Modelos de aprendizagem de máquina para predição da presença de desoxinivalenol em grãos de trigo por meio de análises multiespectrais

Eduarda Zanini[1], Deividi Felipe Zaions[1], Alexandre Lazaretti Zanatta[1]; Carlos Amaral Hölbig[1]; Willingthon Pavan[2]

[1]Programa de Pós Graduação em Computação Aplicada (PPGCA) – Universidade de Passo Fundo (UPF) - 99052-900 - Passo Fundo - RS – Brasil

[2] International Fertilizer Development Center - 46 David Lilienthal Dr, - Muscle Shoals, AL 35661 - USA

{158223, 185579, zanatta, holbig,}@upf.br, wpavan@ifdc.org

Abstract. *This work aimed at an introductory study applied in agriculture, concerning data modeling to predict the condition of wheat grains through multispectral analysis of the AS7265x sensor. Ten commercial samples and 1 healthy sample of wheat grains were analyzed. Among commercial samples, 8 of them are contaminated with deoxynivalenol (DON). Fifty readings were collected from each typifying, totaling 550 data inputs. The paper does a performance analysis of supervised machine learning algorithms: k- nearest neighbors, support vector machine and random forest. As a result, the random forest algorithm had the highest performance among the 3. From the results, we conclude that this solution is a first step to help a wheat farmer in the decision-making process.*

Keywords: *Data Modeling. Machine Learning. K- Nearest Neighbors. Support Vector Machine. Random Forest.*

Resumo. *Este trabalho apresenta um estudo sobre modelagem de dados para predizer o estado dos grãos de trigo através da análise multiespectral do sensor AS7265x. Foram analisadas 10 amostras comerciais e 1 amostra sadia de grãos de trigo. Dentre amostras comerciais, 8 estão contaminadas com desoxinivalenol (DON). Foram coletadas 50 leituras de cada amostra, totalizando 550 entradas de dados. O trabalho faz uma análise do desempenho dos algoritmos supervisionados de machine learning: k- nearest neighbors, support vector machine e random forest. Como resultado, o algoritmo random forest teve o melhor desempenho entre os 3. Com base nos resultados conclui-*

se que, esta solução é um primeiro passo para ajudar um produtor de trigo no processo de tomada de decisão.

Palavras chave: *Modelagem de Dados. Machine Learning. K- Nearest Neighbors. Support Vector Machine. Random Forest.*

1. Introdução

A tecnologia da informação está presente em muitos setores, como por exemplo, na agricultura para avaliação da qualidade da colheita, previsões de rendimento, detecção de doenças, e demais aspectos importantes que auxiliam na tomada de decisão dos produtores e que tornam esse setor cada vez mais competitivo e eficiente.

O trigo é um dos cereais mais produzidos no mundo, justamente por conta da sua grande adaptação com o solo e o clima. No Brasil, o trigo tornou-se uma opção de grande valor econômico, principalmente para os produtores que utilizam mão de obra familiar, pelo seu bom preço de mercado e à significativa demanda pelo produto por ser utilizado em diversos tipos de preparações. (CONAB, 2017)

Nesse sentido, percebeu-se a necessidade de analisar o estado dos grãos do trigo por meio da análise multiespectral de amostras coletadas, aplicando os conceitos de modelagem de dados e *machine learning*. O local de maior relevância para os cultivares de inverno, entre eles o trigo, é a Região Sul do Brasil, justamente por seu clima temperado, que de certa forma, favorece o desenvolvimento desses cereais, pois fornece adaptabilidade a eles em relação aos seus centros de origem. (CONAB, 2020)

A doença giberela (*Gibberella zeae*), é um tipo de fungo que com frequência é encontrado no cultivar do trigo sendo responsável pela redução do rendimento e a qualidade dos grãos e derivados. Também, pode provocar danos à saúde humana e animal, resultando na produção de micotoxinas. As principais micotoxinas são a desoxinivalenol (DON), a zearalenona (ZEA) e o nivalenol (NIV). (Embrapa, 2016)

A espectroscopia é uma ferramenta muito explorada na área da agricultura ao longo de 50 anos (Embrapa, 2018). Como a resposta espectral das plantas se altera conforme seus índices nutricionais, existem outras ferramentas e técnicas para manejo dentro da computação que também auxiliam na informação e segurança para a tomada de decisão como o uso de inteligência artificial.

Diante do exposto, tem-se o seguinte problema de pesquisa: *"Como auxiliar os agricultores no processo de identificação do estado do cultivar de trigo através de análises multiespectrais de amostras e predizer a partir de algoritmos de machine learning se um determinado grão esta sadio ou contaminado com a micotoxina (DON)"*. Assim, o objetivo deste trabalho é introduzir uma abordagem sobre modelagem de dados com algoritmos de *machine learning* para predição do estado dos grãos de trigo por meio da coleta de amostras por um sensor multiespectral, bem como, avaliar o desempenho dos algoritmos propostos para a solução do problema.

Para alcançar o objetivo pretendido, o trabalho divide-se em cinco capítulos. No primeiro capítulo, que é a introdução, é apresentado o tema da pesquisa, sua justificativa de estudo, objetivo geral e os específicos e o questionamento norteador do estudo. Na sequência é apresentada a fundamentação teórica da pesquisa que aborda sobre, espectroscopia na identificação de características de amostras, como é realizado o processo das técnicas de *machine learning* e como funcionam os algoritmos: *k- nearest*

neighbors, *support vector machine*, e *random forest*. No terceiro capítulo apresenta-se a aplicação, descrevendo os métodos utilizados e a discussão sobre os resultados e por fim, no quarto capítulo apresentam-se as considerações finais e o quinto as referências.

2. Revisão de literatura

2.1. Espectroscopia

A espectroscopia é um estudo que remete a compreensão da geração da radiação eletromagnética (uma forma de energia propagada na forma de ondas eletromagnéticas) e da sua interação com a matéria. Se divide em muitas áreas que se dedicam a estudar faixas relativamente estreitas do espectro eletromagnético, de acordo com suas energias e, com os fenômenos que elas podem produzir ao interagir com a matéria. Dentre as diversas formas de interação da radiação com a matéria, a absorção da radiação pelos constituintes da amostra é de grande interesse para a espectroscopia analítica, pois este fenômeno gera os espectros de absorção que contém as informações analíticas qualitativas e quantitativas a respeito de uma amostra. (Embrapa, 2018)

As ondas eletromagnéticas conseguem ser classificadas com base nos seus diversos comprimentos de onda/frequências, que são denominadas como espectro eletromagnético. (Hollas, 2004)

Uma imagem digital é a representação da figura de um objeto pela combinação da intensidade dos raios de luz provenientes da mesma. Uma imagem espectral é aquela que reproduz, a partir da análise de um objeto, o comprimento de onda do objeto em questão. (Habibi, 2014). Os sensores ópticos de câmeras comuns possuem uma faixa de captação do comprimento de onda eletromagnético dentro espectro visível, a qual varia entre 400 nm a 750 nm. (Zhou, 2019). Os sensores multiespectrais conseguem capturar frequências além da faixa de luz visível, ou seja, capturam dados de imagens com uma faixa de captação do comprimento de onda específicas em todo o espectro eletromagnético, e podem ser separados por filtros ópticos ou detectados por meio do uso de equipamentos que são sensíveis a comprimentos de onda específicos, é possível realizar, por exemplo, um levantamento do número de plantas em determinada área, verificar a saúde das plantas e detectar pragas na plantação. (Giannoni et al., 2018)

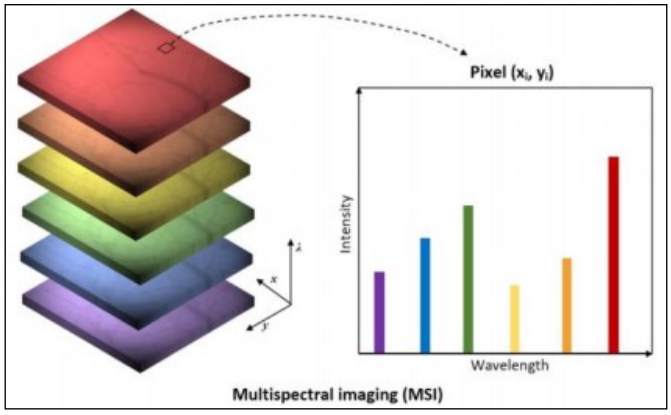

Figura 1. Representação do comprimento de onda de uma imagem multiespectral.

Fonte: Giannoni et al., 2018.

2.2. Machine Learning

O *machine learnig* (ML) tem com o objetivo criar novas possibilidades para resolver, quantificar e compreender processos intensivos de dados em ambientes operacionais de determinadas áreas como a agricultura. O ML envolve um processo de aprendizagem, a partir de dados de treinamento para realizar uma tarefa. Para calcular o desempenho dos modelos e algoritmos de ML, vários modelos estatísticos e matemáticos são usados. Após o final do processo de aprendizagem, o modelo treinado pode ser usado para classificar, prever ou agrupar novos exemplos. As técnicas de ML podem ser divididas em diferentes categorias amplas dependendo do tipo de aprendizado classificam-se em: supervisionado, não supervisionado ou por reforço. (Liakos et al., 2018)

2.2.1. Aprendizado supervisionado

As redes de aprendizado supervisionado aprendem a partir de dados de treinamento pré-classificados, e assim a partir desse classificador é utilizando como exemplo, onde contêm a informação da saída esperada e por fim conseguem classificar os dados de entrada mais precisamente (Coppin, 2015).

2.2.2. Aprendizado não supervisionado

O aprendizado não supervisionado aprende a classificar um conjunto de dados de entrada, sem informação alguma sobre quais são as classificações e sem receber nenhum dado de treinamento, o próprio aprendizado irá identificar padrões e classificá-los (Coppin, 2015).

2.2.3. Aprendizado por reforço

O aprendizado por reforço, posiciona-se entre o aprendizado supervisionado e o aprendizado não supervisionado. Esses métodos costumam ser úteis quando apenas rótulos incompletos estão disponíveis, ou seja, os dados de entrada podem estar apenas parcialmente disponíveis com alguns dados de saídas ausentes (VanderPlas, 2016).

2.2.4. Algoritmo *k-nearest neighbors*

O algoritmo *K-Nearest Neighbor* (KNN), do português "K-Vizinho Mais Próximo", é um método de aprendizado com base em instâncias, que consiste em, armazenar os dados de treinamento e os usar para definir uma classificação para cada dado novo de entrada. (Coppin, 2015) O KNN é um algoritmo supervisionado do tipo classificador não paramétrico, possuí três elementos principais: um conjunto de dados para exemplo, uma métrica de distância e o valor k número de vizinhos. (Oliveira, 2017). Cada instância pode ser formada por um vetor de *n* dimensões, onde *n* é o número de atributos usados para descrever cada instância e as classificações para valores numéricos discretos. Os dados de treinamento são armazenados, e quando uma nova instância é encontrada ela será comparada aos dados de treinamento para encontrar os seus vizinhos mais próximos. E isso é realizado pela computação chamado de distância Euclidiana entre instâncias em um espaço de *n* dimensões (Coppin, 2015).

2.2.5. Algoritmo *Support Vector Machine*

O algoritmo *Support Vector Machine* (SVM), do português "Máquina de Vetor de Suporte" têm como finalidade a especificação de limites de decisão que produzam um ótimo desempenho entre classes por meio da minimização dos erros. O funcionamento do SVM serve para problemas de reconhecimento de padrão, e aplica uma teoria estatística de aprendizagem, encontra uma linha de separação, chamada de hiperplano

entre dados de duas classes. Assim buscando maximizar a distância entre os pontos mais próximos em relação a cada uma das classes. Esse algoritmo possui quatro funções: linear, quadrática, polinomial e função de base radial. (Vapnik, 2009). Neste estudo será utilizada a função linear, justamente com base na simulação da distribuição das amostras coletadas.

2.2.6. Algoritmo *Random Forest*

No algoritmo *random forest* são construídas várias árvores de decisão para classificar um novo dado. A partir disso, será construída a primeira árvore de decisão, sendo definido o primeiro nó da árvore, que será a primeira condição analisada e criará os dois primeiros ramos, para isso, é necessário aplicar a função de entropia ou o índice *Gini*, justamente para escolher a melhor variável para compor o nó raiz. O algoritmo definirá aleatoriamente duas ou mais variáveis, e então realizará os cálculos com base nas amostras selecionadas para definir qual dessas variáveis será utilizada no primeiro nó.

Para escolha da variável do próximo nó, novamente serão escolhidas duas (ou mais) variáveis, excluindo as já selecionadas, e o processo de escolha se repetirá. Desta forma, a árvore será construída até o último nó. Quanto mais árvores criadas, melhores serão os resultados do modelo, até determinado ponto, onde uma nova árvore não conseguirá uma melhora significativa no do modelo. Cada árvore tem o seu resultado. O resultado que mais vezes for gerado será o optado. (Didática Tech, 2020)

3. Metodologia

Para a parte física do dispositivo, foi escolhido o sensor *AS7265x Smart Spectral Sensor* do fabricante AMS para realizar a análise das amostras de trigo, pois possui tecnologia multiespectral que, forma um conjunto de chips do sensor multiespectral de 18 canais *AS7265x*. Sendo visível (VIS) e infravermelho próximo (NIR) de 410nm a 940nm cada com 20nm FWHM. (AMS, 2018). A partir da decisão exposta sobre o uso do sensor, foi realizada a impressão de modelos em impressoras 3D, para realizar a coleta de dados espectrais sobre as amostras de trigo.

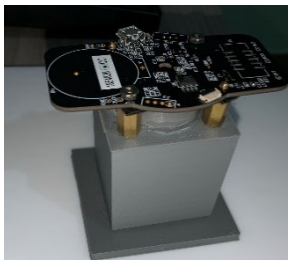

Figura 2. Modelo 3D para coleta das amostras multiespectrais.

Na coleta de dados foram analisadas pelo sensor dez (10) amostras comerciais e uma amostra sadia de grão de trigo. Dentre as amostras comerciais, oito (8) estão contaminadas com níveis diferentes de (DON). Foram coletadas 50 leituras pelo sensor espectral sobre todas as amostras, totalizando 550 entradas de dados. Destes, aplicou-se uma razão 70/30 para treinamento e validação. Na Tabela 1 podem ser observados os índices de (DON) identificados nas amostras.

Tabela 1. Índices de (DON) identificados nas amostras.

	21_195	21_196	21_197	21_198	21_199	21_200	21_201	21_202	21_210	21_227	Sadio

DON (ug/kg)	1788	483,6	2113,8	1508,1	2009,1	1943,4	0	0	799,2	307,5	0

Para a aplicação dos algoritmos utilizou-se das bibliotecas *scikit-learn*, *pandas* e *numpy* na linguagem de programação Python. A partir da modelagem dos dados, para a aplicação dos algoritmos de ML é necessário normalizar estes dados para que os modelos não fiquem com maiores ordens de grandeza, ou seja, normalizar dados tem como objetivo colocar as variáveis dentro do intervalo de 0 e 1. A função que se aplica está operação é dada pela Equação 1:

$$Z = x\text{-}min(x) / (max(x)\text{-}min(x)) \quad (1)$$

A aplicação da função de normalização apresentou os dados dos espectros L 940nm e A 410nm com valores zerados, sendo que, estes dados foram removidos dos testes, pois estavam influenciando no treinamento da rede, o que pode ser caracterizado por *outliers*. Após, foram aplicados os testes de classificação para os três algoritmos: KNN, SVM e *random forest.* Na Tabela 2 pode ser observado um comparativo do desempenho da classificação por decisão binária dos grãos de trigo. Todas as amostras cujos níveis de (DON) foram verificados tem uma classificação de contaminados (1) e todas as demais se classificam como sadio (0).

Tabela 2. Teste de decisão binária (sadio ou contaminado).

	KNN	**SVM**	**Random Forest**
Accuracy	0.84	0.89	0.92
Recall (sadio)	0.59	0.70	0.85
Recall (contaminado)	0.94	0.97	0.94
Precision (sadio)	0.79	0.89	0.85
Precision (contaminado)	0.85	0.89	0.94

Dos testes realizados verificou-se que o algoritmo *random forest* teve um maior desempenho de acurácia, ou seja, a proporção de previsões que o modelo classificou corretamente com ele. Esse algoritmo também conseguiu uma maior precisão de contaminados, a proporção de identificações positivas estava realmente correta para contaminados. No entanto, o SVM teve uma maior precisão para dados sadios. A maior sensibilidade (*Recall*) de sadios também foi o *random forest*. Já a sensibilidade para dados contaminados tanto o algoritmo KNN quando *random forest* resultaram no mesmo valor, e o SVM teve o melhor resultado nesta métrica. Além desses dados a partir da matriz confusão de cada um o *random forest* teve menor índice de dados falsos negativos, ou seja, o resultado que o modelo previu como incorreto é na verdade positivo. Por fim o melhor modelo para esse tipo de classificação dentre os três foi o *random forest*.

Para os testes de classificação a partir dos níveis de DON como classificador, os seguintes resultados estão disponíveis na Tabela 3.

Tabela 3. Teste de classificação.

	KNN	**SVM**	**Random Forest**
Accuracy	0.70	0.68	0.82
Recall (sadio)	0.72	0.63	0.87
Precision (sadio)	0.85	0.89	0.82

Embora o KNN tenha sido efetivo em identificar corretamente as amostras sadias, o *random forest* também teve o melhor desempenho para classificar as amostras de acordo com seu nível de contaminação.

4. Conclusões e trabalhos futuros

Este trabalho mostrou como modelar dados coletados por meio de análises multiespectrais dos grãos de trigo e aplicou os algoritmos de *machine learning*: KNN, SVM *e random forest.* A partir dos dados previu se os níveis desses grãos estão sadios ou contaminados. Conclui-se, que o *random forest* foi o algoritmo que teve a melhor *performance* em comparação com os demais.

Por fim, nota-se que as taxas de acerto e erros relativos obtidos desses algoritmos em geral ainda são razoáveis e que a pesquisa carece de evolução, por exemplo, utilizou-se um "número muito pequeno" de grãos de trigo dissipados ou não com uma micotoxina desoxinivalenol. Por outro lado, essa abordagem atingiu o objetivo do trabalho, pois é um primeiro passo para uma solução que ajude os agricultores no processo de tomada de decisão.

Como trabalho futuro, buscar outras classificações, não se limitando a apenas em duas classes (não contaminado e contaminado). Também, buscar-se-á analisar mais amostras deste sensor e validar os mesmos algoritmos propostos para qualificar a performance dos resultados obtidos, bem como, usar o método *random forest* para outros tipos de classificadores e outros tipos de cultivares.

5. Referências

AMS. (2018). AS7265x Smart 18-Channel VIS to NIR Spectral_ ID 3-Sensor Chipset with Electronic Shutter. 63 p. Disponível: https://ams.com/documents/20143/36005/AS7265x_DS000612_1-00.pdf/08051c8a-a7f6-6231-7993-2d3fe0bf38b8. Acesso: junho/2021.

Companhia Nacional de Abastecimento. (2017). A cultura do trigo. CONAB. 218 p. Disponível: https://www.conab.gov.br/uploads/arquivos/17_04_25_11_40_00_a_cultura_do_trigo_versao_digital_final.pdf. Acesso: junho 2021.

Companhia Nacional de Abastecimento. (2020). Acompanhamento da safra brasileira de grãos, v. 7 - Safra 2019/20 - Décimo segundo levantamento. CONAB. 45 p. Disponível: https://www.conab.gov.br/info-agro/analises-do-mercado-agropecuario-e-extrativista/analises-do-mercado/historico-mensal-de-trigo. Acesso: junho, 2021.

Coppin, B. (2015). Inteligência Artificial / Ben Coppin; tradução e revista técnica Jorge Duarte Pires Valério. [Reimpr.]. Rio de Janeiro: LTC. Tradução de: Artificial intelligence illuminated, 1st ed.

Didática Tech. (2020). Inteligência Artificial & Data Science. O que é e como funciona o algoritmo RandomForest. Disponível: https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-randomforest/. Acesso: junho/2021.

Empresa Brasileira de Pesquisa Agropecuária. (2016). Trigo: o produtor pergunta, a Embrapa responde / Claudia De Mori et al., editores técnicos. Brasília, DF: Embrapa. 309 p. Disponível: https://www.embrapa.br/busca-de-publicacoes/-

/publicacao/1040211/trigo-o-produtor-pergunta-a-embrapa-responde. Acesso: junho, 2021.

Empresa Brasileira de Pesquisa Agropecuária. (2018). Espectroscopia no Infravermelho próximo para avaliar indicadores de qualidade tecnológica e contaminantes em grãos / Casiane Salete Tibola et al., editores técnicos. Brasília, DF: Embrapa. 200 p.

Giannoni, L. et al. (2018). Hyperspectral imaging solutions for brain tissue metabolic and hemodynamic monitoring: past, current and future developments. Computer Science, Medicine. Journal of Optics, v. 20, n. 4. p. 25. Disponível: https://iopscience.iop.org/article/10.1088/2040-8986/aab3a6/meta. Acesso: julho/2021.

Habibi, M. (2014) Image sensors. In: Measurement, Instrumentation, and Sensors Handbook. 2. ed. [S.l.]: CRC Press. cap. 4. p. 1921.

Hollas, J. M. (2004). Modern Spectroscopy. Hoboken: John Wiley & Sons, Ltd. 482 p.

Liakos, K.G. et al. (2018). Machine Learning in Agriculture: A Review. *Sensors*. *18*, 2674. Disponível: https://doi.org/10.3390/s18082674. Acesso: junho/2021.

Oliveira, A.R. and Roesler. (2017). Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes. ELSA-Brasil: accuracy study. Sao Paulo Medical Journal, v.135, n. 3, p. 234-46.

VanderPlas, J. (2016). Python Data Science Handbook. Disponível: https://jakevdp.github.io/PythonDataScienceHandbook/. Acesso: junho/2020.

Vapnik, V. (2009). The Nature of Statistical Learning Theory. 2. nd. New York: Springer. 745 p.

Zhou, X. et al. (2019). A novel combined spectral index for estimating the ratio of carotenoid to chlorophyll content to monitor crop physiological and phenological status. International Journal of Applied Earth Observation and Geoinformation, v. 76, p. 128–142. Disponível: https://www.sciencedirect.com/science/article/abs/pii/S030324341830566X. Acesso: junho/2020.