Modelagem de Variáveis Espectrais para Determinar as Condições Hídricas de Cafeeiros

Deyvis Cabrini Teixeira Delfino
Departamento de Automática
Universidade Federal de Lavras
Lavras, Brasil
deyvis.delfino@estudante.ufla.br

Renan Teixeira Delfino
Recursos Hídricos
Universidade Federal de Lavras
Lavras, Brasil
renan.delfino4@estudante.ufla.br

Margarete Marin Lordelo Volpato
Empresa de Pesquisa Agropecuária de Minas Gerais
EPAMIG
Lavras, Brasil
margarete@epamig.ufla.br

Vânia Aparecida Silva
Empresa de Pesquisa Agropecuária de Minas Gerais
EPAMIG
Lavras, Brasil
vania.silva@epamig.ufla.br

Danton Diego Ferreira
Departamento de Automática
Universidade Federal de Lavras
Lavras, Brasil
danton@ufla.br

***Resumo*—O potencial hídrico é um importante indicador utilizado para estudar as relações hídricas nas plantas, pois reflete o nível de hidratação de seus tecidos. Tradicionalmente, esse indicador é medido diretamente por meio de um equipamento chamado Bomba de Scholander, embora esse processo seja complexo e demorado. Contudo, há várias variáveis numéricas que descrevem as propriedades das plantas e que podem ser adquiridas a partir dos índices de refletância das folhas. Essas variáveis apresentam relações diretas e indiretas com o potencial hídrico. Neste estudo, o objetivo é explorar variáveis espectrais para estimar o potencial hídrico em cafeeiros, utilizando ferramentas de inteligência computacional. Para isso, foram medidas assinaturas espectrais por meio de miniespectrômetro em lavoura de café localizada em Diamantina, cidade localizada no estado de Minas Gerais. Os dados contemplados abrangem o período de 2014, 2015 e 2016. Ademais, os dados apresentam dois grupos de manejo da lavoura, irrigado e sequeiro. Através da plataforma de desenvolvimento de algoritmo *MATLAB* foram desenvolvidas quatro técnicas de *Machine Learning*: Rede Neural Artificial tipo MLP (*Multi-Layer Perceptron*), Árvore de Decisão, *Random Forest* e *KNN* (*K-Nearest Neighbor*). Foram implementados para as quatro técnicas dois métodos distintos, estimação e classificação. Os resultados expõem que as redes neurais artificias foram superiores em ambos os métodos, com raiz do erro quadrático médio (*RMSE*) de 0,4361 e o coeficiente de determinação (R^2) de 0,6923 para estimador e acurácia global 73,3% para o classificador.**

***Index Terms*—Cafeicultura, Aprendizado de Máquina, Potencial hídrico, Análise de Dados.**

I. Introdução

Localizado na quinta posição dos produtos de origem vegetal mais exportados pelo Brasil, o café é uma das commoditys que mais contribuíram para a expansão das vendas externas do agronegócio em 2022, 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.

Com um mercado consumidor cada vez mais exigente, além do volume de sacas produzidas a qualidade do grão de café é algo crucial para garantir uma maior fatia do mercado. Uma maneira de assegurar a alta qualidade do produto é o conhecimento das relações hídricas da planta, com o intuito de mantê-la a mesma sempre hidratada, garantindo assim um produto final de alta qualidade.

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. 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 [1].

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. 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 reportado em [2] foi proposto determinar o efeito do estresse hídrico no milho (*Zea mays L.*) usando índices espectrais, leituras de clorofila e, consequentemente, avaliar os espectros de reflectância. Similarmente, no estudo [3] utilizou-se amostras de duas lavouras de café e índices espectrais para determinar as condições hídricas dos cafeeiros.

Com o intuito de explorar uma abordagem distinta dos trabalhos de [2] e [3], 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 assinatura espectral e investigar qual é o comprimento de onda ou a faixa de comprimentos de onda mais adequada para inferir o potencial hídrico do cafeeiro.

O presente trabalho aborda o desenvolvimento de quatro métodos de *machine learning* para estimar e classificar o potencial hídrico de cafeeiros: Rede Neural Artificial tipo *MLP (Multi-Layer Perceptron)*, Árvore de Decisão, *Random Forest* e *KNN* (*K-Nearest Neighbor*). Em relação às redes neurais artificiais, há inúmeros trabalhos na literatura que fundamentam sua utilização em problemas de classificação e regressão. Dois dos empecilhos para uma RNA eficiente são o número de camadas intermediarias da rede e a quantidade de nós em cada camada. Porém, quando bem projetadas, as redes neurais artificiais resultam em uma resposta adequada tanto para classificação quanto para regressão [4] [17].

Outra ferramenta capaz de solucionar problemas de classificação e regressão é a árvore de decisão, em que sua estrutura em formato de árvore é composta por um conjunto de nós interconectados. Os nós internos testam os atributos de entrada como constantes de decisão e determinam qual será o próximo nó descendente [5].

Com a capacidade de aumentar a complexidade considerando novos dados, porém sem reduzir sua competência de generalização, a *Random Forest* é uma técnica *Ensemble* disponíveis na literatura. Os Métodos *Ensemble* são algoritmos formados por uma coleção de classificadores, nesse caso, um conjunto de árvores de decisão, que realizam o processo de votação das classes por meio do voto majoritário [6].

Por fim, outro recurso relevante contido na literatura é o algoritmo *K-Nearest Neighbor* (*KNN*), do português K-Vizinhos Mais Próximos, pertencente à família de algoritmos *Instance-based Learning* (*IBL*), ou seja, aprendizagem baseada em instâncias. O *kNN* localiza as *k* instâncias mais próximas da instância de consulta e determina sua classe identificando o único rótulo de classe mais frequente [7].

II. Materiais e Métodos

A. Banco 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. A base de dados foi fornecida pela equipe de pesquisadores de campo da EPAMIG (Empresa de Pesquisa Agropecuária de Minas Gerais) e apresentam características espectrais de cafeeiros colhidas através do equipamento Mini-espectrômetro foliar CI-710.

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. O número de amostras contidas no banco de dados são de 445 irrigado e 450 sequeiro, os dois tipos de manejo estão separados em dois bancos de dados distintos e não foi implementado a mescla dos bancos. Cada amostra é composta por 2863 atributos, que são formados pela data de coleta, o número do genótipo/cultivar, o número da repetições dentro do genótipo/cultivar e a sequência da reflectância que corresponde ao comprimento de onda na faixa de 400 á 950nm. Além do mais, ambas as bases de dados possuem como alvo o potencial hídrico (Ψ_W) medido com uma Bomba de Scholander.

B. Pré Processamento

Para o início do pré-processamento será adotado o método de filtragem por mediana desenvolvido por [8], que consiste em suavizar ruido do tipo impulsivo em sinais e imagens digitais [9]. 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 "problemático" pela mediana da janela. O filtro implementado no presente trabalho é 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 será utilizada a Equação (1).

$$P_n = \frac{(P - P_{min})}{(P_{max}) - (P_{min})}, \tag{1}$$

em que P_n corresponde o valor normalizado da variável n, P, P_{min} e P_{max} ao valor original, o menor e o maior valor respectivamente [4].

Outro ponto importante do pré-processamento é a seleção de características. Para esta etapa serão utilizadas duas técnicas. A primeira denominada coeficiente de *Pearson*, que mede o grau da correlação linear entre duas variáveis. Este coeficiente, normalmente representado por ρ, assume apenas valores entre -1 e 1. A Tabela I exibe a interpretação dos valores do coeficiente de *Pearson* (ρ) [10]. A segunda técnica é o algoritmo *Minimum Redundancy Maximum Relevance* (*MRMR*), que ranqueia os dados do mais relevante para o menos relevante usando uma análise de correlação das variáveis, visando identificar redundâncias e a relevância de cada característica [11], [12].

TABELA I
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

O valor do coeficiente de *Pearson* (ρ) foi calculado entre os atributos e a variável alvo (potencial hídrico). Adotou-se o limiar de $\rho = 0,2$ para dados em condição de irrigado, ou seja, atributos com $\rho < 0,2$ foram considerados irrelevantes para o modelo. Com este limiar, os atributos passaram de 2863 dimensões para 465. Já para os dados em condição de sequeiro, o coeficiente de *Pearson* (ρ) teve seu valor selecionado em 0,4, o que representa uma correlação moderada. Os atributos passaram de 2863 dimensões para 53 neste caso. Por fim, o método *Minimum Redundancy Maximum Relevance* (*MRMR*), que ordena os atributos do mais relevante para o menos relevante foi executado e foram extraídos os 10 atributos mais relevantes para ambas as condições, que são mostrados nas Tabelas II e III, para manejo irrigado e sequeiro, respectivamente.

TABELA II
OS 10 ATRIBUTOS MAIS RELEVANTES PARA IRRIGADO.

***Ranking* dos atributos**	**Designação do atributo**
1	mês de coleta
2	reflectância para λ = 657,220nm
3	reflectância para λ = 803,569nm
4	ano de coleta
5	reflectância para λ = 748,703nm
6	reflectância para λ = 6687,210nm
7	reflectância para λ = 776,964nm
8	reflectância para λ = 656,436nm
9	reflectância para λ = 802,283nm
10	reflectância para λ = 658,199nm

em que (λ) corresponde a comprimento de onda.

TABELA III
OS 10 ATRIBUTOS MAIS RELEVANTES PARA SEQUEIRO.

***Ranking* dos atributos**	**Designação do atributo**
1	mês de coleta
2	reflectância para λ = 695,325nm
3	reflectância para λ = 685,855nm
4	reflectância para λ = 691,657nm
5	reflectância para λ = 692,816nm
6	reflectância para λ = 686,823nm
7	reflectância para λ = 694,553nm
8	reflectância para λ = 687,597nm
9	reflectância para λ = 692,430nm
10	reflectância para λ = 690,691nm

em que (λ) corresponde a comprimento de onda..

Para a implementação das técnicas de *Machine Learning* visando a classificação é necessário estabelecer classes para os potenciais hídricos das amostras. As classes foram estipuladas segundo o trabalho de [13] e são mostradas na Tabela IV.

TABELA IV
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 divisão desta partição é feita utilizando a técnica de Validação Cruzada [15] através da plataforma de desenvolvimento de algoritmo *MATLAB*, em que o número de *folds* escolhido foi 10, de forma que para cada *fold* é escolhido para teste e os outros 9 são selecionados para treino.

C. Métricas de Avaliação e Desempenho

Com o objetivo de avaliar o desempenho dos classificadores, as matrizes de confusão correspondentes foram geradas, permitindo o estudo comparativo dos resultados de cada uma das técnicas propostas. Por fim, será empregado como parâmetro a acurácia global que utiliza soma da diagonal principal da matriz de confusão.

Para as técnicas de regressão foi utilizada a raiz do erro quadrático médio, conhecida como *Root Mean Squared Error* (*RMSE*), exibido na Equação (2), que calcula a magnitude da média do erro pela raiz quadrada da média dos quadrados dos erros. Desse modo, atribui-se um peso maior aos erros de maior magnitude e um peso menor aos erros de menor magnitude. Também foi empregado como métrica o coeficiente de determinação, conforme Equação (3), que pode ser compreendido como o percentual da variação dos dados que é explicado pelo modelo. Assim, quanto maior o R^2, mais explicativo é o modelo, ou seja, melhor ele se ajusta aos dados.

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

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

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y})^2}, \tag{3}$$

em que $\overline{y}$ é o valor médio das observações.

III. RESULTADOS E DISCUSSÕES

Após a execução das etapas de pré-processamento e seleção de características, deu-se inicio a implementação dos modelos de *Machine Learning*. Inicialmente, as técnicas Rede Neural Artificial tipo MLP, Árvore de Decisão, *Random Forest* e *KNN*[1] foram executados. É importante salientar que todas as

[1]Levando em conta o método de regressão para técnica *KNN* foi utilizado o algoritmo presente no trabalho reportado em [14].

técnicas foram implementadas para todos os atributos após seleção de características via coeficiente de *Pearson*, também para os 10 e os 5 mais relevantes. Diferentes parâmetros para uma mesma técnica foram implementados com o intuito de otimizar e fornecer o melhor resultado. Cada *fold* dispôs de 30 execuções, totalizando 300 iterações por método de *Machine Learning*.

A. Estimadores

1) Irrigado: Foram extraídos os valores da raiz do erro quadrático médio (*RMSE*) e do coeficiente de determinação R^2 dos 10 *folds* das técnicas de *Machine Learning*. Os resultados da média (μ) e do desvio padrão (σ) das métricas de avaliação são exibidos nas Tabelas V, VI e VII, para os 465, 10 e 5 atributos mais relevantes, respectivamente.

TABELA V
COEFICIENTES DE DESEMPENHO PARA 465 ATRIBUTOS (IRRIGADO).

Método	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	0,6003	±0,0849	0,4937	±0,0899
Árvore de Decisão	0,6260	±0,1306	0,4651	±0,1394
Random Forest	0,6101	±0,1063	0,4742	±0,1103
KNN	0,7743	±0,0816	0,1953	±0,0658

TABELA VI
COEFICIENTES DE DESEMPENHO PARA 10 ATRIBUTOS MAIS RELEVANTES (IRRIGADO).

Método	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	0,5885	±0,0865	0,5098	±0,0963
Árvore de Decisão	0,6039	±0,0939	0,4842	±0,1160
Random Forest	0,6051	±0,1101	0,4895	±0,1204
KNN	0,6421	±0,1033	0,4331	±0,1234

TABELA VII
COEFICIENTES DE DESEMPENHO PARA 5 ATRIBUTOS MAIS RELEVANTES (IRRIGADO).

Método	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	0,5884	±0,0887	0,5123	±0,0946
Árvore de Decisão	0,5990	±0,0921	0,4872	±0,1084
Random Forest	0,6241	±0,1155	0,4701	±0,1248
KNN	0,6307	±0,0976	0,4555	±0,1245

Analisando os resultados das Tabelas V, VI e VII, nota-se que a Rede Neural Artificial alcançou o melhor resultado dentre as 3 divisões de atributos, com o valor 0,5884 ±8,87% de média da raiz do erro quadrático (*RMSE*) e 0,5123 ±9,46% para coeficiente de determinação (R^2), valores esses correspondentes a organização com os 5 atributos mais relevantes.

Considerando o *fold* de melhor desempenho entre os 10 da técnica melhor qualificada, os parâmetros obtidos são rede neural de única camada intermediaria, com 2 neurônios e função de ativação Sigmoide. Com o intuito de edificar a exibição dos resultados obtidos e possibilitar uma melhor análise, a Fig. 1 ilustra a distribuição dos dados 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 0,4361 e o coeficiente de determinação (R^2) de 0,6923.

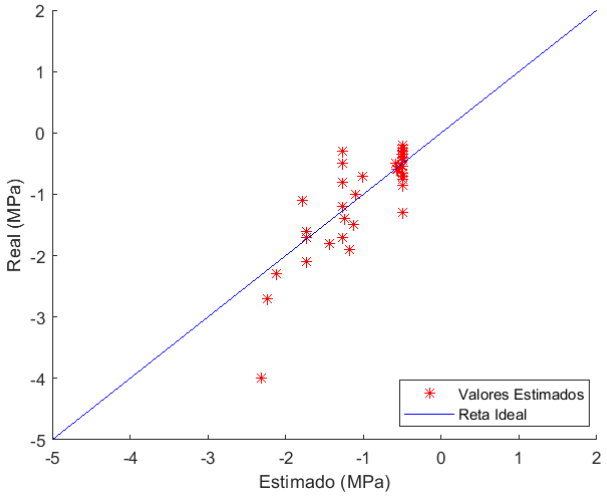

Figura 1. Dados Reais x Dados Estimados (*fold* de melhor desempenho - Irrigado - cinco atributos mais relevantes).

2) Sequeiro: De maneira análoga as amostras em condição irrigada, as técnicas de *Machine Learning* foram implementadas para os dados em estado sequeiro. Foram considerados todos os 53 atributos, posteriormente os 10 e 5 mais relevantes. A média e desvio padrão das técnicas de desempenho foram levantas e exibidas nas Tabelas VIII, IX e X para os 53, 10 e 5 atributos mais relevantes.

TABELA VIII
COEFICIENTES DE DESEMPENHO PARA 53 ATRIBUTOS (SEQUEIRO).

Método	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	1,0647	±0,1122	0,5431	±0,0863
Árvore de Decisão	1,1384	±0,1509	0,4689	±0,1271
Random Forest	1,1324	±0,0943	0,4742	±0,0784
KNN	1,2631	±0,1310	0,3616	±0,0944

TABELA IX
COEFICIENTES DE DESEMPENHO PARA 10 ATRIBUTOS MAIS RELEVANTES (SEQUEIRO).

Método	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	1,1239	±0,1413	0,4802	±0,1318
Árvore de Decisão	1,1070	±0,1209	0,5016	±0,1050
Random Forest	1,1094	±0,1159	0,4940	±0,1080
KNN	1,1578	±0,1647	0,4564	±0,1337

Com base nas Tabelas VIII, IX e X, a técnica Rede Neural Artificial apresenta o melhor resultado considerando os 53 atributos no estado sequeiro com o valor 1,0647 ±11,22% de média da raiz do erro quadrático (*RMSE*) e 0,5431 ±8,63% para coeficiente de determinação (R^2). De modo a exemplificar o resultado, a Fig. 2 exibe o comparativo entre os dados reais

TABELA X
COEFICIENTES DE DESEMPENHO PARA 5 ATRIBUTOS MAIS RELEVANTES (SEQUEIRO).

Método	μ_{RMSE}	σ_{RMSE}	μ_{R^2}	σ_{R^2}
Rede Neural	1,1102	±0,1153	0,4884	±0,1027
Árvore de Decisão	1,1572	±0,1614	0,4668	±0,1389
Random Forest	1,1212	±0,1029	0,4861	±0,0858
KNN	1,1216	±0,1447	0,4871	±0,1166

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 3 neurônios e função de ativação Tangente Hiperbólica. Em que a raiz do erro quadrático médio (*RMSE*) é de 0,8752 e o coeficiente de determinação (R^2) de 0,6987.

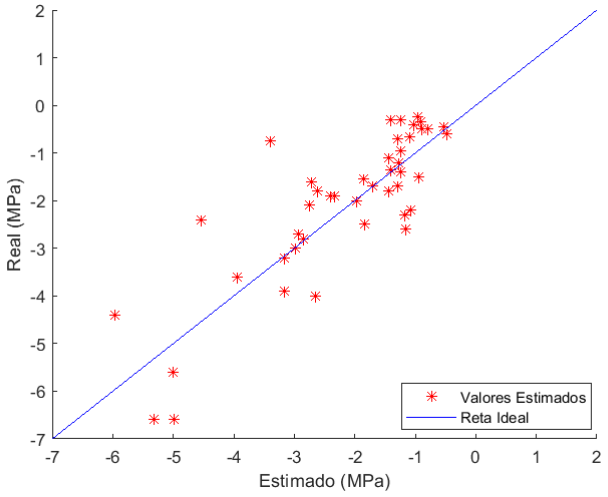

Figura 2. Dados Reais x Dados Estimados (*fold* de melhor desempenho - Sequeiro - 53 atributos).

B. Classificação

1) Irrigado: Para o procedimento de classificação, a métrica utilizada é a acurácia global, que gera a porcentagem de acerto da técnica de *Machine Learning* implementada. As Tabelas XI, XII e XIII exibem a média (μ) e o desvio padrão (σ) da acurácia global dos métodos implementados considerando as 3 variações de atributos para condição de irrigado.

TABELA XI
COEFICIENTES DE DESEMPENHO PARA 465 ATRIBUTOS (IRRIGADO).

Método	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	56,67	±5,24
Árvore de Decisão	61,11	±7,72
Random Forest	58,20	±5,10
KNN	55,77	±6,57

Mediante os resultados das Tabelas XI, XII e XIII, a Rede Neural Artificial alimentada com os 5 atributos mais relevantes

TABELA XII
COEFICIENTES DE DESEMPENHO PARA 10 ATRIBUTOS MAIS RELEVANTES (IRRIGADO).

Método	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	59,79	±4,84
Árvore de Decisão	60,91	±5,55
Random Forest	60,65	±6,10
KNN	61,11	±6,62

TABELA XIII
COEFICIENTES DE DESEMPENHO PARA 5 ATRIBUTOS MAIS RELEVANTES (IRRIGADO).

Método	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	62,22	±6,45
Árvore de Decisão	60,00	±5,90
Random Forest	61,34	±3,94
KNN	60,88	±7,64

demonstra os melhores resultados com média de acerto de 62,22% ±6,45%. O *fold* de melhor resultado levando em conta a técnica de melhor desempenho possui como parâmetro uma rede neural de 3 camadas intermediárias, com 4 neurônios na primeira camada, 2 na segunda e 66 na terceira, e função de ativação Unidade Linear Retificada - do inglês *Rectified Linear Unit* (*ReLU*). A Fig. 3 exibe a matriz de confusão do *fold* com melhor desempenho.

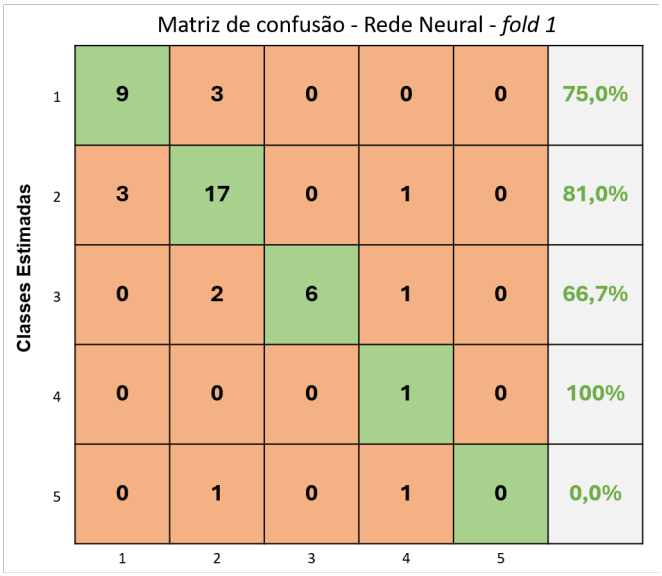

Figura 3. Matriz de confusão 05 mais relevantes (*fold* de melhor desempenho - Irrigado).

Por meio da análise da matriz de confusão exibida na Fig. 3 a acurácia total do modelo para o *fold* 1 é de 73,3% e é viável 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.

2) Sequeiro: De forma similar às amostras em condição irrigada, foram apresentadas aos dados em estado sequeiro as técnicas de *Machine Learning* para o método de classificação. As Tabelas XIV, XV e XVI exibem a média (μ) e o desvio padrão (σ) da acurácia global dos métodos implementados atendendo às três alternativas de atributos.

TABELA XIV
COEFICIENTES DE DESEMPENHO PARA 53 ATRIBUTOS (SEQUEIRO).

Método	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	48,09	±6,52
Árvore de Decisão	49,66	±6,96
Random Forest	48,77	±5,59
KNN	48,33	±6,58

TABELA XV
COEFICIENTES DE DESEMPENHO PARA 10 ATRIBUTOS MAIS RELEVANTES (SEQUEIRO).

Método	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	47,57	±6,80
Árvore de Decisão	48,51	±6,72
Random Forest	50,09	±5,33
KNN	47,83	±5,89

TABELA XVI
COEFICIENTES DE DESEMPENHO PARA 5 ATRIBUTOS MAIS RELEVANTES (SEQUEIRO).

Método	$\mu_{Acurácia}$	$\sigma_{Acurácia}$
Rede Neural	50,31	±6,38
Árvore de Decisão	50,58	±6,39
Random Forest	50,29	±7,15
KNN	46,75	±4,63

Atendendo a condição sequeiro, a técnica de árvore de decisão ostenta os melhores resultados de acordo com as Tabelas XIV, XV e XVI. Dispondo de uma média de acerto de 50,58% ±6,39%, em que os 5 atributos mais relevantes despontam com os melhores resultados. O melhor modelo da técnica de árvores de decisão possui um nível de poda de 13 nós, número de ramificações de 6, valor mínimo de observações por folha de 2 e critério de divisão a Entropia Cruzada. A Fig. 4 exibe a matriz de confusão do *fold* com melhor desempenho para este caso.

Equivalente aos dados em condição de irrigado, a acurácia total do método para o *fold* 4 é de 60,0%. A não classificação considerando as amostras da classes 4 possui como prováveis causas o desbalanceamento das amostras e a similaridade dos atributos da classe 4 para as demais.

Convertendo os números das classes para valores de potencial hídrico via Tabela IV, é perceptível que a faixa entre

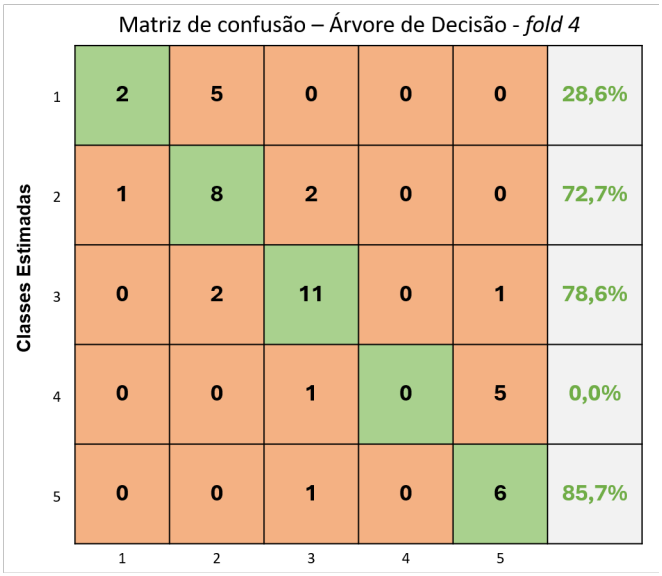

Figura 4. Matriz de confusão 05 mais relevantes (*fold* de melhor desempenho - Sequeiro).

-2,5MPa e -5,1MPa contem as amostras pertencentes a classes com desbalanceamento.

C. Estudo Comparativo

Considerando a importância do potencial hídrico foliar, bem como sua correlação com índices espectrais, a literatura contempla alguns trabalhos que visam a estimação e classificação do (Ψ_W) sem a necessidade de sua complexa medição direta, proporcionando uma excelente base de comparação.

Por meio da análise das apurações do procedimento de estimação atendendo a ambos os estados (irrigado e sequeiro) das amostras, os resultados apontam para um desempenho inferior da Rede Neural Artificial tendo em consideração o coeficiente de determinação (R^2) quando comparado com o trabalho reportado em [16]. O trabalho reportado em [16] coletou dados para estimar a evapotranspiração em 17 posições geográficas distintas no estado do Rio de Janeiro. Para essa finalidade, foi utilizado um modelo de rede neural MLP (Perceptron de Múltiplas Camadas), que recebeu como entrada a latitude, longitude, altitude, temperatura média do ar, amplitude térmica diária e dia do ano. Por meio desse processo, foi possível realizar a estimativa da evapotranspiração para as 17 localidades. O desempenho inferior da Rede Neural Artificial é possivelmente justificada por alguma singularidades das amostras utilizadas.

Tendo em vista a acurácia global como métrica de avaliação para o método de classificação, o resultado obtido no presente estudo é deveras inferior quando confrontado com o trabalho reportado em [3], que por sua vez, utilizou de índices espectrais extraídos dos comprimentos de onda de medições espectrais feitas em campo, o que requer maior complexidade computacional.

IV. Conclusão

De modo geral, os resultados obtidos com a aplicação das metodologias mencionadas 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.

A técnica Rede Neural Artificial alcançou o melhor desempenho para os dados em ambos os estados (irrigado e sequeiro) considerando o método estimação. Entretanto, é possível que o desempenho possa ter sido prejudicado devido ao baixo número de amostras contidas na faixa de potencial hídrico entre -2,5MPa e -5,1MPa.

Levando em consideração o método de classificação, duas técnicas distintas destacaram-se com os melhores desempenhos. A Árvore de Decisão para os dados em estado sequeiro e a Rede Neural Artificial com os dados em condições irrigada. Foi observado que a divisão dos dados em classes resulta em um desequilíbrio na base de dados ocasionando em o pouco número de amostras das classes 4 e 5. Esse problema pode ser abordado através do uso de algoritmos de *oversampling*, os quais pretende-se aplicar em trabalhos futuros.

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.

Agradecimentos

Agradeço principalmente 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.

Referências

[1] ZHANG, Chao; PATTEY, Elizabeth; LIU, Jiangui; CAI, Huanjie; SHANG, Jiali; DONG, Taifeng. Retrieving Leaf and Canopy Water Content of Winter Wheat Using Vegetation Water Indices. Ieee Journal Of Selected Topics In Applied Earth Observations And Remote Sensing, [S.L.], v. 11, n. 1, p. 112-126, jan. 2018. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/jstars.2017.2773625.

[2] GENC, Levent; INALPULAT, Melis; KIZIL, Unal; MIRIK, Mustafa; SMITH, Scot E.; MENDES, Mehmet. Determination of water stress with spectral reflectance on sweet corn (Zea mays L.) using classification tree (CT) analysis. Zemdirbyste-Agriculture, [S.L.], v. 100, n. 1, p. 81-90, 30 mar. 2013. Lithuanian Research Centre for Agriculture and Forestry. http://dx.doi.org/10.13080/z-a.2013.100.011

[3] NUNES, Pedro Henrique et al. Predicting coffee water potential from spectral reflectance indices with neural networks. Smart Agricultural Technology, [S.L.], v. 4, p. 100213-100219, mar. 2023. Elsevier BV. http://dx.doi.org/10.1016/j.atech.2023.100213. Disponível em: https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1152292/1/Predicting-coffee-water-potential.pdf. Acesso em: 16 abr. 2023.

[4] BRAGA, Antonio de Padua; LUDERMIR, Teresa Bernarda; CARVALHO, Andre Carlos Ponce de Leon Ferreira de. Redes neurais artificiais: teoria e aplicações. 2. ed. Rio de Janeiro: Ltc, 2007. 248 p.

[5] WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A.. Data Mining: practical machine learning tools and techniques. 3. ed. Burlington: Morgan Kaufmann Publishers, 2011. 629 p.

[6] BREIMAN, Leo. Random Forests. Machine Learning, [S.L.], v. 45, n. 1, p. 5-32, 2001. Springer Science and Business Media LLC. http://dx.doi.org/10.1023/a:1010933404324.

[7] FARIA, Mauricio Mendes. Detecção de Intrusões em Redes de Computadores com Base nos Algoritmos KNN, K-Means++ e J48. 2016. 146 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Faculdade Campo Limpo Paulista., Campo Limpo Paulista, 2016.

[8] TUKEY, John W.. Exploratory Data Analysis. New York: Pearson, 1977. 503 p.

[9] PRATT, William K.. DIGITAL IMAGE PROCESSING: piks inside. 3. ed. New York: John Wiley Sons, Inc., 2001. 738 p.

[10] VIEIRA, Sonia. Introdução à Bioestatística. 4. ed. Rio de Janeiro: Elsevier, 2008. 357 p.

[11] DING, C.; PENG, H.. Minimum redundancy feature selection from microarray gene expression data. Computational Systems Bioinformatics. Csb2003. Proceedings Of The 2003 Ieee Bioinformatics Conference. Csb2003, [S.L.], v. 3, n. 2, p. 523-528, 08 set. 2003. IEEE Comput. Soc. http://dx.doi.org/10.1109/csb.2003.1227396.

[12] DARBELLAY, G.A.; VAJDA, I.. Estimation of the information by an adaptive partitioning of the observation space. Ieee Transactions On Information Theory, [S.L.], v. 45, n. 4, p. 1315-1321, maio 1999. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/18.761290.

[13] SILVA, Vânia Aparecida; VOLPATO, Margarete Marin Lordelo; FIGUEIREDO, Vanessa Castro; PEREIRA, Alessandro Botelho; MATOS, Christiano Sousa Machado de; SANTOS, Meline de Oliveira. Impacto do déficit hídrico e temperaturas elevadas sobre o estado hídrico do cafeeiro nas regiões Sul e Cerrado de Minas Gerais. 356. ed. Belo Horizonte: Epamig, 2021. 5 p.

[14] CONSONNI, Viviana; BACCOLO, Giacomo; GOSETTI, Fabio; TODESCHINI, Roberto; BALLABIO, Davide. A MATLAB toolbox for multivariate regression coupled with variable selection. Chemometrics And Intelligent Laboratory Systems, [S.L.], v. 213, p. 104313, jun. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.chemolab.2021.104313.

[15] OJALA, Markus; GARRIGA, Gemma C.. Permutation Tests for Studying Classifier Performance. 2009 Ninth Ieee International Conference On Data Mining, Miami, p. 908-913, 06 dez. 2009. IEEE. http://dx.doi.org/10.1109/icdm.2009.108.

[16] ZANETTI, Sidney S.; SOUSA, Elias F.; CARVALHO, Daniel F. de; BERNARDO, Salassier. Estimação da evapotranspiração de referência no estado do Rio de Janeiro usando redes neurais artificiais. Revista Brasileira de Engenharia Agrícola e Ambiental, [S.L.], v. 12, n. 2, p. 174-180, abr. 2008. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s1415-43662008000200010.

[17] CHEN, Yingyi; SONG, Lihua; LIU, Yeqi; YANG, Ling; LI, Daoliang. A Review of the Artificial Neural Network Models for Water Quality Prediction. Applied Sciences, [S.L.], v. 10, n. 17, p. 5776, 20 ago. 2020. MDPI AG. http://dx.doi.org/10.3390/app10175776.