Modelagem de dados introdutória para predição do estado dos grãos de trigo através de análises multiespectrais

Eduarda Zanini

Curso de Ciência da Computação, UPF
Campus 1 - BR 285 - Passo Fundo (RS) - Brasil

158223@upf.br

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 teve como objetivo um estudo introdutório aplicado na agricultura, 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 a 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.

As projeções mostram que a população mundial irá crescer para aproximadamente 9,1 bilhões de habitantes até 2050. E que, a produção de alimentos mundial deve aumentar cerca de 70 % para atender a essa crescente demanda. (FAO, 2009) A presença da tecnologia da informação para a produção agrícola é a principal base para atender esta demanda, pois resultam no aprimoramento das técnicas de cultivo e eficiência de aplicação dos recursos. Possibilita a detecção e medição da quantidade de nutrientes e fertilizantes que precisam ser adicionados ao solo. Para isso, a agricultura de precisão é utilizada, baseada na análise, monitoramento e administração de todas as matérias-primas necessárias que atendem demandas locais nas lavouras.

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. A farinha de trigo, por exemplo, é o principal ingrediente dos diversos usos culinários desse cereal. (CONAB, 2017)

Nesse sentido percebeu-se a necessidade de analisar o estado dos grãos do trigo através da análise multiespectral de amostras coletadas, aplicando os conceitos de modelagem de dados e *machine learning*.

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 uma breve história do trigo, sua composição e cultivo, como a espectroscopia está relacionada a este meio, sobre o sensoriamento multiespectral, como a modelagem de dados é formulada, como é realizado o processo das técnicas de *machine learning* e como funciona 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.

1.1. Identificação e justificativa do assunto

Conforme o acompanhamento constante da safra de grãos, e a monitoração das condições de desenvolvimento do trigo no Brasil. O local de maior relevância para os cultivares de inverno, é a Região Sul. Justamente por seu clima temperado que favorece o desenvolvimento desses cereais, pois fornece adaptabilidade a eles em relação aos seus centros de origem. (CONAB, 2021)

A doença giberela (*Gibberella zeae*), é um tipo de fungo que com frequência é encontrado no cultivar do trigo. Essa doença é um dos fatores prejudiciais economicamente, justamente por reduzir o rendimento e a qualidade dos grãos e derivados. Como também, pode provocar danos à saúde humana e animal, resultando a produção de micotoxinas. As principais micotoxinas são deoxinivalenol (DON), zearalenona (ZEA) e nivalenol (NIV). Na decorrência desses tipos de fungos que os usos de ferramentas adequadas são essenciais para garantir um bom cultivo. (Embrapa, 2016)

Uma das ferramentas utilizadas na agricultura que é considerada confiável é a análise foliar porque avalia o estado da nutrição de uma planta e o balanceamento de nutrientes conforme as exigências da cultura, a toxicidade dos nutrientes nas plantas, e se

o manejo adotado numa lavoura está coerente (Malavolta et al.,1997). Outra ferramenta muito explorada na área da agricultura ao longo de 50 anos é a espectroscopia (Embrapa, 2018). Tendo em vista, que a resposta espectral das plantas se altera conforme seus índices nutricionais. Além dessas ferramentas, também existem outras ferramentas e técnicas para manejo dentro da computação que também ajudam na informação e segurança para a tomada de decisão como o uso de inteligência artificial e *Internet of Things (IoT)*.

Diante do exposto definiu-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).

1.2. Objetivos

Com o intuito de desenvolver a proposta desse estudo foram definidos os seguintes objetivos geral e específicos:

1.2.1. Objetivo geral

Este projeto tem como objetivo geral 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 e avaliar o desempenho dos algoritmos propostos para a solução do problema.

1.2.2. Objetivos específicos

a) Apresentar uma introdução sobre o trigo e suas características dentro da visão do sensor multiespectral na análise de amostras.

b) Introduzir os conceitos de modelagem de dados para a aplicação proposta, e o conceito de *machine learning* junto aos algoritmos: *k- nearest neighbors, support vector machine e random forest*.

c) Comparar o desempenho dos resultados da aplicação de cada algoritmo citado em relação ao problema proposto.

2. Revisão de literatura

O trigo é um alimento fundamental para a nutrição humana. Pertencente ao gênero *Triticum* da família das *gramíneas*, e as suas principais espécies de cultivo são *Triticum monococcum*, *Triticum durum* e *Triticum aestivum* (Embrapa, 2016). Para a economia global este cereal apresenta grande relevância, sendo um dos mais cultivados no mundo, assim como o arroz e o milho. (FAO, 2021)

Este se originou no sudoeste da Ásia em uma região chamada por historiadores de *Crescente Fértil*. A história do trigo tem uma forte relação com a evolução da civilização humana. O cultivo desse cereal iniciou-se há, aproximadamente, 10 mil anos a.C., cooperando no desenvolvimento dos primeiros povoados. (Embrapa, 2016)

A criação do pão foi descoberta a partir do procedimento de fermentação do trigo pelos egípcios por volta de 4000 anos a.C. O grão espalhou-se pelo mundo há cerca de 2.000 anos a.C. O cultivo do trigo expandiu- se nos locais mais frios, como

Rússia e Polônia, e no século XV por meio dos europeus, o trigo chegou na América. (Flandrin e Montanari, 1998)

No Brasil as sementes de trigo chegaram em 1534, e as primeiras lavouras começaram a cultivar o trigo na região de São Vicente. Entretanto, a sua importância econômica no Brasil colonial deu-se em meados do século XVII, quando cultivadas na região do Rio Grande do Sul e em São Paulo. (Rossi e Neves, 2004)

2.1. Composição do trigo

A planta de trigo é estruturada em: raízes, colmo, folhas e inflorescência. Três grupos de raízes formam o sistema radicular do trigo, sendo: raízes seminais, raízes permanentes (coroa) e raízes adventícias. Geralmente, os estádios de desenvolvimento mais conhecidos do trigo são: plântula, afilhamento, alongamento, emborrachamento, espigamento, florescimento, grão em estado leitoso, grão em massa, grão em maturação fisiológica e grão maduro. (Borém e Scheeren, 2015)

Os principais constituintes do grão são: água, proteínas, lipídios e carboidratos. O teor de umidade dos grãos e a composição aminoacídica das proteínas são princípios relevantes para a industrialização do trigo. (Mandarino, 1994)

Segundo apresentado por Arnon e Stout (1939), uma planta só pode completar seu ciclo vital, se lhe fornecer em quantidade suficiente de todos os elementos minerais que lhe são essenciais.

Para isso, deve satisfazer alguns critérios destes elementos:

- Um elemento é considerado essencial para a planta se sua falta impedir que a planta complete seu ciclo de vida;
- Para que um elemento seja essencial, não pode ser substituída ou compensada por outro elemento, somente com o seu fornecimento;
- O elemento tem que estar envolvido ativamente, diretamente na nutrição da planta, sendo que sua ação não pode processar-se de correção eventual de condições químicas ou microbiológicas desfavoráveis do solo ou do meio de cultura, ou seja, por ação indireta.

O fertilizante mais utilizado no cultivo do trigo como fonte de nitrogênio (N) é a ureia, que é usada principalmente na adubação de cobertura, aplicando em duas fases do ciclo de crescimento e desenvolvimento do cereal: no início do perfilhamento e na fase de alongamento do colmo das plantas. Após o nitrogênio (N), o potássio (K) é o elemento com mais alta concentração no tecido vegetativo nos grãos de trigo. Esse atua no controle das concentrações de sais nos tecidos ou células e na resistência à seca da planta de trigo. (Embrapa, 2016)

2.2. Controle de doenças

Dentre as doenças que ocorrem na plantação de trigo, no grupo das chamadas doenças de difícil controle, apesar da disponibilidade de cultivos com melhores resistências genéticas que outras e de produtos para o tratamento químico da parte aérea das plantas, está a giberela. Essa doença ocorre na espiga do trigo, e com maior frequência nas regiões temperadas e subtropicais. Geralmente essa doença é encontrada quando o estádio de espigamento coincide com períodos chuvosos, reduzindo a eficiência do controle químico, além de inviabilizar ações operacionais relacionadas à aplicação de fungicidas na parte aérea das plantas, em razão da impossibilidade da

entrada de máquinas nas lavouras. Essa doença também pode provocar danos à saúde humana e animal, pois resulta na produção de micotoxinas. As principais micotoxinas são deoxinivalenol (DON), zearalenona (ZEA) e nivalenol (NIV). (Embrapa, 2016). A partir destas informações conseguimos verificar que o manejo, tratamento e prevenção para doenças são fundamentais para obter um cultivo de trigo saudável.

2.3. Espectroscopia

A espectroscopia é um estudo que remete a compreensão da geração da radiação eletromagnética e da sua interação com a matéria. Esta 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. (Embrapa, 2018)

A espectroscopia refere-se sobre a medição e interpretação de espectros que surgem da interação da radiação eletromagnética (uma forma de energia propagada na forma de ondas eletromagnéticas) com a matéria. Trata-se sobre a absorção, emissão ou espalhamento de radiação eletromagnética por átomos ou moléculas. 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)

Fótons são as partículas que constituem a luz e podem ser chamados como pacotes mínimos, que carregam a energia contida nas radiações eletromagnéticas, que é calculado pela Equação 1, onde h é constante de *Plank (6,636 x 10 $^{-34}$ J.s)* e v é a frequência da onda eletromagnética. (Embrapa, 2018)

$$E = hv \quad (1)$$

A velocidade da luz está relacionada com o comprimento de onda conforme exposto na Equação 2, onde o comprimento de onda é representado por lambda (λ), e c é a velocidade da luz no vácuo *(299.792.458 m/s)*. (Hollas, 2004)

$$v = c / \lambda \quad (2)$$

Com isso, as ondas eletromagnéticas conseguem ser classificadas com base nos seus diversos comprimentos de onda/frequências, que são denominadas como espectro eletromagnético. Esse, geralmente apresenta-se em ordem crescente de frequências, classificando-se conforme a Figura 1. (Hollas, 2004)

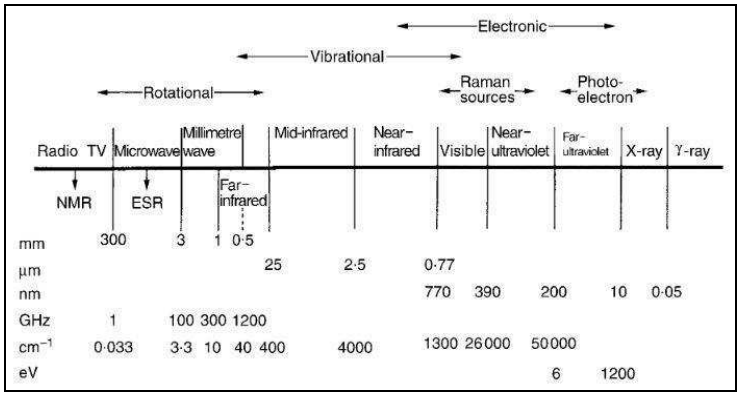

Figura 1. Regiões do espectro eletromagnético.

Fonte: Hollas, 2004

A espectroscopia no infravermelho próximo (NIR) tem se mostrado altamente eficiente na substituição de métodos de análise laboratoriais, justamente porque este método tem mostrado ser mais rápido e sensível, tanto para gases, quanto líquidos e sólidos, como também o fato de não necessitar de preparação da amostra, poupando tempo e reagentes, por ser uma técnica não destrutiva, que permite a seleção e classificação das sementes de acordo com características e atributos específicos, sem alteração de suas propriedades. (Pasquini, 2003) Esse método é um dos mais utilizados no sensoriamento remoto, pois traz informações precisas quanto ao estado fisiológico e de saúde das plantações como uma ferramenta analítica com medições quantitativas e qualitativas. Além de também, não necessitar de um operador qualificado para poder manusear o equipamento, o que reduz custos. (Agelet and Hurburgh, 2014)

Uma imagem digital é a representação da figura de um objeto pela combinação da intensidade dos raios de luz provenientes da mesma. Assim, 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. Com os sensores hiperespectrais, é possível obter os mesmos resultados que um multiespectral e prover detalhes sobre as propriedades físico-químicas dos materiais presentes na superfície imageada, estes sensores são mais adequados para aplicações que são sensíveis a diferenças sutis no sinal ao longo de um espectro contínuo, porém pequenos sinais podem ser perdidos por um sistema que está amostrando bandas de ondas maiores. (Giannoni et al., 2018) Portanto, o uso destes sensores depende muito do tipo de aplicação em que estão sendo usados, os sensores multiespectrais são relevantes para o trabalho, pois adequa-se ao tipo de resultado que estamos buscando onde, é possível predizer o estado dos grãos de trigo.

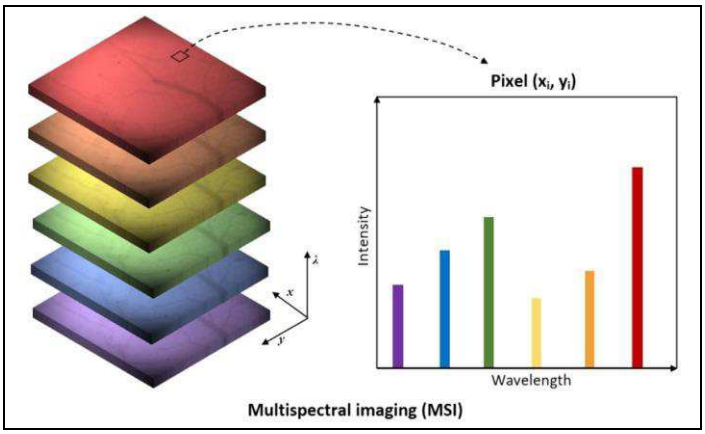

Figura 2. Representação do comprimento de onda de uma imagem multiespectral.

Fonte: Giannoni et al., 2018.

2.4. Modelagem de dados

Após a coleta das informações de um conjunto de dados específicos para análise, como o do trabalho exposto, a modelagem de dados é um dos passos mais importantes para conseguir atingir os objetivos esperados. Esta pode ser classificada em três fases sendo, coleta de dados, tratamento de dados e inferência.

A primeira fase na coleta de dados é o processo de amostragem. A amostra é um conjunto de valores obtidos de uma população de interesse aplicados para representar a população no estudo estatístico. A finalidade desta fase é garantir que a amostra conseguida, faça-se a mais representativa possível. A segunda fase é a de tratamento de dados, onde são empregadas técnicas para representar os dados expostos, reconhecer as falhas nos valores amostrados, e aprimorar o melhor conhecimento do conjunto em estudo. A terceira fase utiliza o conhecimento do cálculo de probabilidades para inferir qual o comportamento da população a partir da amostra, tendo como resposta um modelo probabilístico que representa o conjunto aleatório em estudo, este será incorporado ao modelo de simulação. (Chwif, 2015)

2.5. Machine Learning

A inteligência artificial é o estudo dos sistemas que lidam de uma forma que a um observador entende ser inteligente. Abrange métodos baseados no comportamento inteligente dos seres vivos a fim de resolver problemas complexos. (Coppin, 2015). O *machine learnig* (ML) surgiu junto com tecnologias de *big data* e a computação de alto desempenho com o objetivo de criar novas possibilidades para resolver, quantificar e compreender processos intensivos de dados em ambientes operacionais de determinadas áreas como a agricultura por exemplo. (Liakos et al., 2018)

O ML envolve um processo de aprendizagem com o objetivo de aprender com a experiência, ou seja, a partir de dados de treinamento para realizar uma tarefa. Os dados em ML consistem em um conjunto de exemplos. 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, ou seja, dados de teste usando a experiência obtida durante o processo de treinamento. (Liakos et al., 2018) A partir dessa informação pode-se determinar uma relação da entre os dados e as classificações, na Equação 3, onde a função f é gerada se um grupo de dados, x pertencer à classificação y. (Coppin, 2015)

Então temos:

$$f(x) = y \tag{3}$$

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)

2.5.1. 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. E podendo em algumas situações, generalizar

com grande grau de precisão, a partir de um conjunto de dados de treinamento, chegando ao conjunto completo de entradas possíveis. (Coppin, 2015)

Na Figura 3, conseguimos ilustrar esse processo de classificação:

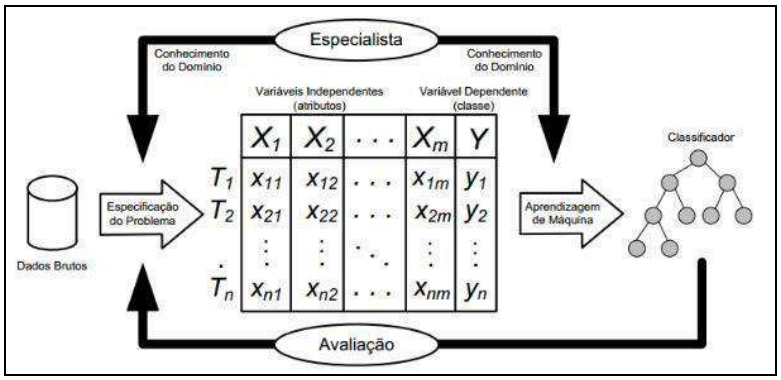

Figura 3. Processo de classificação.

Fonte: Rezende, 2005

Na Figura 3, a partir de um conjunto de dados, *n* cada dado *x* possui *m* atributos, ou seja, $x_i = (x_i\ 1, \ldots, x_i\ m)$. Já as variáveis *y* representam as classes, e partindo dos exemplos e suas devidas classes, o algoritmo extrai um classificador. (Rezende, 2005)

2.5.2. Aprendizado não supervisionado

Estes métodos aprendem sem qualquer intervenção humana. 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 classifica-los. Este método é útil quando, os dados precisam ser classificados, porém as classificações não são conhecidas previamente. (Coppin, 2015)

2.5.3. Aprendizado por reforço

Este tipo de aprendizado esta 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 saída ausentes. (VanderPlas, 2016) Não existe um meio adequado de executar uma tarefa, mas existem regras que o modelo deve seguir para desempenhar corretamente suas tarefas. (Liakos et al., 2018)

2.5.4. 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 este, possuí três elementos principais: um conjunto de dados para exemplo, uma métrica de distância e o valor k número de vizinhos. (Oliveira, 2016)

O seu aprendizado funciona da seguinte forma, cada instância pode ser formada por um vetor de *n* dimensões, onde *n* é o numero 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)

No espaço bidimensional, por exemplo, a distância é calculada entre <*x1, y1*> e <*x2, y2*>, que é dado pela Equação 4:

$$\sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2)} \quad (4)$$

O algoritmo busca a classificação do *k* vizinhos mais próximos da instância a ser classificada e atribuída a ela a classificação mais comumente retornada por aqueles vizinhos. Essa classificação Euclidiana em alguns casos pode ser erronia quando, por exemplo, são usados 7 atributos para definir cada instância, apenas 2 desempenham um papel bom de classificação dessas instâncias. Nestes casos ou as instâncias estão bem afastadas de um espaço das 7 dimensões, ou estão apresentando a mesma classificação. (Coppin, 2015)

Outra abordagem desse algoritmo é o método Shepard que pondera cada um dos vizinhos, de acordo com a sua distância em comparação com a instância que será classificada. Permitindo que cada instância de dado para treinamento contribua para a classificação de uma nova instância. (Coppin, 2015)

2.5.5. 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.

2.5.6. Algoritmo *Random Forest*

Caracteriza-se uma árvore de decisão como, uma função que determina como entrada um vetor de valores de atributos e retorna um valor de saída único booleano, ou seja, dois valores possíveis, em que cada exemplo é classificado como verdadeiro ou falso. E a partir dessa informação, alcança sua decisão executando vários testes. Cada nó interno na árvore corresponde a um teste do valor de um dos atributos de entrada, assim como as ramificações dos nós que são classificadas com os valores possíveis do atributo. Cada nó de folha tem um valor a ser retornado. Cada variável tem um pequeno conjunto de valores possíveis. Os exemplos a partir da raiz são processados seguindo a ramificação apropriada até alcançar uma folha. Um exemplo de árvore de decisão booleana consiste em um par *(x, y)*, onde *x* é um vetor de valores para os atributos de entrada e *y* é um valor único de saída booleano. (Russell, 2013)

O algoritmo de aprendizagem de árvore de decisão adota o conceito de dividir para conquistar, ou seja, sempre testar o atributo que faz mais diferença para a classificação de um exemplo em primeiro lugar. Esse teste divide o problema em subproblemas menores para poder ser solucionado o problema recursivamente. Assim,

conseguimos obter a classificação correta, com um pequeno número de testes, todos os caminhos na árvore serão curtos e a árvore como um todo será pouco profunda. (Russell, 2013)

No algoritmo *random forest* é construída várias árvores de decisão para classificar um novo dado. A partir disso será construída a primeira árvore de decisão. Durante essa construção é preciso definir o primeiro nó da árvore, que será a primeira condição analisada e criará os dois primeiros ramos. Para realizar 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. (Didática Tech, 2020)

Para escolha da variável do próximo nó, novamente serão escolhidas duas (ou mais) variáveis, excluindo as já selecionadas anteriormente, e o processo de escolha se repetirá. Desta forma a árvore será construída até o último nó. Na geração da próxima árvore, os dois processos anteriores se repetirão, levando a criação de uma nova árvore. Quanto mais árvores criadas, melhores serão os resultados do modelo, até determinado ponto, onde uma nova árvore não conseguirá levar a uma melhora significativa no desempenho do modelo. Aplicando esse algoritmo com *machine learning*, podemos formar novos dados e obter previsões. Cada árvore gerada tem o seu resultado, em problemas de classificação o resultado que mais vezes for gerado será o optado. (Didática Tech, 2020)

3. Metodologia

Tendo em vista que o objetivo geral do trabalho é introduzir uma abordagem sobre modelagem de dados com algoritmos de *machine learning* para predição do estado dos grãos de trigo através da coleta de amostras por um sensor multiespectral e avaliar o desempenho dos algoritmos propostos para a solução do problema, busca-se então, alternativas de equipamentos que atendam essas especificações.

Os passos metodológicos do trabalho são apresentados na Figura 4.

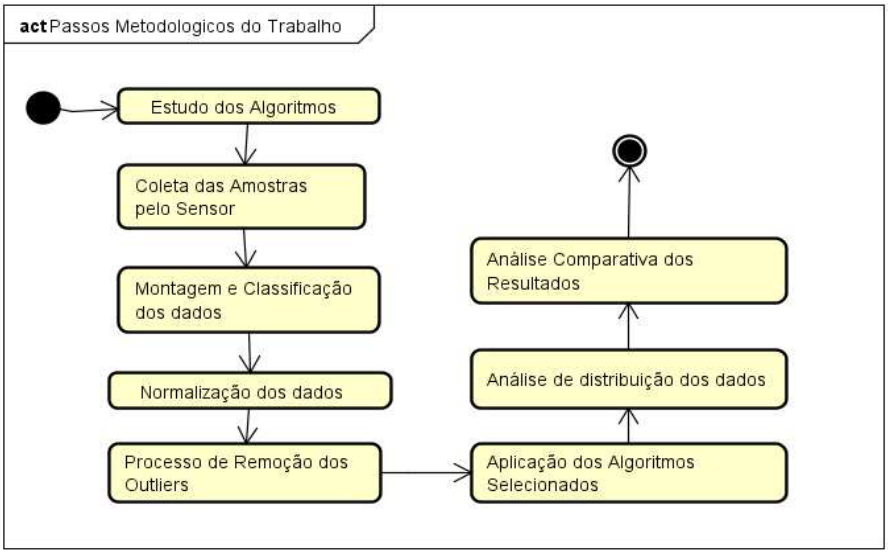

Figura 4. Passos metodológicos do trabalho.

Para tanto, na 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 com um conjunto de 3 chips: *AS72651* com capacidade mestre, *AS72652* e *AS72653*. O *AS72651*, combinado com o *AS72652*, que possuí a resposta espectral de 560nm a 940nm e o *AS72653* que possuí a resposta espectral de 410nm a 535nm. Que por fim, 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. Os componentes *AS72651*, *AS72652* e *AS72653* são pré-calibrados com uma fonte de luz específica. Se qualquer operação for diferente das condições especificadas na documentação, o sensor pode exigir uma nova calibração. (AMS, 2018) Esse processo foi realizado para poder obter a coleta das amostras de forma correta. A partir da decisão exposta sobre o uso do sensor, foi realizada a impressão de modelos em impressoras 3D, para poder realizar a coleta de dados espectrais sobre as amostras de trigo. Esse modelo impresso compõe de três peças 3D (base, tampa e bocal). Na Figura 5, pode-se visualizar o modelo pronto.

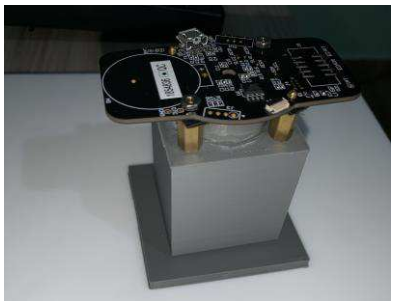

Figura 5. Modelo 3D para coleta das amostras multiespectrais.

Com os materiais selecionados, deve-se então montar o equipamento para realizar a calibração e testes preliminares a fim da validação do dispositivo. Esta fase é muito importante, pois determinará como os 18 canais de comprimento de onda comportam-se. Para realização dessa fase, a própria fornecedora distribui o aplicativo para computador chamado *Ams Spectral Sensor Dashboard*, junto com o kit de avaliação, ao realizar este processo o espectro fica distribuído para análise das amostras, sendo os respectivos dados: R 610nm, S 680nm, T 730nm, U 760nm, V 810nm, W 860nm, G 560nm, H 585nm, I 645nm, J 705nm, K 900nm, L 940nm, A 410nm, B 435nm, C 460nm, D 485nm, E 510nm e F 535nm. (AMS, 2018)

Na coleta de dados foram analisadas pelo sensor 10 amostras comerciais e uma amostra sadia de grão de trigo. Dentre as amostras comerciais, 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, apresentam-se os índices de (DON) identificados nas respectivas 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

A partir da modelagem dos dados, para a aplicação dos algoritmos de *machine lerning* respectivos faz-se necessário normalizar estes dados para que os algoritmos não fiquem com maiores ordens de grandeza, ou seja, normalizar dados tem como objetivo

colocar as variáveis dentro do intervalo de 0 e 1, ou caso tenha resultado negativo -1 e 1. A função que se aplica está operação é dada pela Equação 5:

$$Z = x\text{-}min(x) / (max(x)\text{-}min(x)) \quad (5)$$

Para a aplicação dos algoritmos utilizou-se das bibliotecas *scikit-learn* que é própria para trabalhar com *machine learning*, *pandas* e *numpy* na linguagem de programação Python. A aplicação da função de normalização ao realizada apresentou os dados dos espectros L 940nm e A 410nm com valores zerados, para tanto estes dados foram removidos dos testes, pois, estavam influenciando no treinamento da rede, o que chamamos de *outliers*. Após isso, foram aplicados os testes de classificação para os três algoritmos: KNN, SVM e *random forest.* Na Tabela 2 é possível verificar um comparativo do desempenho da classificação por decisão binária dos grãos de trigo. Todas as amostras cujos níveis de (DON) foram aplicados 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. E 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, ou seja, a proporção da quantidade total de instâncias relevantes contaminados que foram realmente recuperadas. 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*, por mais que os demais tenham classificado de acordo com as suas características.

Para os testes de classificação a partir dos níveis de DON como classificador, os seguintes resultados foram verificados 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

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

Com o algoritmo KNN foram obtidos os seguintes resultados com o valor k=7 para classificação por níveis de DON na Figura 6:

```
-----------------
-----------------
Classificação por níveis de DON
Acurácia KNN:  0.703030303030303
-----------------
-----------------
Dados sobre o teste realizado:
              precision    recall  f1-score   support

           0       0.85      0.72      0.78        54
         307       0.95      0.86      0.90        21
         483       0.92      0.58      0.71        19
         799       0.80      0.80      0.80        15
        1508       0.38      0.67      0.48         9
        1788       0.58      0.54      0.56        13
        1943       0.69      0.75      0.72        12
        2009       0.47      0.70      0.56        10
        2113       0.41      0.58      0.48        12

    accuracy                           0.70       165
   macro avg       0.67      0.69      0.67       165
weighted avg       0.75      0.70      0.72       165

-----------------
-----------------
Matriz Confusão - KNN por níveis de DON
       0  307  483  799  1508  1788  1943  2009  2113
0     39    0    0    0     5     1     4     1     4
307    0   18    0    3     0     0     0     0     0
483    3    0   11    0     3     0     0     2     0
799    1    1    0   12     1     0     0     0     0
1508   0    0    1    0     6     2     0     0     0
1788   0    0    0    0     0     7     0     3     3
1943   2    0    0    0     0     0     9     0     1
2009   0    0    0    0     0     1     0     7     2
2113   1    0    0    0     1     1     0     2     7
```

Figura 6. KNN para classificação por níveis de DON com *k* = 7.

Avaliação do erro ao aumentar o valor de K para classificação por níveis de DON:

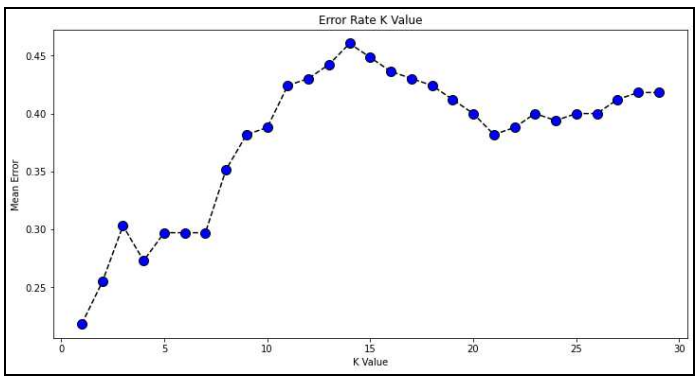

Figura 7. Avaliação do erro ao aumentar o valor de *k* para classificação por níveis de DON.

Com o algoritmo KNN foram obtidos os seguintes resultados com o valor k=15 para decisão binária na Figura 8.

```
Classificação binária
Acurácia KNN1:  0.8424242424242424
-----------------
-----------------
Dados sobre o teste realizado:
              precision    recall  f1-score   support

           0       0.79      0.59      0.68        46
           1       0.85      0.94      0.90       119

    accuracy                           0.84       165
   macro avg       0.82      0.76      0.79       165
weighted avg       0.84      0.84      0.83       165

-----------------
-----------------
Matriz Confusão - Classificação binária KNN
              sadio  contaminado
p_sadio          27            7
p_contaminado    19          112
```

Figura 8. KNN para classificação binária com valor *k* = 15.

Avaliação do erro ao aumentar o valor de K para classificação binária:

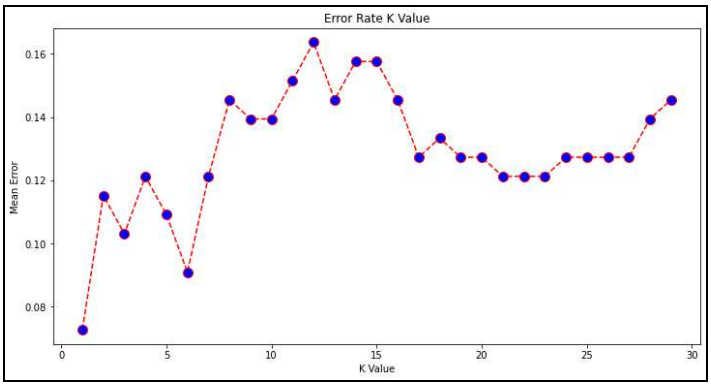

Figura 9. Avaliação do erro ao aumentar o valor de *k* para classificação binária.

Com o algoritmo SVM foram obtidos os seguintes resultados:

```
Teste de Classificação por níveis de DON com o SVM
Acurácia SVM:  0.6848484848484848
-----------------
-----------------
Dados sobre o teste realizado:
              precision    recall  f1-score   support

           0       0.89      0.63      0.74        65
         307       1.00      0.79      0.88        24
         483       0.42      0.56      0.48         9
         799       0.60      0.90      0.72        10
        1508       0.81      0.87      0.84        15
        1788       0.50      0.50      0.50        12
        1943       0.69      0.60      0.64        15
        2009       0.53      0.80      0.64        10
        2113       0.18      0.60      0.27         5

    accuracy                           0.68       165
   macro avg       0.62      0.69      0.63       165
weighted avg       0.77      0.68      0.71       165

Matriz Confusão - SVM por níveis de DON
       0  307  483  799  1508  1788  1943  2009  2113
0     41    0    5    1     0     3     4     2     9
307    0   19    0    5     0     0     0     0     0
483    0    0    5    0     2     1     0     1     0
799    1    0    0    9     0     0     0     0     0
1508   0    0    2    0    13     0     0     0     0
1788   0    0    0    0     0     6     0     3     3
1943   2    0    0    0     1     2     9     0     1
2009   1    0    0    0     0     0     0     8     1
2113   1    0    0    0     0     0     0     1     3
-----------------
-----------------
```

Figura 10. SVM para classificação dos níveis de DON.

```
Classificação binária por SVM
Acurácia SVM:  0.8909090909090909
-----------------
-----------------
Dados sobre o teste realizado:
              precision    recall  f1-score   support

           0       0.89      0.70      0.78        46
           1       0.89      0.97      0.93       119

    accuracy                           0.89       165
   macro avg       0.89      0.83      0.85       165
weighted avg       0.89      0.89      0.89       165


Matriz Confusão - Classificação binária SVM
               sadio  contaminado
p_sadio           32            4
p_contaminado     14          115
```

Figura 11. SVM para classificação binária.

Com o algoritmo *random forest* foram obtidos os seguintes resultados:

```
Classificação por níveis de DON
-----------------
-----------------
Dados sobre o teste realizado:
              precision    recall  f1-score   support

           0       0.82      0.87      0.84        46
         307       0.90      0.95      0.92        19
         483       0.86      1.00      0.92        12
         799       0.92      0.80      0.86        15
        1508       0.93      0.81      0.87        16
        1788       0.55      0.50      0.52        12
        1943       0.86      0.92      0.89        13
        2009       0.57      0.53      0.55        15
        2113       0.94      0.88      0.91        17

    accuracy                           0.82       165
   macro avg       0.82      0.81      0.81       165
weighted avg       0.82      0.82      0.82       165

Matriz Confusão - Random Forest por níveis de DON
       0  307  483  799  1508  1788  1943  2009  2113
0     40    0    0    0     2     2     1     3     1
307    0   18    0    2     0     0     0     0     0
483    0    0   12    0     1     0     0     1     0
799    0    1    0   12     0     0     0     0     0
1508   1    0    0    0    13     0     0     0     0
1788   1    0    0    1     0     6     0     3     0
1943   0    0    0    0     0     1    12     0     1
2009   3    0    0    0     0     3     0     8     0
2113   1    0    0    0     0     0     0     0    15
```

Figura 12. *Random forest* para classificação dos níveis de DON.

```
Classificação binária
              precision    recall  f1-score   support

           0       0.92      0.74      0.82        46
           1       0.91      0.97      0.94       119

    accuracy                           0.91       165
   macro avg       0.91      0.86      0.88       165
weighted avg       0.91      0.91      0.91       165


Matriz Confusão - Classificação binária Random Forest
               sadio  contaminado
p_sadio           34            3
p_contaminado     12          116
```

Figura 13. *Random forest* para classificação binária.

4. Conclusões e trabalhos futuros

A apresentação desenvolvida neste trabalho mostra como modelar dados coletados através de analises multiespectrais dos grãos de trigo e aplicar os algoritmos de *machne learning*: KNN, SVM *e random forest* e a partir de informações resultantes prever se o nível 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 ainda carece de evolução. Por outro lado, essa abordagem consegue satisfazer 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, busca-se 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. Partindo deste princípio o trabalho pode-se usar como fonte de mais pesquisas.

5. Referências

Agelet, L. E. and Hurburgh Jr., C. R. (2014). Limitations and current applications of near infrared spectroscopy for single seed analysis. Talanta, v. 121, p. 288-299. Disponível: https://www.sciencedirect.com/science/article/abs/pii/S0039914013010205. Acesso: junho, 2021.

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.

Arnon, D. I. and Stout, P. R. (1939). The essentiality of certain elements in minute quantity for plants with special reference to copper. Plant Physology, v. 14, n. 1, p. 371–375.

Borém, A. e Scheeren, P. L. (2015) Trigo: do plantio à colheita. Viçosa, MG: Ed. UFV. 260 p. Disponível: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1022684/trigo-do-plantio-a-colheita. Acesso: junho/2021.

Chwif, L. (2015). Modelagem e simulação de eventos discretos: teoria & aplicações / Leonardo Chwif, Afonso C. Medina. 4. ed. Rio de Janeiro: Elsevier.

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.

Flandrin, J. L. e Montanari, M. (1998). História da Alimentação. São Paulo: Estação Liberdade. p. 16.

Food and Agriculture Organization of the United Nations. (2009). Global agriculture towards 2050. FAO. Disponível: http://www.fao.org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdf. Acesso: novembro/2020.

Food and Agriculture Organization of the United Nations (2021). Crop Prospects and Food Situation - Quarterly Global Report. FAO. Disponível: http://www.fao.org/documents/card/en/c/cb3672en. Acesso: junho/2021.

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.

Malavolta, E., Vitti, G.C. e Oliveira, S.A. (1997). Avaliação do estado nutricional das plantas: princípios e aplicações. 2 ed. Piracicaba: Potafós. 319 p.

Mandarino, J. M. G. (1994). Componentes do trigo: características físico-químicas, funcionais e tecnológicas. Londrina, EMBRAPA-CNPSo. 36 p.

Marques, J. et al. (2018). Operação de um manipulador por meio da detecção de gestos baseada em Aprendizado de Máquina.

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.

Pasquini, C. (2003). Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. Journal of the Brazilian Chemical Society, v. 14, n. 2, p. 198-219.

Rezende, S. O. (2005). Sistemas Inteligentes: Fundamentos e Aplicações. 1. ed. São Paulo: Manole. 525 p.

Rossi, R. M. e Neves, M. F. (2004). Estratégias para o trigo no Brasil PENSA/UNIEMP, São Paulo. Editora Atlas. 224 p.

Russel, S. J. (2013). Inteliência Artificial / Stuart Russel, Peter Norvig; tradução Regina Célia Simille. Rio de Janeiro: Elsevier. Tradução de: Artificial Intelligence, 3rd ed.

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.