e 2016), visando capturar o efeito de variações climáticas sazonais da região do município de Diamantina, localizada no norte do estado de Minas Gerais, e também de cafeeiros com dois tipos de manejo, irrigado e sequeiro. Os dados com característica de irrigado são originários de cafeeiros que foram submetidos a métodos de irrigação, de modo que a água é fornecida ao plantio de maneira artificial, cujo principal objetivo é viabilizar o cultivo. Considerando o manejo sequeiro, os cafeeiros não foram expostos a irrigação artificial, ficando sujeito apenas a hidratação resultante das precipitações naturais. A base de dados foi fornecida pela equipe de pesquisadores de campo da EPAMIG (Empresa de Pesquisa Agropecuária de Minas Gerais). A Figura 3.1 exibe o fluxograma da coleta os dados.

De acordo com a Figura 3.1, inicialmente a amostra é colhida e passa por duas medições, a primeira via Bomba de Scholander, a qual determina o potencial hídrico e gera um banco de dados que é incrementado com a data de coleta, genótipo e repetição da amostra. Logo em seguida, a amostra passa pelo equipamento mini espectrômetro foliar CI-710 que fornece a características espectrais do cafeeiro.

O Mini Espectrofotômetro de Folha CI-710 é um instrumento crucial na análise não invasiva da fisiologia foliar de plantas. Este espectrofotômetro oferece uma medição precisa da absorção de luz pelas folhas em uma ampla faixa espectral de 400 nm a 1000 nm, permitindo a quantificação de pigmentos como clorofila e carotenoides, fundamentais para a fotossíntese e processos bioquímicos foliares. Com uma resolução espectral de 1 nm e uma taxa de aquisição de dados de até 1000 pontos por segundo, o CI-710 oferece detalhes finos das respostas espectrais das folhas. Além disso, sua operação não invasiva preserva a integridade das amostras, tornando-o ideal para estudos longitudinais e a avaliação do estado fisiológico das plantas ao longo do tempo. Com uma portabilidade notável e recarga da bateria via USB, o CI-710 é facilmente transportável e pode ser integrado a diferentes ambientes de pesquisa. Sua versatilidade e precisão tornam-no essencial

Figura 3.1 – Coleta de Dados.



Fonte: Do Autor (2023)

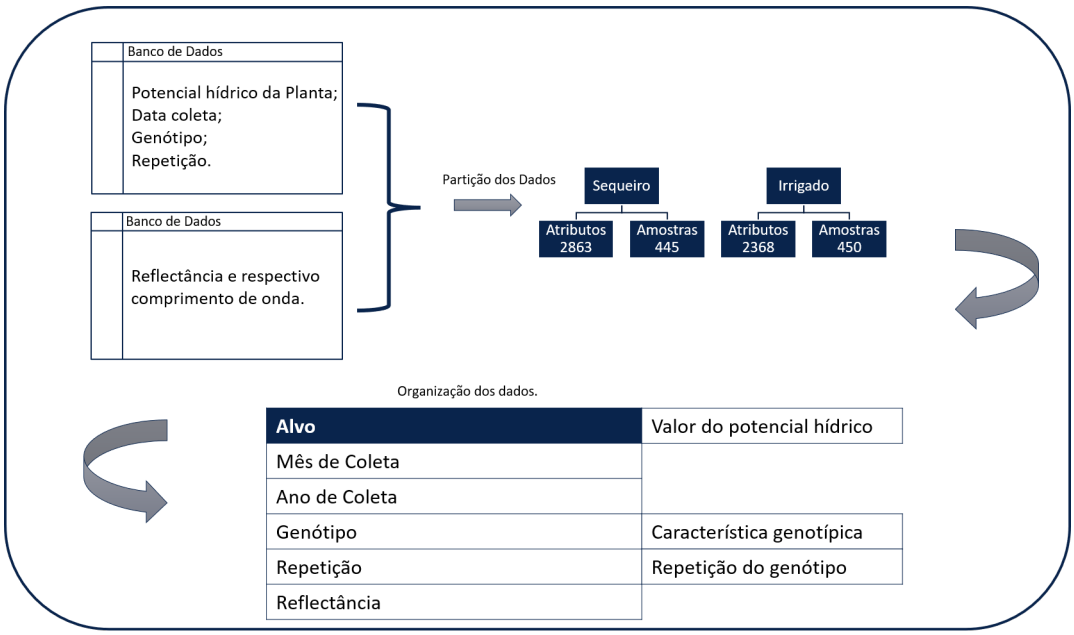
em pesquisas em fisiologia vegetal, agricultura e estudos sobre respostas das plantas a condições ambientais variadas.

O número de amostras armazenadas no banco de dados é de 450 irrigado e 445 sequeiro, os dois tipos de manejo estão separados em dois bancos de dados distintos e não foi implementado a mescla dos bancos. A Figura 3.2 ilustra o processo de estruturação dos dados. Após a aquisição da base de dados, um trabalho de organização foi efetuado de modo que cada amostra é composta por 2863 atributos, os quais são formados pela data de coleta, o número do genótipo (que representa a característica genética da planta como, espessura da parede celular, sensibilidade estomática e etc..), o número da repetições dentro do genótipo e a sequência da reflectância que corresponde ao comprimento de onda na faixa de 400 á 950 nm. Além do mais, ambas as bases de dados possuem como alvo o potencial hídrico (Ψ_W) medido com uma Bomba de Scholander. A representação da base de dados é mostrada no Apêndice C.

3.3 Pré-processamento

Para o início do desenvolvimento foi necessário o pré-processamento dos dados. A Figura 3.3 exibe o fluxograma das etapas do pré-processamento. Os dados foram alocados em forma matricial, posteriormente, via analise de *Boxplot* foram eliminados os *outliers*, o método de filtragem por mediana foi implementado

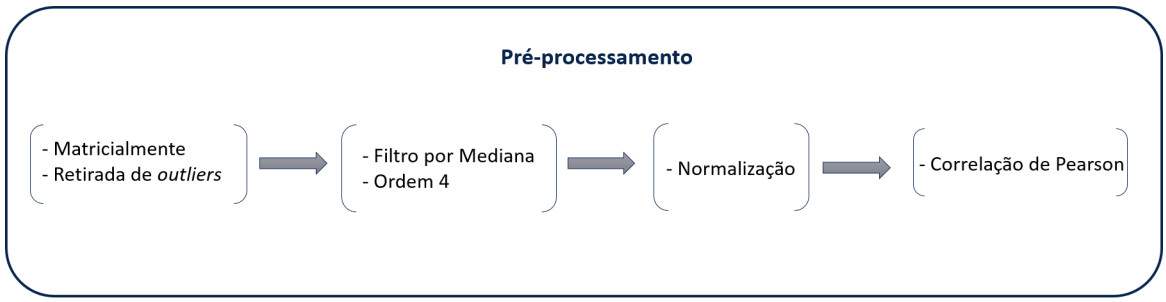
Figura 3.2 – Organização dos Dados.



Fonte: Do Autor (2023)

com o intuito de minimizar ruídos e em seguida o processo de normalização das amostras foi executado, de modo a evitar eventuais priorizações de dados considerando apenas seu valor absoluto mais elevado em relação aos demais (BRAGA; CARVALHO; LUDERMIR, 2007).

Figura 3.3 – Pré-Processamento.



Fonte: Do Autor (2023)

Dessa maneira, foi adotado o método de filtragem por mediana desenvolvida por (TUKEY, 1977), que consiste em suavizar o ruído do tipo impulsivo em sinais e imagens digitais (PRATT, 2001). O filtro por mediana atua determinando uma janela de ação de N amostras. Posteriormente os N valores são dispostos em ordem crescente a mediana é o valor que foi ordenado bem no centro da amostra e o filtro por mediana

substitui o valor central. A plataforma de desenvolvimento de algoritmo *MatLab* dispõe de comando *medfilt1* que proporciona a implementação do filtro por mediana. O filtro implementado no trabalho foi de ordem 4.

Em seguida, a regra de normalização adotada para o conjunto de dados é de escala [0, 1]. Para a realização desse procedimento foi utilizada a Equação (3.1), em que P_n é o valor normalizado da variável n, P é o valor original, P_{min} é o menor valor dentre os valores do banco de dados referente à variável n e P_{max} é o maior valor.

$$P_n = \frac{(P - P_{min})}{(P_{max}) - (P_{min})} \tag{3.1}$$

Outro ponto importante do pré-processamento é a seleção de características. Para esta etapa foi utilizado a técnica do coeficiente de *Pearson*, que mede o grau da correlação linear entre duas variáveis de escala métrica. Este coeficiente, normalmente representado por ρ, assume apenas valores entre -1 e 1. A Tabela 3.1 exibe a interpretação dos valores do coeficiente de *Pearson* (ρ) (VIEIRA, 2008).

Tabela 3.1 – Valores de coeficiente de correlação (ρ).

Valor de ρ (+ ou -)	**Interpretação**
0,00 a 0,19	Correlação muito fraca
0,20 a 0,39	Correlação fraca
0,40 a 0,69	Correlação moderada
0,70 a 0,89	Correlação forte
0,90 a 1,00	Correlação muito forte

Para implementação das técnicas de *Machine Learning* visando a classificação foi necessário estabelecer classes para os potenciais hídricos das amostras. Isso posto, as classes foram estipuladas segundo o trabalho de (SILVA et al., 2021) e são mostradas na Tabela 3.2.

Tabela 3.2 – Classes e faixas de potencial hídrico.

Valores do Potencial Hídrico (Ψ_W) (MPa)	**Classe**
Ψ_W até -0,5 MPa	1
Ψ_W entre -0,5 e -1,4 MPa	2
Ψ_W entre -1,5 e -2,4 MPa	3
Ψ_W entre -2,5 e -3,5 MPa	4
Ψ_W menores que -3,5 MPa	5

Realizada esta etapa, foi promovida uma divisão dos dados em um conjunto de treino e teste. A partição é feita utilizando a técnica de Validação Cruzada *K-fold*, em que o número de *folds* escolhido foi 10, de forma que para cada *fold* escolhido para teste os outros 9 são selecionados para treino.

3.3.1 Avaliação de Desempenho

Com o objetivo de avaliar o desempenho dos classificadores, foi utilizado a métrica da acurácia balanceada que avalia o desempenho de modelos em conjuntos de dados desbalanceados, onde as classes possuem quantidades diferentes de amostras. A acurácia balanceada busca mitigar o viés introduzido pela distribuição desigual das classes ao calcular uma média das acurácias individuais de cada classe, conferindo peso igual a todas, independentemente do tamanho (BRODERSEN et al., 2010; KELLEHER; NAMEE; D'ARCY, 2015).

Para as técnicas de regressão foi utilizada a raiz do erro quadrático médio, conhecida como *Root Mean Squared Error* (RMSE), que tem uma escala dependente dos dados e maior penalização para maiores erros (CHAIN; DRAXLER, 2014; SHEKHAR; XIONG, 2017; REN et al., 2017). Também foi empregada a métrica coeficiente de determinação, também conhecido como R^2, podendo ser compreendido como o percentual da variância dos dados que é explicada pelo modelo. Assim, quanto maior o R^2, mais explicativo é o modelo, ou seja, melhor ele se ajusta à amostra (CHICCO; WARRENS; JURMAN, 2021; GUJARATI; PORTER, 2008).

Por fim, a implementação da Análise de Variância (ANOVA), uma técnica estatística sofisticada e robusta utilizada para examinar as variações em dados experimentais, particularmente na comparação de médias entre grupos distintos. A ANOVA é enraizada nos princípios da decomposição da variabilidade total, delineando meticulosamente as fontes de variação e permitindo inferências sobre a existência de diferenças significativas entre as médias dos grupos sob escrutínio. A média das métricas de desempenho das quatro técnicas de *Machine Learning* foram submetidas ao teste ANOVA (MONTGOMERY, 2017; MICKEY; DUNN; CLARK, 2004; SEBER; LEE, 2003).

4 RESULTADOS E DISCUSSÕES

O presente capítulo expõe a análise detalhada dos resultados obtidos a partir da aplicação das metodologias propostas. A análise dos dados é apresentada de forma a evidenciar as descobertas, padrões e nuances revelados durante a pesquisa. São abordados os resultados obtidos em diferentes etapas do estudo, desde a coleta e preparação dos dados até a aplicação e avaliação das técnicas *Machine Learning* utilizadas.

4.1 Analise dos Dados

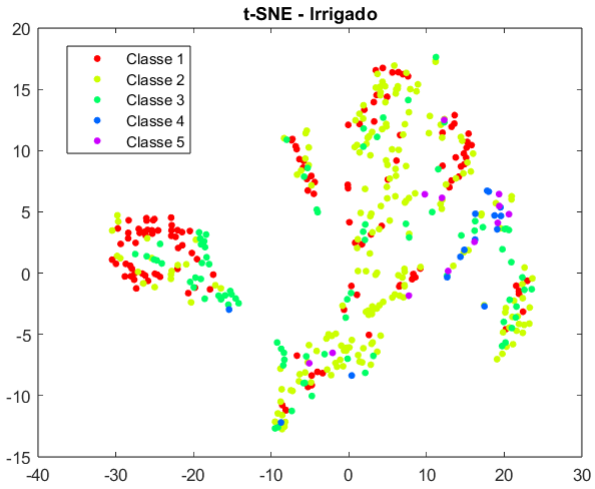
Conhecer a distribuição dos dados é crucial com o intuito de compreender a complexidade do problema. A maneira como estão distribuídos pode auxiliar na escolha e no ajuste adequado dos modelos de aprendizado de máquina. Entender a distribuição dos dados permite identificar padrões, tendências e anomalias. Outro ponto importante a se discutir é a respeito de bases de dados desbalanceadas que podem levar a modelos tendenciosos. Conhecer a distribuição das classes ou categorias-alvo ajuda a lidar com esse desbalanceamento, aplicando técnicas específicas para melhorar a performance dos modelos.

Com o intuído de conhecer melhor a distribuição dos dados a técnica t-SNE (*t-distributed Stochastic Neighbor Embedding*) é um alternativa viável de análise. O algoritmo consiste em uma redução de dimensionalidade usado principalmente para visualizar conjuntos de dados de alta dimensionalidade em duas ou três dimensões, de forma a preservar as relações entre os pontos (MAATEN; HINTON, 2009).

O t-SNE é especialmente útil para visualizar agrupamentos ou padrões em conjuntos de dados complexos. Ele funciona mapeando pontos de dados de alta dimensão para um espaço de menor dimensão, onde pontos similares são mapeados próximos uns dos outros, enquanto pontos menos similares são mapeados mais distantes. Ao reduzir a dimensionalidade dos dados para duas ou três dimensões de forma a preservar as relações entre os pontos, o t-SNE permite uma representação visual dos dados que pode revelar agrupamentos, clusters ou padrões que não seriam facilmente identificados em dimensões mais elevadas, tornando o t-SNE factível para o problema em questão, uma vez que a base de dados usada neste trabalho abrange 2863 dimensões. A Figura 4.1 e a Figura 4.2 exibem o t-SNE para os dados em estado irrigado e sequeiro, respectivamente.

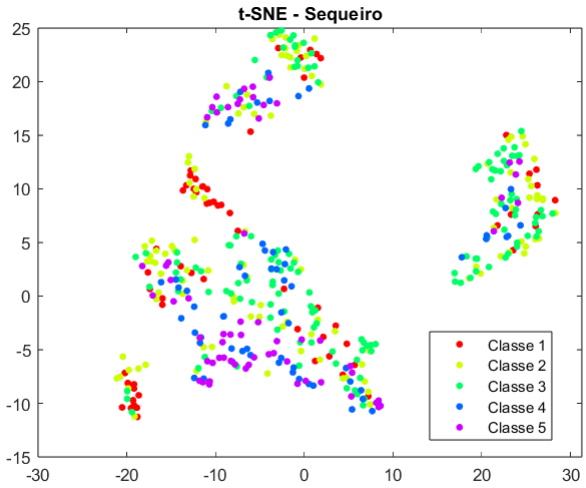
Analisando a Figura 4.1 e 4.2 é possível notar as adversidades de desbalanceamento das classes. Quando treinados com conjuntos desbalanceados, modelos tendem a demonstrar melhor desempenho na classificação das amostras da classe majoritária, porém um desempenho inferior para as amostras da classe minoritária.

Figura 4.1 – t-SNE Irrigado.



Fonte: Do Autor (2023)

Figura 4.2 – t-SNE Sequeiro.



Fonte: Do Autor (2023)

Na Figura 4.1 e 4.2 a sobreposição de dados é outra característica presente nos conjuntos de dados abordados e que está fortemente ligada à queda no desempenho de modelos de aprendizado de máquina. A sobreposição ocorre quando exemplos de classes diferentes estão “misturados” em determinadas regiões do espaço dimensional, dificultando o estabelecimento de um limite de decisão confiável que separe as classes de maneira eficiente. Isso vai exigir o uso de classificadores mais complexos e capazes de gerar fronteiras de separação complexas e não lineares.

4.2 Seleção de Variáveis

A divisão dos dados separa as amostras em sequeiro e irrigado, sem mescla e levando em conta todas as amostras colhidas. Considerando o modo de divisão por manejo o coeficiente de *Pearson* (ρ) foi calculado entre os atributos e a variável alvo (potencial hídrico). Inicialmente para os dados em condição de irrigado, atendendo o método de regressão adotou-se o limiar de $\rho = 0,21$. Considerando a técnica de classificação o coeficiente de *Pearson* adotado foi de $\rho = 0,27$, ou seja, atributos com $\rho < 0,21$ para regressão e $\rho < 0,27$ para classificação, foram considerados irrelevantes para o modelo. Em suma, os atributos passaram de 2863 dimensões para 15 e 22, para regressão e classificação, respectivamente.

Já para os dados contidos no manejo sequeiro, o coeficiente de *Pearson* (ρ) teve seu valor selecionado em $\rho = 0,42$ para ambos os métodos (regressão e classificação), o que representa uma correlação moderada segundo a Tabela 3.1. Os atributos passaram de 2863 dimensões para 20 e 21 atendendo regressão e classificação, respectivamente. Por fim, com o intuito de exemplificar os 10 atributos mais relevantes para ambas as condições são mostrados na Tabela 4.1 e Tabela 4.2, para manejo irrigado e sequeiro.

Tabela 4.1 – Os 10 atributos mais relevantes para irrigado.

***Ranking* dos atributos**	***Designação do atributo***	
	Regressão	**Classificação**
1	mês de coleta	mês de coleta
2	reflectância para λ = 779,751 nm	reflectância para λ = 780,123 nm
3	ano de coleta	ano de coleta
4	reflectância para λ = 784,944 nm	reflectância para λ = 784,944 nm
5	reflectância para λ = 783,091 nm	reflectância para λ = 783,462 nm
6	reflectância para λ = 781,051 nm	reflectância para λ = 779,380 nm
7	reflectância para λ = 779,380 nm	reflectância para λ = 782,906 nm
8	reflectância para λ = 784,574 nm	reflectância para λ = 781,051 nm
9	reflectância para λ = 782,906 nm	reflectância para λ = 779,937 nm
10	reflectância para λ = 779,566 nm	reflectância para λ = 784,759 nm

em que (λ) corresponde a comprimento de onda.

Fonte: Próprio Autor.

Todos os atributos mais relevantes extraídos após a aplicação do *Pearson* são apresentados no Apêndice A.

Os resultados obtidos através das Tabelas 4.1 e 4.2 exibem como faixas significativas comprimentos de onda em torno de 690nm, uma vez que cafeeiros em condição sequeiro apresentam como atributos mais relevantes reflectâncias correspondentes a esse comprimentos de onda, já para manejo irrigado as reflectâncias correspondentes 780nm de comprimento de onda são as mais úteis.

Tabela 4.2 – Os 10 atributos mais relevantes para sequeiro.

***Ranking* dos atributos**	***Designação do atributo***	
	Regressão	**Classificação**
1	mês de coleta	mês de coleta
2	reflectância para λ = 690,885 nm	reflectância para λ = 690,498 nm
3	reflectância para λ = 694,167 nm	reflectância para λ = 692,430 nm
4	reflectância para λ = 692,816 nm	reflectância para λ = 688,758 nm
5	reflectância para λ = 691,271 nm	reflectância para λ = 691,657 nm
6	reflectância para λ = 690,691 nm	reflectância para λ = 689,725 nm
7	reflectância para λ = 693,202 nm	reflectância para λ = 691,464 nm
8	reflectância para λ = 691,464 nm	reflectância para λ = 689,145 nm
9	reflectância para λ = 692,044 nm	reflectância para λ = 691,851 nm
10	reflectância para λ = 691,078 nm	reflectância para λ = 691,078 nm

em que (λ) corresponde a comprimento de onda.

Fonte: Próprio Autor.

Após a execução das etapas de pré-processamento e seleção de características, deu-se inicio a implementação dos modelos de *Machine Learning*.

4.3 Divisão por Manejo

Inicialmente, as técnicas Rede Neural Artificial, Árvore de Decisão, *Random Forest* e *KNN*[1] foram executadas. É importante salientar que todas as técnicas foram implementadas para todos os atributos após seleção de características via coeficiente de *Pearson*. Diferentes parâmetros para uma mesma técnica foram implementados com o intuito de otimizar e fornecer o melhor resultado.

De modo aprimorar o desempenho dos modelos de *Machine Learning*, o ajuste dos hiperparâmetros procura encontrar a combinação ideal que otimize o desempenho do modelo. O presente trabalho adotou alteração de hiperparâmetro de maneira automática. Na plataforma de desenvolvimento de algoritmo *MatLab* a otimização de hiperparâmetros em redes neurais pode abranger parâmetros como número de camadas que pode variar de 1 a 3, número de neurônios por camada (altera de 1 a 300) e função de ativação que pode variar em "*relu*" representa a função de ativação *Rectified Linear Unit*, "*tanh*" é a função tangente hiperbólica, "*sigmoid*" é a função logística e "*none*" indica que nenhuma função de ativação será aplicada, o que pode ser relevante para certos casos ou arquiteturas de rede específicas. O algoritmo utilizado para treinamento foi o *Quasi-Newton*, *Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm* (LBFGS), o qual usa um método de busca linear padrão.

[1] Levando em conta o método de regressão para técnica *KNN* foi utilizado o algoritmo presente no trabalho reportado em (CONSONNI et al., 2021).

Considerando a técnica árvores de decisão, os parâmetros que podem ser alterados com a finbalidade de melhorar o desempenho são; o número máximo de divisões de decisão (ou nós de ramificação), número mínimo de amostras por folha e o critério de divisão, o qual é usado para construir a árvore de decisão. Para problemas com duas classes, é utilizado "*gdi*" para o índice de diversidade de Gini e "*deviance*" conhecida como entropia cruzada. Para três ou mais classes, também é considerado o "*twoing*" (regra de "*twoing*"). Esses critérios determinam como a árvore é dividida em cada nó, influenciando diretamente a construção da estrutura da árvore.

Para a *Random Forest*, a otimização inclui se houver três ou mais classes no problema, os métodos elegíveis são "*Bag*", "*AdaBoostM2*" e "*RUSBoost*", o número de ciclos de aprendizado (número de iterações), por padrão, a busca ocorre entre números inteiros positivos, normalmente distribuídos de maneira logarítmica, variando de 10 a 500 ciclos de aprendizado e a taxa de aprendizado para algoritmos que busca entre números reais positivos, normalmente distribuídos de maneira logarítmica, variando de 0,001 (ou 1e-3) a 1.

Por fim, no KNN, os parâmetros a serem ajustados são; distância, que define a métrica de distância usada para calcular a proximidade entre os vizinhos, que são; "*cityblock*", "*chebychev*", "*correlation*", "*cosine*", "*euclidean*", "*hamming*", "*jaccard*", "*mahalanobis*", "*minkowski*", "*seuclidean*" e "*spearman*". Cada uma dessas métricas de distância possui suas próprias características. Outro parâmetro é o método de ponderação da influência dos vizinhos com base na distância, onde há três opções: "*equal*" (igual), "*inverse*" (inverso) e "*squaredinverse*" (inverso quadrado). O número de vizinhos também é um parâmetro a ser considerado, a busca ocorre entre valores inteiros positivos, normalmente distribuídos de maneira logarítmica, por padrão, variando de 1 a um valor determinado baseado no número de observações do conjunto de dados. Cada *fold* dispôs de 30 execuções, totalizando 300 iterações por método de *Machine Learning*.

4.3.1 Cafeeiro Irrigado

4.3.1.1 Regressão

Foram extraídos os valores da raiz do erro quadrático médio (*RMSE*) e do coeficiente de determinação R^2 dos 10 *folds* para as técnicas de *Machine Learning* empregadas na aplicação de regressão. Os resultados da média (μ) e do desvio padrão (σ) das métricas de avaliação são exibidos na Tabela 4.3, Tabela 4.4 e Tabela 4.5, para os 5, 10 e 15 atributos mais relevantes, respectivamente.

Tabela 4.3 – Divisão por Manejo, os 05 atributos mais relevantes para irrigado.

Técnica	**Regressão**			
	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	0,5830	±0,1114	0,5272	±0,1175
Árvore de Decisão	0,5962	±0,1224	0,4988	±0,1380
Random Forest	0,5900	±0,1005	0,5110	±0,1017
KNN	0,6442	±0,1205	0,4348	±0,1178

Fonte: Do Autor (2023).

Tabela 4.4 – Divisão por Manejo, os 10 atributos mais relevantes para irrigado.

Técnica	**Regressão**			
	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	0,6144	±0,1412	0,4758	±0,1866
Árvore de Decisão	0,5806	±0,1095	0,5254	±0,1195
Random Forest	0,5913	±0,1094	0,5107	±0,1150
KNN	0,6237	±0,1111	0,4884	±0,1058

Fonte: Do Autor (2023).

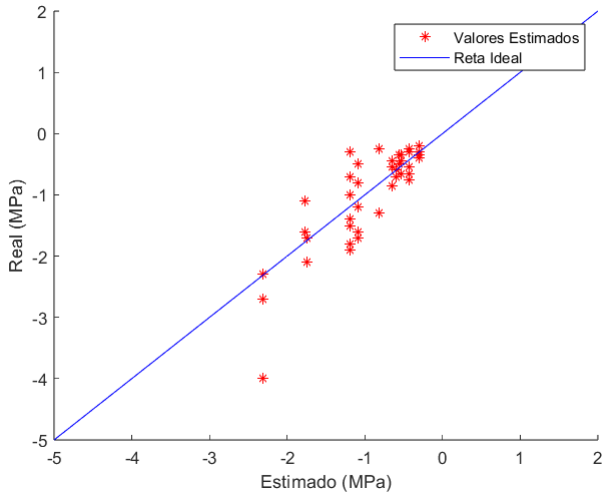
Tabela 4.5 – Divisão por Manejo, todos atributos mais relevantes para irrigado.

Técnica	**Regressão**			
	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	0,5934	±0,1127	0,4996	±0,1286
Árvore de Decisão	0,5785	±0,1063	0,5253	±0,1197
Random Forest	0,5873	±0,1012	0,5158	±0,1159
KNN	0,6199	±0,0950	0,4663	±0,0956

Fonte: Do Autor (2023).

Analisando os resultados da Tabela 4.3, Tabela 4.4 e Tabela 4.5, nota-se que a Árvore de Decisão alcançou o melhor resultado dentre as 3 opções de atributos, com o valor 0,5785 ±0,1012 de média da raiz do erro quadrático (*RMSE*) e 0,5253 ±0,1197 para coeficiente de determinação (R^2), valores esses correspondentes a organização com os 15 atributos mais relevantes levando em conta o método de regressão. Considerando o *fold* de melhor desempenho entre os 10 da técnica melhor qualificada, os parâmetros obtidos são número de ramificações de 35, valor mínimo de observações por folha de 19 e critério de divisão Índice Gini. Com o intuito de edificar a exibição dos resultados obtidos e possibilitar uma melhor análise, a Figura 4.3 ilustra os valores preditos no *fold* de melhor desempenho em relação a reta ideal, em que quanto mais distantes os dados estimados estão da reta, maiores os erros associados aos dados em questão. Para este caso, o *RMSE* é de 0,4342 e o coeficiente de determinação (R^2) de 0,6993.

Figura 4.3 – Dados Reais x Dados Estimados (*fold* de melhor desempenho - Irrigado - 15 atributos mais relevantes).



Fonte: Do Autor (2023).

Considerando os resultados para o manejo irrigado e o método de regressão. O teste ANOVA foi aplicado para as métricas de desempenho dos 10 *folds* de todas as técnicas de *Machine Learning*. O α = 0,05 de nível de significância foi adotado, para o *RMSE* o valor-p = 0,9762 e R^2 o valor-p = 0,7562. Nestes resultados, a hipótese nula afirma que os valores médios do *RMSE* são iguais e os do R^2 também são iguais. Como o valor de p é maior que o nível de significância de 0,05, isso indica que não há evidências estatísticas suficientes para rejeitar a hipótese nula. Nesse caso, não há diferenças estatisticamente significativas entre as médias dos grupos testados.

4.3.1.2 Classificação

Para o procedimento de classificação, a métrica utilizada é acurácia balanceada, que oferece uma visão mais realista do desempenho do modelo de *Machine Learning* para cada classe do banco de dados. Os resultados da média (μ) e do desvio padrão (σ) dos 10 *folds* são exibidos nas Tabelas 4.6, 4.7 e 4.8.

A Árvore de Decisão alimentada com os 10 atributos mais relevantes demonstra os melhores resultados com média de acerto de 50,64% $\pm$4,59%. O *fold* de melhor resultado levando em conta a técnica de melhor desempenho possui como parâmetro o número de ramificações de 9, valor mínimo de observações por folha de 7 e critério de divisão *twoing*. A Figura 4.4 exibe a matriz de confusão do *fold* 3 com melhor desempenho.

Tabela 4.6 – Divisão por Manejo, os 05 atributos mais relevantes para irrigado.

Técnica	**Classificação**	
	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	47,09	±11,27
Árvore de Decisão	49,91	±14,12
Random Forest	46,52	±10,62
KNN	44,42	±9,79

Fonte: Do Autor (2023).

Tabela 4.7 – Divisão por Manejo, os 10 atributos mais relevantes para irrigado.

Técnica	**Classificação**	
	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	46,48	±10,62
Árvore de Decisão	50,64	±8,11
Random Forest	44,59	±12,88
KNN	46,47	±10,03

Fonte: Do Autor (2023).

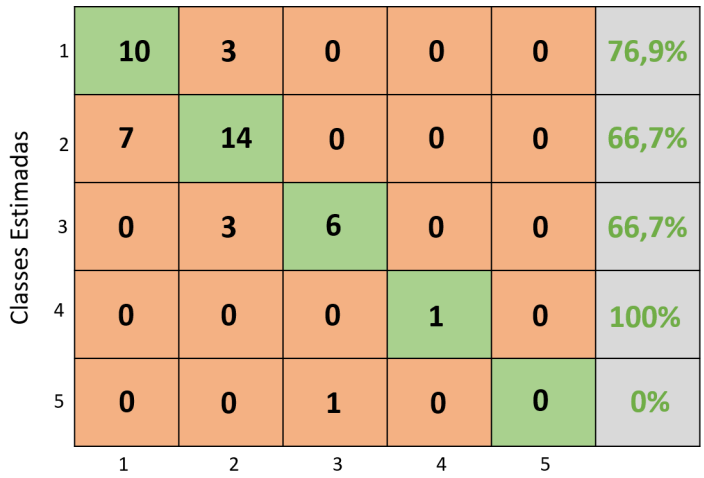
Tabela 4.8 – Divisão por Manejo, todos atributos mais relevantes para irrigado.

Técnica	**Classificação**	
	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	39,78	±6,87
Árvore de Decisão	49,18	±13,26
Random Forest	40,06	±9,49
KNN	40,73	±6,37

Fonte: Do Autor (2023).

Por meio da análise da matriz de confusão exibida na Figura 4.4 a acurácia balanceada do modelo para o *fold* 3 é de 62,05%. É fácil constatar que a maioria dos dados são pertencentes as classes 1, 2 e 3, validando o melhor desempenho do classificador nesses eventos. Levando em consideração a classe 5, o algoritmo não foi capaz de classificar corretamente nenhuma amostra que pertencesse a essa classe. O desbalanceamento de amostras faz com que o algoritmo não seja capaz de aprender os padrões representativos dessa classe. Como resultado o cálculo da proporção ou taxa de acerto para a classe 5 envolve a divisão do número de previsões corretas pela contagem total de amostras da classe. Convertendo os números das classes para valores de potencial hídrico via Tabela 3.2, é perceptível que valores acima de -2,5MPa contem as amostras pertencentes as classes 4 e 5 com menor número de amostras. O desbalanceamento é exibido na Tabela 4.9 para os dados de treino levando em conta o manejo irrigado.

Figura 4.4 – Matriz de confusão 10 mais relevantes (*fold* 3 - melhor desempenho - Irrigado).



Fonte: Do Autor (2023).

Tabela 4.9 – Número de amostras por classe, dados de treino manejo irrigado.

Classe	Número de amostras
1	117
2	187
3	77
4	14
5	10

Fonte: Do Autor (2023).

O desbalanceamento de classes na base treino afeta o desempenho do modelo, pois devido a esse desequilíbrio o modelo não é capaz de aprender de maneira efetiva acarretando em um desempenho não satisfatório.

O teste ANOVA para o método de classificação adotou os mesmo parâmetros de regressão. O valor-p para a acurácia balanceada é igual a 0,2672. O presente resultado indica que não há evidências estatísticas suficientes para afirmar que existe diferenças entre as médias dos grupos testados.

4.3.2 Sequeiro

De maneira análoga às amostras em condição irrigada, foram apresentadas aos dados em estado sequeiro as técnicas de *Machine Learning*.

4.3.2.1 Regressão

Foram considerados os 5 e 10 atributos mais relevantes, posteriormente todos os 20 atributos. A Tabela 4.10, Tabela 4.11 e Tabela 4.12 exibem a média (μ) e o desvio padrão (σ) da raiz do erro quadrático médio (*RMSE*), do coeficiente de determinação R^2.

Tabela 4.10 – Divisão por Manejo, os 05 atributos mais relevantes para sequeiro.

Técnica	**Regressão**			
	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	1,1160	±0,1157	0,4917	±0,1061
Árvore de Decisão	1,0840	±0,1388	0,5250	±0,1182
Random Forest	1,0927	±0,1126	0,5173	±0,0930
KNN	1,1229	±0,1278	0,5018	±0,0946

Fonte: Do Autor (2023).

Tabela 4.11 – Divisão por Manejo, os 10 atributos mais relevantes para sequeiro.

Técnica	**Regressão**			
	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	1,0887	±0,1423	0,5129	±0,1101
Árvore de Decisão	1,0938	±0,1399	0,5141	±0,1216
Random Forest	1,1361	±0,1287	0,4783	±0,0999
KNN	1,1577	±0,1255	0,4725	±0,0913

Fonte: Do Autor (2023).

Tabela 4.12 – Divisão por Manejo, 20 atributos mais relevantes para sequeiro.

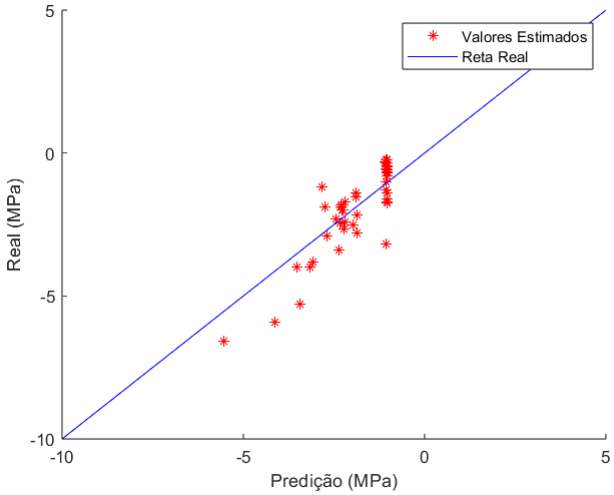
Técnica	**Regressão**			
	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	1,0569	±0,1246	0,5441	±0,1099
Árvore de Decisão	1,0959	±0,1334	0.5169	±0,1153
Random Forest	1,0982	±0,1196	0,5117	±0,1045
KNN	1,1476	±0,1197	0,4728	±0,0952

Fonte: Do Autor (2023).

Com base na Tabela 4.10, Tabela 4.11 e Tabela 4.12, a técnica Rede Neural Artificial apresenta o melhor resultado considerando os 20 atributos no estado sequeiro com o valor 1,0569 ±0,12462 de média da raiz do erro quadrático (*RMSE*) e 0.5441 ±0.1099 para coeficiente de determinação (R^2). De modo a exemplificar o resultado, a Figura 4.5 exibe o comparativo entre os dados reais e estimados para o melhor *fold* dentre os 10 da Rede Neural Artificial com melhor desempenho. Os parâmetro obtidos após as iterações

é uma rede neural de única camada intermediária, com 2 neurônios na camada e função de ativação Sigmoide. Onde a raiz do erro quadrático médio (*RMSE*) é de 0,7841 e o coeficiente de determinação (R^2) de 0,7690.

Figura 4.5 – Dados Reais x Dados Estimados (*fold* 9, melhor desempenho - Irrigado - 20 atributos mais relevantes).



Fonte: Do Autor (2023).

De maneira análoga a condição irrigado, o teste ANOVA foi aplicado para as métricas de desempenho dos 10 *folds* de todas as técnicas de *Machine Learning* do manejo sequeiro. O α = 0,05 de nível de significância foi adotado, para o *RMSE* o valor-p = 0,9763 e (R^2) o valor-p = 0,9881. Nestes resultados, a hipótese nula afirma que os valores médios do *RMSE* são iguais e os do R^2 também são iguais. Como o valor de valor-p é maior que o nível de significância de 0,05, isso indica que não há evidências estatísticas suficientes para rejeitar a hipótese nula. Nesse caso, não há diferenças estatisticamente significativas entre as médias dos grupos testados.

4.3.2.2 Classificação

Para o procedimento de classificação as Tabelas 4.13, 4.14 e 4.15 exibem a média (μ) e o desvio padrão (σ) da acurácia balanceada dos métodos implementados considerando as 3 variações de atributos para condição de sequeiro.

Atendendo o método de classificação, a técnica de *Random Forest* ostenta os melhores resultados de acordo com as Tabelas 4.13, 4.14 e 4.15. Dispondo de uma média de acerto de 47,08% $\pm$7,50%, em que os 21 atributos mais relevantes despontam com os melhores resultados. A técnica possui como parâmetro o método *Bag*, número de iterações igual a 94, tamanho mínimo das folhas igual a 2, número máximo de

Tabela 4.13 – Divisão por Manejo, os 05 atributos mais relevantes para sequeiro.

Técnica	**Classificação**	
	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	44,48	±7,41
Árvore de Decisão	45,84	±5,65
Random Forest	44,33	±5,61
KNN	43,31	±4,46

Fonte: Do Autor (2023).

Tabela 4.14 – Divisão por Manejo, os 10 atributos mais relevantes para sequeiro.

Técnica	**Classificação**	
	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	44,84	±5,58
Árvore de Decisão	43,06	±5,97
Random Forest	46,17	±6,15
KNN	45,01	±7,56

Fonte: Do Autor (2023).

Tabela 4.15 – Divisão por Manejo, 21 atributos mais relevantes para sequeiro.

Técnica	**Classificação**	
	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	44,25	±7,02
Árvore de Decisão	44,79	±6,85
Random Forest	47,08	±7,50
KNN	45,80	±7,28

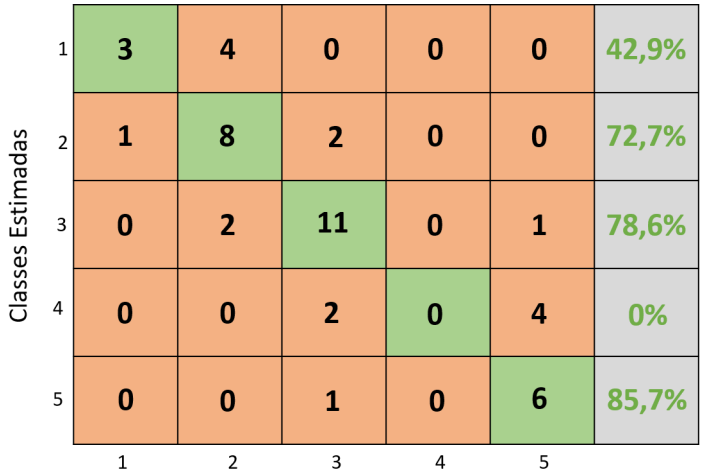
Fonte: Do Autor (2023).

divisões de 9 e critério de divisão *twoing*. A Figura 4.6 exibe a matriz de confusão do *fold* 4 com melhor desempenho para este caso, com uma acurácia balanceada de 55,97%. Equivalente aos dados em condição de irrigado, a baixa assertividade na classificação considerando as amostras da classe 4 possui como provável causa o desbalanceamento das amostras. O desbalanceamento é exibido na Tabela 4.16 para os dados de treino levando em conta o manejo sequeiro.

Com um desbalanceamento um pouco mais suave, porém ainda existente, as amostras de treino em condição de sequeiro exibidas na Tabela 4.16 provavelmente interferem no desempenho das técnicas de *Machine Learning*.

O método de classificação utilizou os mesmos parâmetros de regressão no teste ANOVA. O valor-p para a acurácia balanceada é 0,9797. Esse resultado sugere a falta de evidências estatísticas para afirmar diferenças entre as médias dos grupos testados.

Figura 4.6 – Matriz de confusão 21 mais relevantes (*fold* 4 - melhor desempenho - Sequeiro).



Fonte: Do Autor (2023).

Tabela 4.16 – Número de amostras por classe, dados de treino manejo sequeiro.

Classe	Número de amostras
1	61
2	101
3	126
4	55
5	58

Fonte: Do Autor (2023).

4.4 Oversampling - SMOTE

A fim de melhorar a performance das técnicas de *Machine Learning*, o algoritmo SMOTE é executado, algoritmo esse que cria dados sintéticos com intuito de balancear as classes minoritárias. Antes de implementar o algoritmo, a técnica de validação *holdout* é implementada e os dados devidamente pré-processados são separados em 80% para treino e 20% para teste. Apenas os dados de treino foram utilizados no SMOTE, de modo que os modelos uma vez treinados, o teste ocorreu apenas em dados reais.

4.4.1 Cafeeiro Irrigado

4.4.1.1 Regressão

Quando implementado o SMOTE para os dados de treino do manejo irrigado obteve-se um balanceamento de amostras. A Tabela 4.17 exibe o número de amostras para treino antes e depois da implementação

do SMOTE e também o número de dados teste. A Tabela 4.18 mostra o resultado das métricas avaliativas raiz do erro quadrático médio (*RMSE*) e o coeficiente de determinação (R^2) para as três determinações de atributos.

Tabela 4.17 – Divisão dos Dados SMOTE - Irrigado.

Regressão	
Dados Brutos Treino	353
Dados Sintéticos Treino	806
Total de Dados Treino	1165
Dados Teste	97

Fonte: Do Autor (2023).

Tabela 4.18 – SMOTE, métodos de desempenho regressão, irrigado.

Técnica	**Regressão**					
	05 mais relevantes		**10 mais relevantes**		**15 mais relevantes**	
	RMSE	R^2	*RMSE*	R^2	*RMSE*	R^2
Rede Neural	0,7487	0,5617	0,6999	0,6103	0,6884	0,6233
Árvore de Decisão	0,7644	0,5315	0,7524	0,5428	0,7877	0,5114
Random Forest	0,7919	0,4961	0,7422	0,5519	0,7360	0,5595
KNN	0,8290	0,4515	0,8490	0,4328	0,8184	0,4805

Fonte: Do Autor (2023).

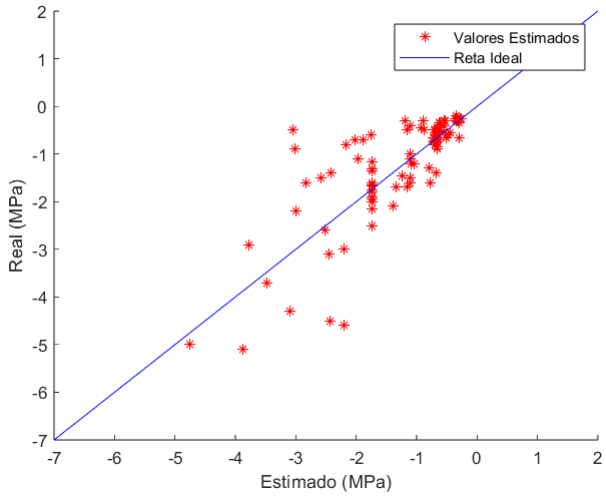
Pela Tabela 4.18 após a implementação de treino e teste para o 05, 10 e 15 atributos do manejo irrigado e considerando o método de regressão a técnica com melhor destaque é a Rede Neural Artificial levando em conta os 15 atributos mais relevantes com a raiz do erro quadrático médio (*RMSE*) de 0,6233 e o coeficiente de determinação (R^2) de 0,6884. Os parâmetro obtidos após as iterações é uma rede neural de 2 camadas intermediárias, com 2 neurônios na primeira camada e 69 na segunda, função de ativação Unidade Linear Retificada - do inglês *Rectified Linear Unit* (*ReLU*). A Figura 4.7 ilustra a distribuição dos dados em relação a reta ideal.

4.4.1.2 Classificação

Ainda no manejo irrigado e levando em consideração o método de classificação. A execução do SMOTE foi capaz de equilibrar as classes minoritárias do banco de dados. A Tabela 4.19 mostra o número de amostras totais e por classes, antes e após o SMOTE.

Com a execução dos modelos de *Machine Learning* para os 05, 10 e 22 atributos os resultados para acurácia balanceada são exibidos na Tabela 4.20. A técnica que obteve melhor desempenho foi *Random*

Figura 4.7 – Dados Reais x Dados Estimados (SMOTE - Irrigado - 15 atributos mais relevantes).



Fonte: Do Autor (2023).

Tabela 4.19 – Divisão dos Dados SMOTE - Irrigado.

Classificação						
	Total de Dados	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Dados Brutos Treino	353	103	165	69	10	6
Dados Sintéticos Treino	492	61	0	117	155	159
Total de Dados Treino	845	164	165	186	165	165
Dados Teste	97	32	36	18	5	6

Fonte: Do Autor (2023).

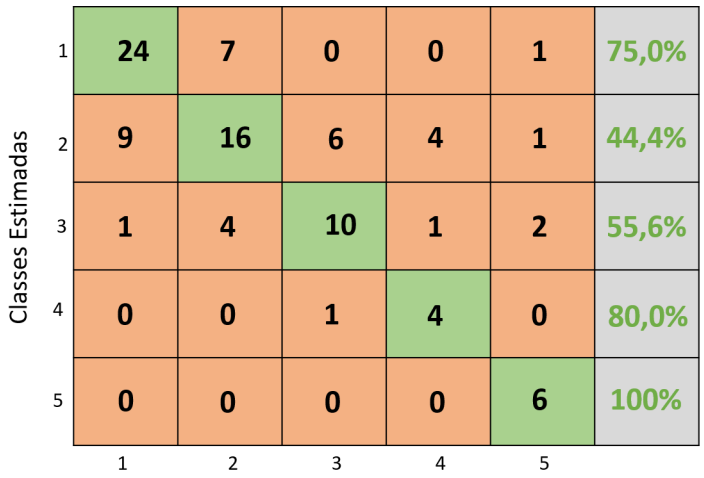
Tabela 4.20 – SMOTE Acurácia Balanceada irrigado.

Técnica	**Classificação**		
	05 mais relevantes	**10 mais relevantes**	**22 mais relevantes**
	Acurácia	**Acurácia**	**Acurácia**
Rede Neural	62,94	58,98	65,81
Árvore de Decisão	63,08	64,33	61,27
Random Forest	62,52	62,13	71,00
KNN	57,97	56,26	55,05

Fonte: Do Autor (2023).

Forest para os 22 mais relevantes com o método *Bag*, número de iterações igual a 32, tamanho mínimo das folhas igual a 2, número máximo de divisões de 17 e critério de divisão Entropia Cruzada. Com uma acurácia balanceada de 71,0% a matriz de confusão é exibida na Figura 4.8.

Figura 4.8 – Matriz de confusão 22 mais relevantes (SMOTE - Árvore de Decisão - Irrigado).



Fonte: Do Autor (2023).

4.4.2 Sequeiro

4.4.2.1 Regressão

Considerando o manejo sequeiro e a implementação do algoritmo SMOTE o balanceamento das amostras é mostrado na Tabela 4.21. A Tabela 4.22 exibe o resultado das métricas de avaliação raiz do erro quadrático médio (*RMSE*) e o coeficiente de determinação (R^2).

Tabela 4.21 – Divisão dos Dados SMOTE - Sequeiro.

Regressão	
Dados Brutos Treino	340
Dados Sintéticos Treino	617
Total de Dados Treino	957
Dados Teste	105

Fonte: Do Autor (2023).

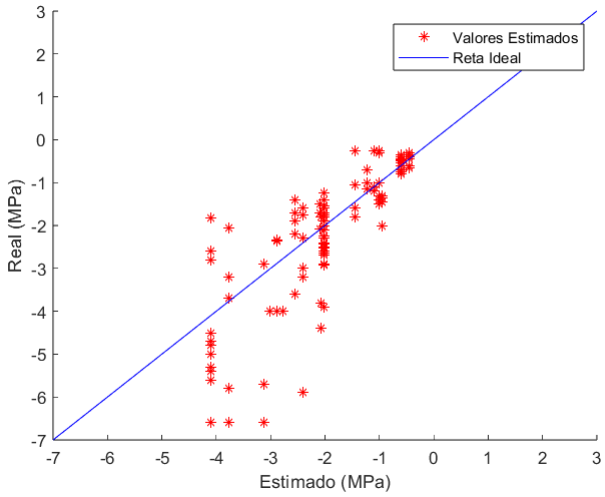
Pela análise da Tabela 4.22 a Árvore de Decisão desponta com melhor resultado para os 10 atributos mais relevantes. A Figura 4.9 mostra a distribuição dos dados em relação a reta ideal com raiz do erro quadrático médio (*RMSE*) igual a 1,0137 e o coeficiente de determinação (R^2) de 0,6520 para os parâmetros de número de ramificações de 307 e valor mínimo de observações por folha de 35 e critério de divisão Índice Gini.

Tabela 4.22 – SMOTE, métodos de desempenho regressão, sequeiro.

Técnica	**Regressão**					
	05 mais relevantes		**10 mais relevantes**		**15 mais relevantes**	
	RMSE	R^2	*RMSE*	R^2	*RMSE*	R^2
Rede Neural	1,1375	0,5451	1,0938	0,5645	1,0626	0,6021
Árvore de Decisão	1,0338	0,6330	1,0137	0,6520	1,0784	0,6108
Random Forest	1,0571	0,6305	1,0748	0,6127	1,0890	0,6002
KNN	1.0319	0,6095	1,1150	0,5446	1,1414	0,5256

Fonte: Do Autor (2023).

Figura 4.9 – Dados Reais x Dados Estimados (SMOTE - Irrigado - 15 atributos mais relevantes).



Fonte: Do Autor (2023).

4.4.2.2 Classificação

Por fim, de maneira similar ao manejo irrigado o algoritmo SMOTE foi capaz de balancear as classes do sequeiro. A Tabela 4.23 estampa o total de dados brutos e sintéticos, como também as classes e seus respectivos números de amostras.

Tabela 4.23 – Divisão dos Dados SMOTE - Sequeiro.

Classificação						
	Total de Dados	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Dados Brutos Treino	340	53	88	109	45	45
Dados Sintéticos Treino	210	53	21	0	68	68
Total de Dados Treino	550	106	109	109	113	113
Dados Teste	105	17	21	31	15	21

Fonte: Do Autor (2023).

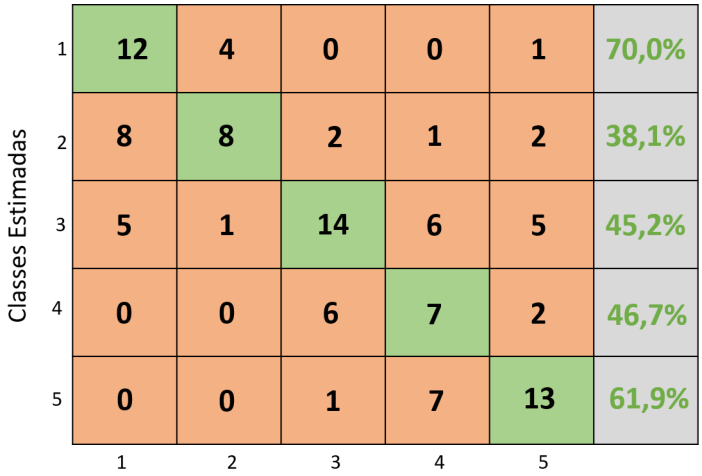
Tabela 4.24 – SMOTE acurácia balanceada sequeiro.

Técnica	**Classificação**		
	05 mais relevantes	**10 mais relevantes**	**22 mais relevantes**
	Acurácia	**Acurácia**	**Acurácia**
Rede Neural	49,25	48,56	48,00
Árvore de Decisão	52,48	50,43	48,52
Random Forest	49,55	50,57	49,24
KNN	46,51	46,73	47,23

Fonte: Do Autor (2023).

A Tabela 4.24 exibe os valoes da acurácia balanceada para as quatro técnicas. Abordando os 05 atributos mais relevantes com uma acurácia balanceada de 52,48%, a árvore de decisão obteve a melhor performance. A Figura 4.10 ilustra a matriz de confusão para a técnica em si, seu parâmetros são, o número de ramificações de 356, valor mínimo de observações por folha de 6 e critério de divisão entropia cruzada, sua matriz de confusão é mostrada na Figura 4.10.

Figura 4.10 – Matriz de confusão 05 mais relevantes (SMOTE - Árvore de Decisão - Sequeiro).



Fonte: Do Autor (2023).

Quando comparado as duas propostas (sem e com SMOTE) a implementação do algoritmo SMOTE não apresentou melhora significativa no desempenho das técnicas de *Machine Learning*, exceto para classificação do manejo irrigado, na qual a acurácia balanceada sofreu uma melhora de aproximadamente 9%. O Apêndice C exibe as tabelas comparativas das melhores técnicas. Quando considerado uma análise comparativa por classes via matriz de confusão, o acréscimo de dados sintéticos no treino e uma validação cruzada via *holdout*, que proporcionou uma maior quantidade de amostras para teste, ocasionou em um melhor desem-

penho nas classes minoritárias (4 e 5) da condição irrigado e também para a classe 4 em condição sequeiro, a qual na proposta sem o algoritmo SMOTE não apresentou acerto de classificação de amostras segundo a Figura 4.6.

Com o intuito de melhorar o rendimento, técnicas alternativas podem ser consideradas, como *Safe-Leve*-SMOTE que foca a geração de amostras em regiões confiáveis da classe minoritária e evita regiões com muitos ruídos ou com sobreposição de dados. Até mesmo considerar amostras da classe minoritária estão cercadas de um número muito maior de vizinhos da classe majoritária do que da classe minoritária, estas amostras podem ser consideradas como *outliers* e ignoradas.

5 CONCLUSÃO

A presente pesquisa buscou investigar o comprimento de onda ou a faixa de comprimentos de onda mais adequada para inferir o potencial hídrico do cafeeiro, explorando técnicas de análise de dados e inteligência artificial. Ao longo deste estudo, foram alcançados resultados que proporcionam uma visão mais profunda sobre reflectância e potencial hídrico.

Os resultados indicam que, em condições de sequeiro, as reflectâncias em torno do comprimento de onda de 690 nm são mais apropriadas para inferir o potencial hídrico dos cafeeiros. No entanto, em situações de manejo irrigado, as reflectâncias em torno do comprimento de onda de 780 nm mostram-se mais relevantes para essa inferência.

Além disso, durante o desenvolvimento desta pesquisa, foram identificados desafios e limitações, tais como o desbalanceamento da base de dados, o baixo número de amostras referentes as classes 4 e 5 da base de dados e a sobreposição dos dados, o que acarretou em um desempenho não satisfatório das técnicas de *Machine Learning* implementadas. Na tentativa de sanar as adversidades do desbalanceamento de classes o algoritmo SMOTE foi implementado, porém, sem melhora considerável nos resultados. De modo que a sobreposição dos dados e baixo número de amostras para determinadas classes apresentam como possíveis causas do desempenho.

São alternativas para trabalhos futuros a implementação de técnicas como o *Safe-Leve*-SMOTE ou considerar classes minoritárias como *outliers*, ou até mesmo técnicas mais sofisticadas como Few-Shot Learning, que lida com a tarefa de aprendizado com poucas amostras e busca capacitar os modelos de aprendizado de máquina a aprender com exemplos limitados, tornando-os mais flexíveis e adaptáveis a novos conjuntos de dados. Além do desenvolvimento de um sensor portátil embasado na presente pesquisa e com um custo final consideravelmente inferior aos encontrado no mercado atualmente.

Enfim, é importante considerar aspectos que não foram abordados no presente estudo como o custo computacional, facilidade de implementação e entendibilidade das técnicas, além dos resultados obtidos.

REFERÊNCIAS

ALZUBAIDI, L. et al. **Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Jour-nal of Big Data**, n. 8, p. 74, 2021.

ANGIULLI, F.; PIZZUTI, C. **Outlier Mining in Large High-Dimensional Data Sets. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**, v. 17, n. 2, p. 203–215, 2005.

ATKINS, P.; PAULA, J. de. **Physical Chemistry for the Life Sciences**. New York: Oxford University Press, 2006.

AZEVEDO, F. M. d.; BRASIL, L. M.; OLIVEIRA, R. C. L. d. **Redes neurais com aplicações em controles e em sistemas especialistas**. Florianópolis: Bookstore, 2000.

BISHOP, C. M. **Neural networks for Pattern Recognition.** 4. ed. Birmingham: Clarendon Press, 1995.

BORGES, F. A. S. **Extração de características combinadas com árvores de decisão para detecção e classificação dos distúrbios de qualidade da energia elétrica. Universidade de São Paulo**, p. 118, 2013.

BRAGA, A. d. P.; CARVALHO, A.; LUDERMIR, T. **Redes Neurais Artificias: Teorias e Aplicações**. 2. ed. Rio de Janeiro: LTC, 2007.

BREIMAN, L. **Random forests. Machine Learning. Springer Science and Business Media LLC**, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. **WALD LECTURE II. LOOKING INSIDE THE BLACK BOX.** 1. ed. [S.l.]: UCB Statistics., 2003.

BRODERSEN, K. H. et al. **The Balanced Accuracy and Its Posterior Distribution**. **International Conference on Pattern Recognition**, v. 9, p. 3121–3124, ago. 2010.

CARVALHO, I. R. et al. **DEMANDA HÍDRICA DAS CULTURAS DE INTERESSE AGRONÔMICO**. **ENCICLOPÉDIA BIOSFERA**, v. 9, n. 17, p. 969–985, dez. 2013. Disponível em: <https://www.conhecer.org.br/enciclop/2013b/CIENCIAS%20AGRARIAS/DEMANDA%20HIDRICA.pdf>. Acesso em: 26 mar. 2023.

CECCATO, P. et al. **Detecting Vegetation Leaf Water Content Using Reflectance in the Optical Domain**. **Remote Sensing of Environment**, v. 77, p. 22–33, jul. 2001. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0034425701001912>. Acesso em: 13 mac. 2023.

CHAIN, T.; DRAXLER, R. R. **Root mean square error (RMSE) or mean absolute error (MAE)? Geoscientific Model Development**, v. 7, n. 1, p. 1525 – 1534, fev. 2014.

CHANG, R. **Physical Chemistry for the Biosciences**. Sausalito: University Science Books, 2005.

CHEMURA, A.; MUTANGA, O.; DUBE, T. **Remote sensing leaf water stress in coffee (Coffea arabica) using secondary effects of water absorption and random forests**. **Physics and Chemistry of the Earth, Parts A/B/C**, v. 100, n. 1, p. 317–324, ago. 2017.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. **The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation**. **PeerJ Computer Science**, n. 7, p. e623, jul. 2021.

CONSONNI, V. et al. **A MATLAB toolbox for multivariate regression coupled with variable selection. Chemometrics and Intelligent Laboratory Systems**, v. 213, p. 9, 2021. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0169743921000812?via%3Dihub>. Acesso em: 19 fev. 2023.

COVER, T. M.; HART, P. E. **Nearest Neighbor Pattern Classification. Department of Computer Science and Technology**, v. 13, n. 1, p. 21–27, 1967.

CYBENKO, G. **"Continuous Valued Networks With Two Hidden Layers Are Sufficien. Department of Computer Science**, abr. 1988.

DAMATTA, F. M.; RAMALHO, J. D. C. **Impacts of drought and temperature stress on coffee physiology and production: a review**. **Brazilian Journal of Plant Physiology**, v. 18, n. 1, p. 55–81, mar. 2006. Disponível em: <https://www.scielo.br/j/bjpp/a/bDfpJwLr4xLcznSwy4b9zkf/?lang=en>. Acesso em: 11 fev. 2023.

DARBELLAY, G. A.; VAJDA, I. **Estimation of the Information by an Adaptive Partitioning of the Observation Space**. **IEEE TRANSACTIONS ON INFORMATION THEORY**, v. 45, n. 4, p. 1315–1321, 1999.

DATT, B. **Remote Sensing of Water Content in Eucalyptus Leaves**. **Australian Journal of Botany**, v. 47, n. 909, p. 1–16, out. 1999. Disponível em: <https://www.semanticscholar.org/paper/Remote-Sensing-of-Water-Content-in-Eucalyptus-Datt/d3984d1d546efaba0e737b1d3da080b8a2a2db98>. Acesso em: 05 abr. 2023.

DELALIEUX, S. et al. **Detection of Biotic Stress (Venturia inaequalis) in Apple Trees Using Hyperspectral data: Non-parametric statistical approaches and physiological implications**. **Europ. J. Agronomy**, v. 27, p. 130–143, fev. 2007. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1161030107000287?via3Dihub>. Acesso em: 03 mar. 2023.

DING, C.; PENG, H. **Minimum Redundancy Feature Selection from Microarray Gene Expression Data**. **IEEE TRANSACTIONS ON INFORMATION THEORY**, v. 3, n. 2, p. 185–205, 2005.

DUBATH, P. et al. **Random forest automated supervised classification of Hipparcos periodic variable stars. Monthly Notices of the Royal Astronomical Society, Blackwell Publishing Ltd**, v. 414, n. 3, p. 2602–2617, 2011.

FACELI, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina.** 1. ed. Rio de Janeiro: LTC., 2011.

FLECK, L. et al. **REDES NEURAIS ARTIFICIAIS: PRINCÍPIOS BÁSICOS. Revista Eletrônica Científica Inovação e Tecnologia**, v. 1, n. 13, p. 47–57, jan. 2016. Disponível em: <https://periodicos.utfpr.edu.br/recit/article/viewFile/4330/Leandro>. Acesso em: 03 mar. 2023.

FLORENZANO, T. **Imagens de Satélite para Estudos Ambientais**. fev. 2002. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/5692586/mod_resource/content/2/Imagens20de20sateCC81lite20para20estudos20ambientais.pdf>. Acesso em: 25 fev. 2023.

FURTADO, H.; MACAU, E. E. N.; VELHO, H. F. d. C. **ASSIMILAÇÃO DE DADOS COM REDES NEURAIS ARTIFICIAIS EM EQUAÇÕES DIFERENCIAIS. International Conference on Artificial Intelligence (ICAI)**, v. 1, p. 488–494, ago. 2011.

GARCIA, S. H. et al. **Simulação de estresse hídrico em feijão pela diminuição do potencial osmótico**. **Revista de Ciências Agroveterinárias.**, v. 11, n. 1, p. 35–41, ago. 2011. Disponível em: <https://revistas.udesc.br/index.php/agroveterinaria/article/view/5234/3408>. Acesso em: 11 abr. 2023.

GENC, L. et al. **Determination of Water Stress With Spectral Reflectance on Sweet Corn (Zea mays L.) Using Classification Tree (CT) Analysis**. I, n. 1, p. 1–10, dez. 2013.

GIANNINI, T. C. et al. **Desafios atuais da modelagem preditiva de distribuição de espécies. Revista do Jardim Botânico do Rio de Janeiro**, v. 63, n. 3, p. 733–749, 2012.

GOMES, N. M.; LIMA, L. A.; CUSTóDIO, A. A. d. P. **Crescimento vegetativo e produtividade do cafeeiro irrigado no sul do Estado de Minas Gerais**. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 11, n. 6, p. 564–570, dez. 2007. Disponível em: <https://www.scielo.br/j/rbeaa/a/V7wPyMkTThXxLv3q4tj7nzx/?lang=pt>. Acesso em: 11 abr. 2023.

GUJARATI, D. N.; PORTER, D. C. **Basic Econometrics**. 2. ed. New York City: McGraw-Hill/Irwin, 2008.

GUTIERREZ, M.; REYNOLDS, M. P.; KLATT, A. R. **Association of Water Spectral Indices With Plant and Soil Water Relations in Contrasting Wheat Genotypes**. **Journal of Experimental Botany**, v. 61, n. 12, p. 3291–3303, maio 2010. Disponível em: <http://jxb.oxfordjournals.org/open_access.html>. Acesso em: 11 mar. 2023.

HALMENSCHLAGER, C. **Um algoritmo para indução de árvores e regras de decisão.** p. 112, 2002. Disponível em: <https://lume.ufrgs.br/bitstream/handle/10183/2755/000325797.pdf?sequence=1,20Acesso: 20julho/2015>. Acesso em: 06 mar. 2023.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. 2. ed. Tradução de Paulo Martins Angel, Porto Alegre: Bookman, 2001.

HAYKIN, S. **Neural networks and learning machines**. 3. ed. Ontario: Pearson Prentice Hall., 2008.

HO, T. K. **Random Decision Forests. Proceedings of 3rd International Conference on Document Analysis and Recognition**, v. 1, p. 278–282, 1995. Disponível em: <https://doi.org/10.1109/icdar.1995.598994>.

HOPFIELD, J. J. **Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences USA**, v. 79, p. 2524–2528, abr. 1982. Disponível em: <https://www.pnas.org/doi/pdf/10.1073/pnas.79.8.2554>. Acesso em: 21 mar. 2023.

HOPKINS, W. G.; HüNER, N. P. A. **Introduction to Plant Physiology**. 4. ed. Ontario: University of Western Ontario, 2008.

IMANDOUST, S. B.; BOLANDRAFTAR, M. **Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. International Journal of Engineering Research and Applications**, v. 3, n. 5, p. 605–610, 2013.

JACQUEMOUDD, S.; FREDERIC, B. **PROSPECT: A Model of Leaf Optical Properties Spectra** . **Remote Sensing of Environment**, v. 34, p. 75–91, jun. 1990. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/003442579090100Z>. Acesso em: 19 mar. 2023.

KACIRA, M. et al. **Plant Response-Based Sensing for Control Strategies in Sustainable Greenhouse Production**. **Journal of Agricultural Meteorology (Japan)**, v. 61, p. 15–22, jan. 2005. Disponível em: <https://www.researchgate.net/publication/250167547_Plant_Response-Based_Sensing_for_Control_Strategies_in_Sustainable_Greenhouse_Production>. Acesso em: 19 mar. 2023.

KATSOULAS, N. et al. **Crop Reflectance Monitoring as a Tool for Water Stress Detection in Greenhouses: A review**. **ELSERVIER**, I, n. 1, p. 1–25, out. 2016. Disponível em: <www.elsevier.com/locate/issn/15375110>. Acesso em: 02 mar. 2023.

KELLEHER, J. D.; NAMEE, B. M.; D'ARCY, A. **Fundamentals of Machine Learning for Predictive Data Analytics**. 1. ed. Massachusetts: MIT Press., 2015.

KIRASICH, K.; SMITH, T.; SADLER, B. **Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review**, v. 1, n. 3, p. 25, 2018.

KOTSIANTIS, S. B. **Supervised Machine Learning: A Review of Classification Techniques. Department of Computer Science and Technology**, p. 249–268, 2007.

KOUIROUKIDIS, N.; EVANGELIDIS, G. **The Effects of Dimensionality Curse in High Dimensional kNN Search. Panhellenic Conference on Informatics**, n. 15, p. 41–45, 2011.

KROSE, B.; SMAGT, P. Van der. **An Introduction to Neural Networks**. 8. ed. Amsterdam: University of Amsterdam., 1996.

KUSNER, M. J. et al. **Stochastic Neighbor Compression. Proceedings of the 31st international conference on machine learning (ICML-14)**, n. 15, p. 622–630, 2014.

LACERDA, C. F.; FILHO, J. E.; PINHEIRO, C. B. **FISIOLOGIA VEGETAL. UNIDADE III**. 1. ed. Fortaleza: http://www.fisiologiavegetal.ufc.br/apostila.htm, 2007. Acesso em: 13 fev. 2023.

LIMA, M. C. d.; FARIA, E. R.; BARIONI, M. C. N. **HubISC: um novo algoritmo baseado em hubness para classificação de fluxo de dados de imagens. Proceedings of the 37th Brazilian Symposium on Data Bases**, n. 37, p. 138–150, 2022.

MAATEN, L. s. Van der; HINTON, G. **VisualizingDatausingt-SNE**. **J. Machine Learning Research.**, v. 9, p. 2579–2605, nov. 2009. Disponível em: <chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>. Acesso em: 02 mar. 2023.

MACIEL, D. A. et al. **Leaf water potential of coffee estimated by landsat-8 images**. **PLOS ONE**, p. e0230013, mar. 2020.

MANI P.AND VAZQUEZ, M. M.-B. J. et al. **The hubness phenomenon in high-dimensional spaces. Association for Women in Mathematics Series**, p. 15–45, 2019.

MARIZ, F. M. **AVALIAÇÃO E COMPARAÇÃO DE VERSÕES MODIFICADAS DO ALGORITMO KNN. UNIVERSIDADE FEDERAL DE PERNAMBUCO CENTRO DE INFORMÁTICA**, p. 54, 2017.

MCCULLOCH, W. S.; PITTS, W. **A logical calculus nervous activity. Bulletin of Mathematical Biophysics**, v. 5, p. 115–133, set. 1943. Disponível em: <https://www.nature.com/articles/187922a0>. Acesso em: 13 mar. 2023.

MEDEIROS, N. S. R. et al. **ANÁLISE DO DESEMPENHO DO ALGORTIMO K-NEAREST NEIGHBORS NA CLASSIFICAÇÃO DE PATOLOGIAS DE COLUNA VERTEBRAL. Congresso Nacional de Pesquisa e Ensino em Ciências**, n. 5, p. 12, 2020. Disponível em: <https://editorarealize.com.br/edicao/detalhes/anais-do-v-conapesc>. Acesso em: 22 fev. 2023.

MICKEY, R. M.; DUNN, O. J.; CLARK, V. A. **Applied Statistics: Analysis of Variance and Regression**. 4. ed. Hoboken: Wiley, 2004.

MINSKY, M.; PAPERT, S. **Perceptrons: an introduction to computational geometry.** 1. ed. Massachusetts: MIT Press, 1969.

MIRANDA, F. A.; FREITAS, S. R. C. D.; FAGGION, P. L. **INTEGRAÇÃO E INTERPOLAÇÃO DE DADOS DE ANOMALIAS AR LIVRE UTILIZANDO-SE A TÉCNICA DE RNA E KRIGAGEM. Boletim de Ciências Geodésicas**, v. 15, n. 3, p. 428–443, jul. 2009. Disponível em: <https://www.redalyc.org/pdf/3939/393937709008.pdf>. Acesso em: 12 mar. 2023.

MITCHELL, T. M. **Machine Learning.** 1. ed. [S.l.]: McGraw-Hill Science/Engineering/Math., 1997.

MONARD, M. C.; BARANAUSKAS, J. A. **Conceitos Sobre Aprendizado de Máquina. Sistemas Inteligentes Fundamentos e Aplicações.** 1. ed. Barueri: Manole Ltda., 2003.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 9. ed. Hoboken: Wiley, 2017.

MORAES. **Manual de Referencia. In: VII Simpósio Brasileiro de Sensoriamento Remoto**. Curitiba, 1993. Disponível em: <http://mtc-m21b.sid.inpe.br/col/sid.inpe.br/mtc-m21b/2017/10.18.16.17/doc/INPE-7605.pdf>. Acesso em: 12 mar. 2023.

MURTHY, S. K. **Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Kluwer Academic Publishers**, v. 2, p. 345–389, 1998.

NUNES, P. et al. **Estudo do Potencial Hıdrico em Cafeeiros Utilizando Tecnicas de Aprendizado de Maquina**. [S.1]:, p. 1–8, 2020.

NUNES, P. H. et al. **Predicting coffee water potential from spectral reflectance indices with neural networks**. **Smart Agricultural Technology**, n. 7, p. 100213, ago. 2023.

OSHIRO, T. M. **Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica. Revista do Jardim Botânico do Rio de Janeiro**, p. 101, 2013.

PONZONI, F. J. **Sensoriamento Remoto no estudo da vegetação: diagnosticando a Mata Atlântica. Cap 8.** 1. ed. São Jose dos Campos: Instituto Nacional de pesquisas Espaciais - Divisão de Sensoriamento Remoto, 2002. Disponível em: <http://mtc-m12.sid.inpe.br/col/sid.inpe.br/sergio/2005/06.14.13.11/doc/CAP8_FJPonzoni.pdf>. Acesso em: 07 mar. 2023.

PRATT, W. K. **DIGITAL IMAGE PROCESSING.** 3. ed. [S.l.]: JOHN WILEY SONS., 2001.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.-L. **Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 3, p. 1–15, 2019.

QUINLAN, J. R. **Learning efficient classification procedures ans their application to chessend games. Machine Learning: An AI Approach**, v. 1, p. 463–482, 1983.

RAMOS, J. A. d. P. **Árvores de Decisão Aplicadas À Detecção de Fraudes Bancárias. Universidade de Brasília**, p. 65, 2014.

REN, T. et al. **Study of Dynamometer Cards Identification Based on Root-Mean-Square Error Algorithm**. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 32, n. 2, p. 1525 – 1534, nov. 2017.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. Barueri: Manole, 2003.

ROMASZEWSKI, M.; GŁOMB, P.; CHOLEWA, M. **Adaptive, Hubness-Aware Nearest Neighbour Classifier with Application to Hyperspectral Data. Springer International Publishing**, n. 8, p. 113–120, 2018.

SALISBURY, F.; ROSS, C. **Plant Physiology. 682p**. 4. ed. California: Wadsworth Publishing Company, 1992.

SAMMUT, C.; WEBB, G. I. **Encyclopedia of Machine Learning.** 1. ed. Sydney, Australia: Springer References., 2011.

SANTOS, A. R. **Noções Teóricas e Práticas de Sensoriamento Remoto. Cap 4.** Espirito Santo: https://www.mundogeomatica.com.br, 2013.

SANTOS, S. A. dos et al. **Evaluation of the Water Conditions in Coffee Plantations Using RPA**. **AgriEngineering**, v. 5, n. 1, p. 65–84, dez. 2022.

SEBER, G. A. F.; LEE, A. J. **Linear Regression Analysis**. 2. ed. Hoboken: Wiley, 2003.

SHANNON, C. E. **A Mathematical Theory of Communication. The Bell System Techinical Journal**, v. 3, n. XXVII, p. 379–423, 1948.

SHEKHAR, S.; XIONG, H. **Encyclopedia of GIS**. 2. ed. New York City: Springer, 2017.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes Neurais Artificiais para engenharia e ciências aplicadas.** 2. ed. São Paulo: Artliber, 2016.

SILVA, V. A. et al. **Impacto do déficit hídrico e temperaturas elevadas sobre o estado hídrico do cafeeiro nas regiões Sul e Cerrado de Minas Gerais**. **Circular Técnica - Empresa de Pesquisa Agropecuária de Minas Gerais Departamento de Informação Tecnológic**, n. 356, p. 5, 2021.

SLATYER, R.; TAYLOR, S. **Terminology in Plant-Soil-Water Relations**. **Nature**, v. 187, n. 4741, p. 922–924, set. 1960. Disponível em: <https://www.nature.com/articles/187922a0>. Acesso em: 05 abr. 2023.

SOUDANI, K. et al. **Ground-based Network of NDVI Measurements for Tracking Temporal Dynamics of Canopy Structure and Vegetation Phenology in Different Biomes**. **Remote Sensing of Environment**, n. 123, p. 234–245, abr. 2012. Disponível em: <https://www.sciencedirect.com/journal/remote-sensing-of-environment>. Acesso em: 28 mar. 2023.

STEFFEN, C. A. **INTRODUÇÃO AO SENSORIAMENTO REMOTO**. São Jose dos Campos: Instituto Nacional de pesquisas Espaciais - Divisão de Sensoriamento Remoto, 2016. Disponível em: <http://www3.inpe.br/unidades/cep/atividadescep/educasere/apostila.htm>. Acesso em: 26 mar. 2023.

TAIZ, L.; ZEIGER, E. **Plant Physiology. 481p**. 5. ed. Sunderland: Sinauer Associates Inc., 2010.

TANG, B.; HE, H. **Extended Nearest Neighbor Method for Pattern Recognition. IEEE Computational intelligence magazine**, v. 10, n. 3, p. 52–60, 2015.

TOSIN, R. et al. **Estimation of grapevine predawn leaf water potential based on hyperspectral reflectance data in Douro wine region**. **VITIS - Journal of Grapevine Research**, v. 59, n. 1, p. 9–18, fev. 2020.

TUKEY, J. W. **Exploratory Data Analysis.** 1. ed. [S.l.]: Addison-Wesley, Reading., 1977.

VIEIRA, S. a. **Introdução à Bioestatística**. 4. ed. Rio de Janeiro, Brasil: Elsevier., 2008.

VO, K.; HERNANDEZ, M.; PATEL, N. **Electromagnetic Radiation**. **LibreTexts**, abr. 2022. Disponível em: <https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Spectroscopy/Fundamentals_of_Spectroscopy/Electromagnetic_Radiation>. Acesso em: 21 mar. 2023.

WANG H.AND ZHOU, Z. W. Y.; YAN, X. **Feature selection for image classifica-tion based on bacterial colony optimization. Advances in Swarm Intelligence: 12th International Conference**, v. 10, n. 3, p. 430–439, 2021.

WANG, Z. et al. **Robust high dimensionalstream classification with novel class detection. IEEE 35th International Conference on Data Engineering (ICDE)**, p. 1418–1429, 2019.

WEISS, S. M.; KULIKOWSKI, C. A. **Computer syztem that leran: Classification and prediction methods from statistics, neural net, machine learninge, and expert systems.** 1. ed. Ontario: Pearson Education., 1990.

WITTEN, I. H.; FRANK, E.; HALL, M. **Data Mining: Practical Machine Learning Tools and Techniques.** 3. ed. Burlington: Morgan Kaufmann Publishers., 2011.

WRIGHT, M. N.; KöNIG, I. R. **Splitting on categorical predictors in random forests. PeerJ**, v. 7, n. e6339, p. 1–15, 2019.

WU, Q. et al. **HIBoost: A hubness-aware ensemble learning algorithm for high-dimensional imbalanced data classification. Journal of Intelligent Fuzzy Systems**, p. 12, 2020.

ZAKALUK, R.; RANJAN, R. S. **Predicting leaf water potential of potato using spectral reflectance indices**. **CANADIAN BIOSYSTEMS ENGINEERING**, v. 50, n. 1, p. 7.1–7.12, jan. 2008.

ZHAO, Y.; ZHANG, Y. **Comparison of decision tree methods for finding active objects. Advances in Space Research**, v. 41, n. 12, p. 1955–1959, 2007.

APÊNDICE A – Atributos mais Relevantes

Os atributos mais relevantes após a aplicação da técnica de *Pearson* são exibidos nas Tabelas 1 e 2. Uma vez que, em condição irrigado os atributos passaram de 2863 dimensões para 15 e 22, para regressão e classificação, respectivamente. Já para sequeiro os atributos passaram de 2863 dimensões para 20 e 21 atendendo regressão e classificação, respectivamente.

Tabela 1 – Atributos mais relevantes para irrigado.

***Ranking* dos atributos**	***Designação do atributo***	
	Regressão	**Classificação**
1	mês de coleta	mês de coleta
2	reflectância para λ = 779,751 nm	reflectância para λ = 780,123 nm
3	ano de coleta	ano de coleta
4	reflectância para λ = 784,944 nm	reflectância para λ = 784,944 nm
5	reflectância para λ = 783,091 nm	reflectância para λ = 783,462 nm
6	reflectância para λ = 781,051 nm	reflectância para λ = 779,380 nm
7	reflectância para λ = 779,380 nm	reflectância para λ = 782,906 nm
8	reflectância para λ = 784,574 nm	reflectância para λ = 781,051 nm
9	reflectância para λ = 782,906 nm	reflectância para λ = 779,937 nm
10	reflectância para λ = 779,566 nm	reflectância para λ = 784,759 nm
11	reflectância para λ = 781,236 nm	reflectância para λ = 779,751 nm
12	reflectância para λ = 784,759 nm	reflectância para λ = 785,315 nm
13	reflectância para λ = 779,937 nm	reflectância para λ = 783,091 nm
14	reflectância para λ = 785,130 nm	reflectância para λ = 779,566 nm
15	reflectância para λ = 785,315 nm	reflectância para λ = 785,130 nm
16	-	reflectância para λ = 781,236 nm
17	-	reflectância para λ = 782,720 nm
18	-	reflectância para λ = 779,194 nm
19	-	reflectância para λ = 783,277 nm
20	-	reflectância para λ = 780,308 nm
21	-	reflectância para λ = 784,574 nm
22	-	reflectância para λ = 780,865 nm

em que (λ) corresponde a comprimento de onda.

Fonte: Próprio Autor.

Tabela 2 – Atributos mais relevantes para sequeiro.

***Ranking* dos atributos**	***Designação do atributo***	
	Regressão	**Classificação**
1	mês de coleta	mês de coleta
2	reflectância para λ = 690,885 nm	reflectância para λ = 690,498 nm
3	reflectância para λ = 694,167 nm	reflectância para λ = 692,430 nm
4	reflectância para λ = 692,816 nm	reflectância para λ = 688,758 nm
5	reflectância para λ = 691,271 nm	reflectância para λ = 691,657 nm
6	reflectância para λ = 690,691 nm	reflectância para λ = 689,725 nm
7	reflectância para λ = 693,202 nm	reflectância para λ = 691,464 nm
8	reflectância para λ = 691,464 nm	reflectância para λ = 689,145 nm
9	reflectância para λ = 692,044 nm	reflectância para λ = 691,851 nm
10	reflectância para λ = 691,078 nm	reflectância para λ = 691,078 nm
11	reflectância para λ = 693,974 nm	reflectância para λ = 692,237 nm
12	reflectância para λ = 692,430 nm	reflectância para λ = 689,338 nm
13	reflectância para λ = 691,657 nm	reflectância para λ = 690,691 nm
14	reflectância para λ = 692,623 nm	reflectância para λ = 690,112 nm
15	reflectância para λ = 692,237 nm	reflectância para λ = 692,044 nm
16	reflectância para λ = 693,781 nm	reflectância para λ = 688,951 nm
17	reflectância para λ = 691,851 nm	reflectância para λ = 690,885 nm
18	reflectância para λ = 693,588 nm	reflectância para λ = 689,532 nm
19	reflectância para λ = 693,009 nm	reflectância para λ = 691,271 nm
20	reflectância para λ = 693,395 nm	reflectância para λ = 690,305 nm
21	-	reflectância para λ = 689,918 nm

em que (λ) corresponde a comprimento de onda.

Fonte: Próprio Autor.

APÊNDICE B – Tabelas comparativas

As Tabelas 3 e 4 exibem os melhores resultados das duas melhores técnicas para condição de irrigado. Já as Tabelas 5 e 5 para condição de sequeiro.

Tabela 3 – Comparativo regressão Irrigado.

Técnica	**Regressão**			
	Irrigado			
	Sem SMOTE		**Com SMOTE**	
	RMSE	R^2	*RMSE*	R^2
Rede Neural	-	-	0,6233	0.6884
Árvore de Decisão	0,4342	0,6993	-	-
Random Forest	-	-	-	-
KNN	-	-	-	-

Fonte: Do Autor (2023).

Tabela 4 – Comparativo classificação irrigado.

Técnica	**Classificação**	
	Irrigado	
	Sem SMOTE	**Com SMOTE**
	Acurácia	*Acurácia*
Rede Neural	62,05	
Árvore de Decisão	-	-
Random Forest	-	71,00
KNN	-	-

Fonte: Do Autor (2023).

Tabela 5 – Comparativo regressão sequeiro.

Técnica	**Regressão**			
	Sequeiro			
	Sem SMOTE		**Com SMOTE**	
	RMSE	R^2	*RMSE*	R^2
Rede Neural	0,7841	0,7690	-	-
Árvore de Decisão	-	-	1,0137	0,6520
Random Forest	-	-	-	-
KNN	-	-	-	-

Fonte: Do Autor (2023).

Tabela 6 – Comparativo classificação sequeiro.

Técnica	**Classificação**	
	Sequeiro	
	Sem SMOTE	**Com SMOTE**
	Acurácia	*Acurácia*
Rede Neural	-	
Árvore de Decisão	-	52,48
Random Forest	55,97	-
KNN	-	-

Fonte: Do Autor (2023).

APÊNDICE C – Representação Base de Dados

Nesta seção, a representação da base de dados é exibida na Tabela 7 e Tabela 8, para manejo irrigado e sequeiro.

Tabela 7 – Representação Base de Dados - Irrigado.

Mês de Coleta	*Ano de Coleta*	*Genótipo*	*Repetição*	*Reflectância*	*Potencial Hídrico*
4	2014	7	3	2,5341	-0,2
4	2014	12	2	4,2734	-0,2
4	2014	19	2	3,6884	-0,2
⋮	⋮	⋮	⋮	⋮	⋮
9	2016	20	1	4,7708	-4,6
9	2016	5	1	2,8035	-5,0
9	2016	20	3	5,1406	-5,1

Fonte: Próprio Autor.

Tabela 8 – Representação Base de Dados - Sequeiro.

Mês de Coleta	*Ano de Coleta*	*Genótipo*	*Repetição*	*Reflectância*	*Potencial Hídrico*
4	2014	20	2	3,2350	-0,25
4	2014	20	3	4,5399	-0,25
3	2015	3	2	3,7777	-0,25
⋮	⋮	⋮	⋮	⋮	⋮
9	2014	21	3	4,4712	-6,6
9	2014	21	4	3,2956	-6,6
9	2014	23	1	2,3725	-6,6

Fonte: Próprio Autor.