ESTIMATIVA DE ALCALÓIDES EM TABACO UTILIZANDO ESPECTROSCOPIA VIS-NIR-SWIR E MACHINE LEARNING

João Vitor Ferreira Gonçalves [1], Karym Mayara de Oliveira[2], Renato Herrig Furlanetto[3], Caio Almeida de Oliveira[4], Weslei Augusto Mendonça[5], Marcos Rafael Nanni[6], Marcelo Luiz Chicati[7].

[1]Mestrando, Universidade Estadual de Maringá, Avenida Colombo, 5790, Maringá – PR, joaofergo@gmail.com; [2]Doutoranda, Universidade Estadual de Maringá, Avenida Colombo, 5790, Maringá – PR, eng.karymoliveira@gmail.com; [3]Pós Doutorando, Universidade da Flórida, Estados Unidos, re.herrigfurlane@ufl.edu; [4]Mestrando, Universidade Estadual de Maringá, Avenida Colombo, 5790, Maringá – PR, caio_fck@outlook.com; [5]Graduando, Universidade Estadual de Maringá, Avenida Colombo, 5790, Maringá – PR, ra113400@uem.br; [6]Docente, Universidade Estadual de Maringá, Avenida Colombo, 5790, Maringá – PR, mrnanni@uem.br; [7]Docente, Universidade Estadual de Maringá, Avenida Colombo, 5790, Maringá – PR, mlchicati@uem.br.

RESUMO

O tabaco é uma cultura importante para economia brasileira e a estimativa de atributos bioquímicos foliares está diretamente ligada ao seu valor agregado. O objetivo desse trabalho foi avaliar a estimativa do teor de alcaloides presente em tabaco, utilizando espectroscopia aliada a modelos de *machine learning*. O experimento foi realizado em estufa, a reflectância foi coletada aos 41 e 69 dias após plantio, concomitantemente foram realizadas análises químicas de alcalóides totais. Os modelos de *machine learning* utilizados foram PLS *(Partial Least Square*), *Random Forest* (RF) e *Neural Network* (NN). As métricas de avaliação dos modelos foram coeficiente de determinação (R^2), *Root Mean Squared Error* (RMSE) e Razão de Previsão para Desvio (RPD). Os resultados obtidos na fase de predição foram R^2 0.68, 0.64 e 0.71 para PLS, RF e NN, respectivamente e RMSE de 0.47, 0.49 e 0.46, respectivamente. Esses resultados sugerem o potencial do uso de espectroscopia no monitoramento de alcaloides em tabaco.

Palavras-chave **—** Alcalóides, refletância espectral, machine learning, quimiometria.

ABSTRACT

Tobacco is an important crop for the Brazilian economy and the estimation of leaf biochemical attributes is directly linked to its added value. The objective of this work was to evaluate the estimation of the content of alkaloids present in tobacco, using spectroscopy combined with machine learning models. The experiment was carried out in a greenhouse, the reflectance was collected at 41 and 69 days after planting, concomitantly, chemical analyzes of total alkaloids were performed. The machine learning models used were PLS (Partial Least Square), Random Forest (RF) and Neural Network (NN). The evaluation metrics of the models were the coefficient of determination (R^2), Root Mean Squared Error (RMSE) and Ratio of Prediction to Deviation (RPD). The results obtained in the prediction phase were R2 0.68, 0.64 and 0.71 for PLS, RF and NN, respectively, and RMSE of 0.47, 0.49 and 0.46, respectively. These results suggest the potential of using spectroscopy to monitor alkaloids in tobacco.

Key words — Alkaloids, spectral reflectance, machine learning, chemometrics.

1. INTRODUÇÃO

A cultura do Tabaco (*Nicotiana tabacum L.*) no Brasil é relevante socioeconomicamente, e movimenta principalmente a agricultura familiar devido aos métodos de cultivo e grande mão de obra necessária, o que potencializa o valor agregado. A produção total de mais de 590 mil toneladas de tabaco obtida no ano safra 2021/2022, exportação em mais de 464 mil toneladas, faturamento total em mais de 26 bilhões de reais e mais de 2 milhões de empregos diretos e indiretos gerados no setor, confirmam sua importância para a economia do país [1].

Do ponto de vista científico destacam-se as pesquisas que são relacionadas a composição bioquímica do tabaco, pois a identificação e quantificação dos compostos são diretamente relacionados a qualidade comercial do produto. De todas as propriedades bioquímicas estudadas, a que se destaca é a quantidade total de alcalóides, onde a principal fração é a Nicotina (95% do total) [2]. Fatores ambientais, genéticos, manejo e processos de maturação, são influenciadores dos teores de alcalóides presentes na folha.

Visto a qualidade e valor agregado ao produto, além da obrigatoriedade legal, a indústria do tabaco tem interesse direto no monitoramento dessas propriedades bioquímicas. Na prática, esses valores são mensurados por rotinas de análises químicas laboratoriais cujos resultados são precisos, contudo, demanda elevada mão de obra a campo e em laboratório, além da utilização de reagentes, os quais podem gerar resíduos nocivos ao meio ambiente. Neste contexto, o sensoriamento remoto se enquadra como ferramenta de monitoramento de safras que demanda reduzida quantidade de pessoal, além de constituir-se em método não destrutivo e com rápida entrega de resultados [3].

Técnicas de espectroscopia no visível (Vis), infravermelho próximo (NIR) e infravermelho de ondas curtas (SWIR) permitem coletar a assinatura espectral da folha, e essa assinatura expressa determinadas propriedades bioquímicas, sendo possível a estimativa de clorofila, compostos nitrogenados, dentre outros. Diversos pesquisadores têm buscado aplicações das técnicas de espectroscopia na agricultura com objetivo de avaliação de deficiência nutricional de plantas, estimativa de produtividade e até mesmo avaliação de estresse hídrico. Quanto a cultura do tabaco, autores buscam determinar teores de açucares, alcalóides e outros compostos bioquímicos [2, 4, 5] relevantes ao valor agregado a cultura.

A estimativa de propriedades bioquímicas de plantas por meio da espectroscopia Vis-NIR-SWIR permite monitoramento de safras e otimização do processo produtivo. Neste contexto, este trabalho tem como objetivo avaliar a estimativa do teor de alcalóides presente em folhas de tabaco, utilizando-se de espectroscopia Vis-NIR-SWIR aliados a modelos de *machine learning*.

2. MATERIAL E MÉTODOS

2.1. Experimento

O experimento foi realizado na cidade de Mafra, região sul do Brasil, em ambiente com umidade e temperatura (±26ºC) controlados. O delineamento utilizado foi inteiramente casualizado, composto por 30 tratamentos (variedades) e 10 repetições.

Para determinação laboratorial da concentração de alcalóides, foi coletada a terceira folha totalmente expandida de cada planta, aos 41 e 69 dias após plantio (DAP). A extração em laboratório foi realizada por éter metil terc – butílico (MTBE) e a análise feita por cromatografia gasosa com detecção de ionização por chama (GC/FID).

Os dados espectrais foram coletados com espectrorradiômetro FieldSpec 3 Jr. (Analytical Spectral Devices, Boulder, CO, EUA). A obtenção de dados espectrais foi realizada nas mesmas folhas e datas utilizadas para análise química. Foram obtidos valores médios de reflectância em 3 pontos por folha (ponta, meio e base da folha), de todas as plantas, totalizando 300 leituras por dia, e 600 leituras ao longo do ciclo da cultura. O intervalo espectral analisado foi de 350 a 400 nm e 2450 a 2500 nm, totalizando 2050 comprimentos de onda.

A análise descritiva do teor de alcalóides foi realizada através do software XLSTAT (Addinsoft, 2022, student version). Já a construção dos algorítmos de machine learning foram realizados em linguagem Python, através da plataforma de código aberto Jupyter Notebook e bibliotecas Numpy, Pandas, Scikit Learning e TensorFlow.

Os dados foram divididos aleatóriamente em subconjuntos de 70% (treinamento) e 30% (teste). Os modelos utilizados para a predição de alcalóides foram *Random Forest (RF), Neural Networks (NN) e Partial Least Square* (PLS). Na fase de treinamento a avaliação dos modelos foi realizada por *cross validation*, com 10 dobras. Os parâmetros de análise dos modelos foram *Root Mean Squared Error* para etapa de *cross validation* (RMSEcv) e predição (RMSEp), coeficiente de determinação para *cross validation* (R^2_{CV}) e predição (R^2p) e a razão de previsão para desvio (RPD).

Segundo Sayes et al [6], o R^2 demonstra de fato a porcentagem de variância entre o mensurado e o predito, e pode ser classificado como, discriminação entre altas e baixas concentrações (0,50 a 0,65), passível de uso para previsões quantitativas (0,66 a 0,81) e boas previsões quantitativas (acima de 0,82). Ainda segundo o mesmo autor, um RPD abaixo de 1,5 indica calibração ruim, entre 1,5 e 2,0 revela possibilidade de distinção em níveis altos e baixos, entre 2,0 e 2,5 há possibilidade de previsões quantitativas, por fim 2,5 e acima é classificado como boa previsão.

3. RESULTADOS

3.1. Estatística descritiva

A análise descritiva apresentada na Tabela 1 demonstra os valores mensurados de alcalóides para a primeira leitura, segunda leitura e soma das leituras.

Alcalóides (%)			
Estatística	**41 DAP**	**69 DAP**	**41 e 69 DAP**
Observações	300	300	600
Mínimo	0.36	0.86	0.36
Máximo	1.81	5.18	5.18
Média	0.99	2.31	1.65
CV[1]	19.09	31.52	51.37

[1]CV: coeficiente de variação.

Tabela 1. Estatística descritiva de alcaloides em dias após plantio (DAP).

O valor do coeficiente de variação (CV) foi de 51,37% para as duas datas em conjunto. Pimentel-Gomes[7] sugere a classificação da seguinte forma: CV baixo (<10%), médio (entre 10 e 20%), alto (entre 20 e 30%) e muito alto (acima de 30%). Desta maneira, o valor foi classificado como muito alto. Valores classificados como muito alto sugerem pouca homogeneidade dos dados.

3.2. Análise visual do espectro

A Figura 1 apresenta o comportamento espectral médio das 600 leituras. Observa-se uma maior intensidade de reflectância para a segunda data na região do visível (400-700 nm) até infravermelho próximo (700-1100 nm), já na região do infravermelho de ondas curtas (1100-2500 nm) observou-se uma inverção do comportamento.

Para a Figura 2 tem-se o espectro médio dos 30 genótipos de tabaco. Observa-se uma distinção clara na intensidade da reflectância entre os tratamentos. Essa variação é devida pela

estrutura foliar, pigmentos como clorofila e carotenóides e outras características fisiológicas.

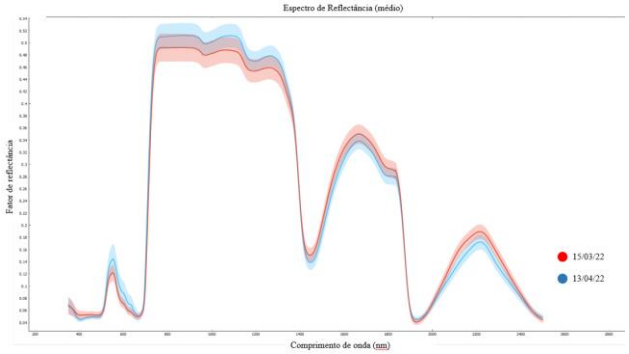


Figura 1. Curva espectral média de tabaco aos 41 e 69 dias após plantio (em vermelho e azul, respectivamente).

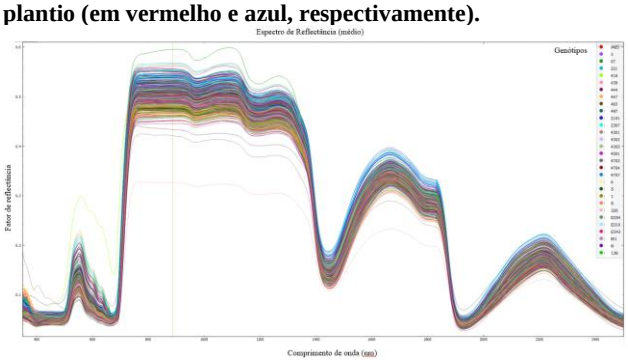


Figura 2. Curva espectral média dos tratamentos.

3.3. Machine learning

A Tabela 2 apresenta as métricas de avaliação para validação cruzada na fase de treinamento, para os três modelos utilizados. Os melhores modelos são aqueles que têm menor RMSEcv e maior R^2 e RPD.

Modelo	RMSEcv	R^2cv	RPD
Neural Network	0,45	0,72	1,90
Random Forest	0,46	0,71	1,85
PLS	0,48	0,68	1,78

Tabela 2. Resultados da fase de treinamento.

O melhor resultado nessa fase foi obtido para o modelo Neural Network, pois apresentou o melhor valor nas três métricas de avaliação. Segundo a classificação citada no ítem 2.3., o modelo NN apresentou R^2 passível de realizar previsões quantitativas, entretanto, o valor obtido para RPD (1.90) o classifica apenas como passível de uso para distinção de baixas e altos valores.

A Tabela 3 demonstra as mesmas métricas de avaliação da fase de predição.

Modelo	RMSEp	R^2	RPD
Neural Network	0,46	0,71	1,80
PLS	0,47	0,68	1,76
Random Forest	0,49	0,64	1,69

Tabela 3. Resultados da fase de predição.

Pela menor quantidade de dados nessa fase, espera-se que os modelos apresentem valores de RMSEp maiores do que os obtidos na fase de treinamento e menores valores de R^2 e RPD. Para o modelo Neural Network, observou-se um pequeno aumento de RMSEp e uma leve diminuição nos valores de R^2 e RPD para a fase de predição.

O modelo NN destacou-se dos demais nessa fase, atingindo um valor de R^2 de 0,71, sendo classificado como apto a realizar predições, contudo, o valor de 1,80 para RPD o enquadra apenas como apto a realizar classificação entre baixos e altos valores. O modelo RF, foi o que apresentou a piora mais acentuada entre as fases de treinamento e de predição.

As figuras 3, 4 e 5 mostram os gráficos de correlação para os 3 modelos avaliados. Conforme mencionado anteriormente, o melhor modelo na fase de predição foi a NN (Figura 3), a dispersão dos pontos apresenta certa homogeneidade de distribuição por toda faixa de dados em comparação aos demais (Figura 4 e 5).

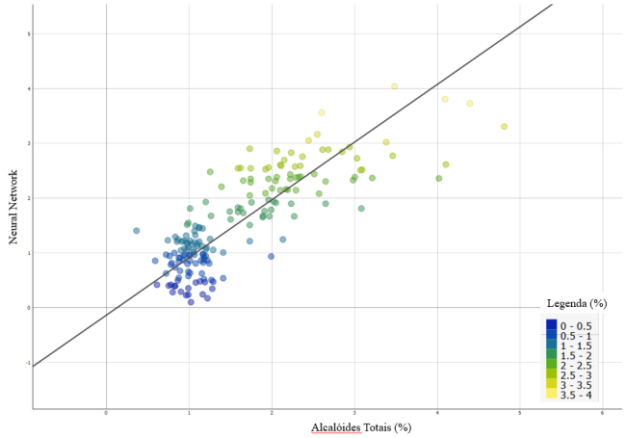


Figura 3. Correlação dos pontos preditos e observados para Neural Network.

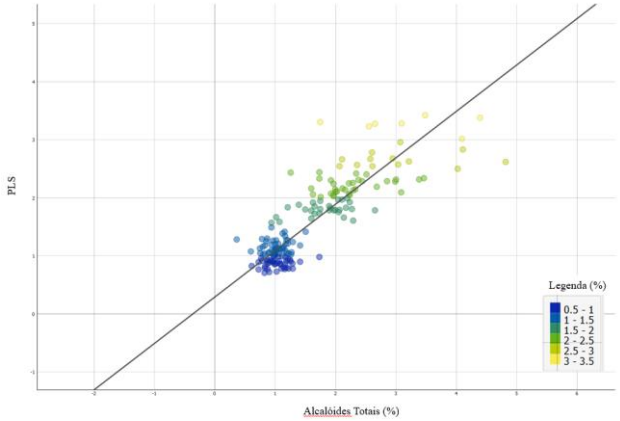


Figura 4. Correlação dos pontos preditos e observados para PLS.

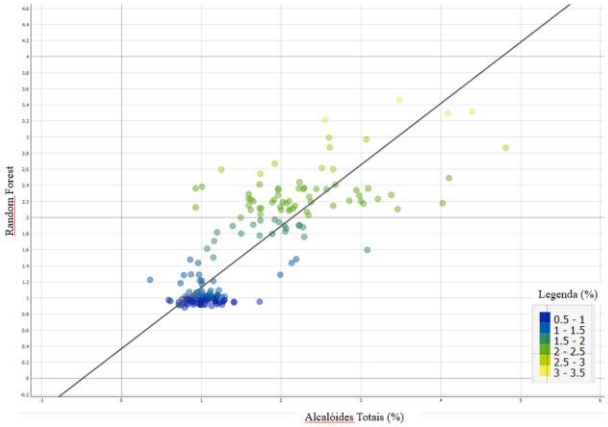


Figura 5. Correlação dos pontos preditos e observados para Random Forest.

4. DISCUSSÃO

4.1. Análise descritiva

A alta variabilidade para o teor de alcalóide demonstrada na Tabela 1 (CV = 51,37%), é resultado da alta diversidade genética do experimento, que contém 30 genótipos distintos. Além disso, as 600 observações são resultado de leituras em duas datas de coleta diferentes, cujo desenvolvimento da cultura promove variações bioquímicas e estruturais nas plantas, o que também pode ser visto nas diferenças entre as curvas espectrais médias para cada data (Figura 1).

4.2. Análise visual do espectro de reflectância

Como demonstrado na Figura 1 e 2, o comportamento médio da reflectância entre as datas de obtenção de dados e entre os genótipos foram distintos. Alguns aspectos fisiológicos relacionados a estrutura foliar, contribuem para a clara diferença entre os genótipos apresentados na Figura 2. Em relação a dimensão (espessura ou tamanho) e a quantidade de alcalóides produzidos, são eles, compartimento estrutural, estrutura celular, parede celular, espessura foliar, taxa fotossintética e acúmulo de celulose e lignina [8].

4.3. Machine learning

Conforme descrito no item 3.3 e apresentados na Tabela 1, os 3 modelos apresentaram comportamento semelhante, com uma boa vantagem para a Neural Network. Os valores de R^2cv para todos os modelos foram classificados como aptos a realizarem predições quantitativas (>0.66). Para a fase de predição também foi observado um leve aumento de RMSEp e diminuição R^2p, explicado possivelmente pela menor quantidade de dados utilizados nessa fase[6].

5. CONCLUSÕES

Os modelos de machine learning utilizados foram capazes de estimar os teores foliares de alcaloides em tabaco, por meio da reflectância espectral foliar, apresentando R^2 de 0,72 (NN), 0,71 (PLS) e 0,68 (RF) na etapa de predição de dados. O modelo Neural Network apresentou maior precisão na estimativa do atributo.

Esses resultados sugerem o potencial do uso de reflectância espectral Vis-NIR-SWIR no monitoramento de alcaloides em tabaco, o que pode auxiliar na determinação de zonas de manejo da cultura com base na variabilidade dos teores foliares e otimizar o processo industrial de classificação da qualidade do tabaco.

6. REFERÊNCIAS

[1] Associação dos Plantadores de Fumo em Folha no Rio Grande do Sul. Dados sobre a Fumicultura no Brasil, 2022. Disponível em: < https://afubra.com.br/fumicultura-brasil.html>.

[2] L. F. S. Santos. Aplicação das técnicas de microextração µMSPD e BAµE para determinação de alcaloides em matrizes in natura e processadas de tabaco (Nicotiana tabacum L.) por cromatografia líquida acoplada a espectrometria de massas. Universidade Federal de Sergipe, Tese de doutorado apresentada ao Programa de Pós-Graduação em Química. São Cristovão – SE, 2019.

[3]A. Gracia-Romero, S. C. Kefauver, O. Vergara-Díaz, M. A. Zaman-Allah, B. M. Prasanna, J. E. Cairns, J. L. Araus. Comparative performance of ground vs. aerially assessed RGB and multispectral indices for early-growth evaluation of maize performance under phosphorus fertilization, Frontiers in plant science, v. 8, pp. 2004, 2017.

[4] Z. Li, C. Li, Y. Gao, W. Ma, Y. Zheng, Y. Niu, Y. Guan, J. Hu. Identification of oil, sugar and crude fiber during tobacco (Nicotiana tabacum L.) seed development based on near infrared spectroscopy. Biomass Bioenergy, volume (111.):páginas(39-45), abril de 2018.

[5] M. Rodrigues, R. B. de Oliveira, G. L. A. A. dos Santos, K. M. de Oliveira, A. S. Reis, R. H. Furlanetto, L. A. Y. B. Junior, F.S. Coelho, M. R. Nanni. Rapid quantification of alkaloids, sugar and yield of tobacco (Nicotiana tabacum L.) varieties by using Vis–NIR–SWIR spectroradiometry. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. Volume (274.):páginas (121082), março de 2022.

[6] W. Sayes, A. M. Mouazen, H. Ramon. Potential for Onsite and Online Analysis of Pig Manure using Visible and Near Infrared Reflectance Spectroscopy. Biosystems Engineering. Volume (91):páginas 393-402, Agosto de 2005.

[7] F. Pimentel-Gomes, Statistics Course Experimental, (2009).

[8] R. Falcioni, T. Moriwaki, W. C. Antunes, M. R. Nanni. Rapid Quantification Method for Yield, Calorimetric Energy and Chlorophyll a Fluorescence Parameters in Nicotiana tabacum L.

Using Vis-NIR-SWIR Hyperspectroscopy. Plants. Volume (18):páginas 2406, Setembro de 2022.