TÉCNICAS DE CIÊNCIAS DOS DADOS APLICADAS À AGRICULTURA

Eduardo Vicente do Prado[1]

[1]Centro Universitário Amparense – UNIFIA, Amparo – SP. Brasil.

Resumo:
A utilização de imagens de satélites, de dados climáticos, de dados de sensores instalados no solo e nas plantas, geram um volume de dados muito grande no dia a dia de uma propriedade agrícola. Processar e analisar todos esses dados não é uma tarefa fácil. É necessário utilizar técnicas modernas para análise e processamento de grandes volumes de dados, gerando informações úteis que serão usadas para tomadas de decisão no processo produtiva da propriedade agrícolas. Entre essas técnicas modernos, destaca-se a Ciência de Dados. O objetivo deste trabalho foi apresentar as técnicas da Ciência de Dados mais utilizadas no tratamento, processamento e análise de grandes volumes de dados agrícolas na geração de informações que darão suportes nas tomadas de decisões do produtor agrícola.
Palavras-chave: Internet das Coisas, Aprendizado de Máquinas, Redes Neurais Artificiais, Sensores.

Abstract:
The use of satellite images, climate data, data from sensors installed in the soil and plants, generate a very large volume of data in the day to day of an agricultural property. Processing and analyzing all this data is not an easy task. It is necessary to use modern techniques for analyzing and processing large volumes of data, generating useful information that will be used for decision-making in the production process of the agricultural property. Among these modern techniques, Data Science stands out. The objective of this work was to present the most used Data Science techniques in the treatment, processing and analysis of large volumes of agricultural data in the generation of information that will support the decision-making of the agricultural producer.
Keywords: Internet of Things, Machine Learning, Artificial Neural Network, Sensors,

1. INTRODUÇÃO

De acordo com relatório da FAO, a população mundial ultrapassará 9,0 bilhões de pessoas até o ano 2050 (FAO, 2009). Desafios terão de ser superados para atingir o nível de produtividade agrícola para atender à demanda mundial prevista por alimentos, rações, fibras e combustível em 2050. A aplicação de novas tecnologias às atividades produtivas e o oferecimento de serviços tecnológicos destinados ao setor agropecuário têm um grande potencial para incrementar rentabilidade, reduzir perdas e desperdícios e contribuir para o aumento da quantidade e qualidade dos produtos e dos processos produtivos, bem como para minimizar os prejuízos decorrentes de infestações de pragas, manifestações

de doenças e eventos climáticos (THE ECONOMIST, 2016). Para atingir um nível de crescimento adequado à expectativa de elevação da população mundial, o setor agropecuário deve adotar algumas práticas, como o monitoramento e o controle de variáveis que influenciam a produtividade. A Agricultura Digital (AD) se apresenta como uma entre muitas soluções para os grandes desafios que a agricultura e o mundo enfrentam atualmente. A principal característica AD é o uso intensivo de dados. Os dados estão disponíveis em quantidade e frequências espacial e temporal em grande escala e passaram a ser insumos para os processos de tomada de decisão (SARAIVA et al., 2020). O volume de dados gerados é muito grande (*Big Data*). As abordagens tradicionais de análises de dados não são mais adequadas. Além dos grandes volumes de dados, a diversidade nos formatos, origens dos dados e velocidade com que são gerados trouxeram a necessidade de evolução na forma como são analisados. A Ciência de Dados vem atender essa demanda e se caracteriza por uma abordagem multidisciplinar em que, além dos profissionais de Estatística e Ciência da Computação, os de outras áreas são necessários, como os da agricultura. Os métodos e técnicas de Inteligência Artificial (IA) têm ganhado mais importância na análise de *Big Data*, tanto pela busca de maior eficiência quanto para revelar informações que não seriam óbvias com a aplicação dos métodos de análises de dados tradicionais. A AD se refere ao emprego de métodos computacionais de alto desempenho, redes de sensores, comunicação máquina para máquina (M2M), conectividade entre dispositivos móveis, computação em nuvem, métodos e soluções analíticas para processar grandes volumes de dados e construir sistemas de suporte à tomada de decisões de manejo. Estão incluídas no escopo da AD, a Agricultura de Precisão (AP), a automação e a robótica agrícola, além de técnicas de *Big Data* e a Internet das Coisas (IoT). Práticas produtivas agropecuárias eficientes envolvem o gerenciamento de uma ampla matriz de variáveis que incluem: dados sobre tempo e clima, nutrientes e umidade do solo; informações sobre infestação de ervas daninhas, pragas e doenças; o planejamento de atividades e o levantamento dos custos relacionados às atividades de produção, podendo maximizar rendimentos e lucros e minimizar custos e riscos associados à atividade.

2. ALGUMAS DEFINIÇÕES

2.1. Ciências de Dados

A Ciência de Dados se utiliza de processos científicos e computacionais para extrair conhecimento, padrões e tendências a partir de conjuntos de dados de vários formatos, estruturados ou não-estruturados. Seu objetivo é aprimorar a tomada de decisão para entender fenômenos por intermédio da análise automatizada de dados. Surgiu a partir da convergência de muitos campos de estudo

tradicionais como: análise causal; estatística; métodos de visualização de dados, etc. A CD pode levar a descobertas transformadoras com a integração entre campos científicos distintos. Na agricultura é importante uma abordagem sistêmica envolvendo o setor produtivo, considerando a adoção de vários tipos de tecnologias da informação e comunicação (TIC) como: o uso de sensores para a captura de dados; técnicas de integração de dados; robôs e Sistema de Aeronave Remotamente Pilotada (RPAS) entre outros. A agricultura deve utilizar os dados para entender fatores que afetam a produtividade e a geração de perdas nas cadeias produtivas. O conceito de *Smart Farming* envolve o monitoramento automatizado das variáveis que influenciam a produção de forma precisa e barata e o suporte à tomada de decisão do produtor pelo processamento e análise destas variáveis. Este paradigma está associado à existência de dificuldades como: a implementação de sensores ao ar livre e em grandes distâncias; o acesso a redes de telefonia móvel e internet e à própria resistência do produtor em adotar estas novas tecnologias. Dentre as aplicações de CD ao contexto da *Smart Farming*, cita-se: análise de dados de redes de sensores de forma a identificar padrões e possibilidades de intervenção; integração de sensores, simuladores e atuadores; realidade aumentada; modelos de alertas para a ocorrência de pragas e doenças a partir de dados de clima; sistemas de recomendações de conteúdo e sistemas de rastreabilidade baseados em cadeias de blocos (*Blockchain*).

A CD revela tendências e produz as informações que os produtores podem usar para tomar melhores decisões e produzir mais, com mais qualidade, diminuir custos e melhorar a aplicação e uso de insumos. Talvez o mais importante seja que a CD permite que os modelos de Aprendizado de Máquinas (ML) aprendam com as grandes quantidades de dados que estão sendo fornecidos a eles, em vez de depender principalmente de analistas para ver o que podem descobrir a partir dos dados. Os dados são à base da inovação, mas seu valor vem das informações que os analistas podem extrair e depois usar. A CD abrange a preparação de dados para análise, incluindo limpeza, agregação e manipulação para realizar análises avançadas. Os aplicativos analíticos e os analistas podem revisar os resultados para descobrir padrões e permitir que os produtores tomem decisões mais assertivas, em tempo real. A CD utiliza várias técnicas de IA para realizar suas análise e retirar informações dos dados. Entre as técnicas mais utilizadas estão o *Machine Learning*, o *Deep Learning* (DL) e Redes Neurais Artificiais (RNA).

2.2. *Big Data*

O *Big Data* na agricultura pode ser entendido como um grande volume de dados gerado com grande velocidade na interseção das geotecnologias, informação sobre a produção no campo, sobre tempo e clima e sobre o mercado. A inclusão de informações de diferentes estratégias de gestão, junto a outras variáveis de interesse, aumenta a profundidade da análise para definir estratégias de gestão. O BD

inclui dados espaciais, específicos ao local e associados com AP, incluindo dados de sensores no solo e de plantas, de máquinas e implementos agrícolas; e os metadados sobre práticas de gestão e tecnologias, como profundidade de plantio, colocação da semente, cultivar, condição de uso de máquinas, datas de plantio, aplicação de fertilizantes; dados de eventos climáticos e evapotranspiração, entre outros, que não são controláveis pelo produtor.

2.3. Inteligência Artificial

A Inteligência Artificial é uma ciência multidisciplinar que desenvolve e aplica técnicas computacionais que simulam o comportamento humano em atividades específicas. Um sistema de IA deve ser capaz de fazer três coisas: (i) armazenar conhecimento, (ii) aplicar o conhecimento armazenado para resolver problemas e (iii) adquirir novo conhecimento através da experiência. Um sistema de IA deve possuir três componentes fundamentais: representação, raciocínio e aprendizagem.

2.4. Aprendizado de Máquina (*Machine Learning*)

O Aprendizado de Máquina pode ser entendido como a capacidade dos computadores aprenderem com base em dados disponíveis (VALENTE et al., 2020). Os algoritmos de ML são classificados em algoritmos de classificação ou regressão. Classificação: gera resultados categóricos ou discretos. Regressão: gera valores contínuos. Os modelos de ML podem utilizar dados categóricos e numéricos, como binários, inteiros e reais, provenientes de diferentes tipos de sensores. Os algoritmos de ML também podem ser classificados em aprendizado supervisionado e não supervisionado**.** Supervisionado: para cada amostra de entrada é apresentado ao algoritmo o resultado esperado na saída, conhecido como rótulo (*K-Nearest Neightbours* KNN; SVM; *Randon Forest*; RNA). Não supervisionado: apenas as amostras de entradas são apresentadas ao algoritmo e os dados utilizados para ajustar ou treinar os modelos não são rotulados. Espera-se que o algoritmo classifique os dados automaticamente. Os principais algoritmos são: *K-Means* e *Fuzzy K-Means* (VALENTE et al., 2020). Um dos benefícios da AD é a obtenção de grandes volumes e variedades de dados do sistema agrícola. Esses dados devem ser processados para extração de informações úteis na tomada de decisão. Modelos de ML utilizam esses dados para responder a perguntas como: qual o nível de severidade de uma doença, a dosagem de fertilizantes, a lâmina de água a ser aplicada, a previsão de chuva, entre outras. ML é uma ramificação da IA baseado em estatística computacional e procedimentos de otimização, que explora técnicas de aprendizado auto aperfeiçoadas para a solução de problemas ou execução de tarefas específicas. Diferente de outras abordagens da IA, ramificações como a do ML tentam construir sistemas que não precisam ser programados para realizar as tarefas. Ainda, no caso específico do ML,

são construídos modelos matemáticos a partir de dados de amostragem, chamados de dados de treino, para que os parâmetros do modelo sejam adaptados progressivamente até que seu desempenho em tarefas específicas seja melhorado sem qualquer intervenção humana.

2.5. Internet das Coisas (*Internet of Things*)

A Internet das Coisas refere-se à tecnologia de conectar equipamentos, objetos ou itens do dia a dia, capazes de trocar dados entre si, via internet. Tornar a agricultura inteligente significa introduzir um conjunto de metodologias inovadoras e tecnologias conectadas com o objetivo de otimizar e aumentar a eficiência do uso de insumos, produção e lucratividade de forma sustentável. A terra se torna um substrato onde diferentes tipos de sensores podem adquirir dados heterogêneos. Esses sensores são conectados em uma espécie de rede rural, por sua vez conectada à internet. Os dados, em tempo real, são armazenados em um banco de dados contendo todo o conhecimento necessário sobre as características do terreno. A IoT pode ser aplicada de diversas formas na agricultura, desde sistemas de telemetria, *softwares*, levantamento de dados, monitoramento e controle de automação de maneira eficiente. A telemetria possibilita que dados dos sensores instalados em campo e nos equipamentos sejam coletados e compartilhados de forma remota e em tempo real. Os equipamentos e sensores conectados à central podem ser visualizados em tempo real, desde que estejam com sinal de internet, e as operações podem ser corrigidas e atualizadas.

2.6. Redes Neurais Artificiais

As Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência. Uma RNA é composta por várias unidades de processamento. Essas unidades, geralmente são conectadas por canais de comunicação que estão associados a determinado peso sináptico. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma RNA vem das interações entre as unidades de processamento da rede.

2.7. Aprendizado Profundo (*Deep Learning*)

O Aprendizado Profundo (DP) é uma subárea de ML e utiliza RNA em sua estrutura. Como uma RNA é um paradigma conexionista ponderado por pesos, a capacidade de desenvolver modelos inteligentes está nas conexões de uma quantidade significativa de neurônios artificiais e não nos neurônios em si, as operações aritméticas crescem de forma exponencial no tocante a DP. Nesse ponto é

importante o conceito de BD, uma vez que esse tipo de arquitetura tende a funcionar melhor com mais dados de entrada. O que difere uma RNA de uma Rede Neural Profunda é a quantidade de neurônios e de conexões, sendo maior no segundo caso. Uma RNA simples possui até cinco camadas, uma Rede Neural Profunda possui mais de cinco. Um computador comum não tem capacidade de processamento suficiente para treinar Redes Neurais Profundas.

2.8. Outros algoritmos

Valente et al. (2020) destacam outros dois algoritmos de ML: o *Decision Tree* (árvore de decisão) e o *Random Forest* (floresta aleatória). *Decision Tree*: constrói modelos de regressão ou classificação na forma de uma estrutura em árvore de decisão. Este algoritmo divide o conjunto de dados em subconjuntos menores. Espera-se que o novo subconjunto de dados tenha resultados mais homogêneos em relação ao conjunto original (tenha menor impureza). O algoritmo seleciona um atributo e determina um limiar para o atributo que melhor separa o conjunto de dados. O atributo que gerar o subconjunto com menor impureza será colocado no topo da árvore (nó raiz). Assim, irá ocorrer a subdivisão dos nós da árvore até que se chegue a uma decisão final (nó folha), na qual não haverá mais subdivisões. A subdivisão poderá ser interrompida quando o resultado gerado tiver impureza maior que a dos dados originais. Outro critério de parada na subdivisão poderá ser pela definição de algum parâmetro, como a máxima profundidade da árvore. Os algoritmos de ML apresentam melhor acurácia quando utilizam mais de um algoritmo para a tomada de decisão; também, pode-se conseguir uma melhor acurácia, gerando vários modelos com um mesmo algoritmo modificando a fonte de dados. É com base nessa estratégia que funciona o algoritmo *Random Forest*. Este algoritmo gera vários modelos de *Decision Tree* com base na escolha aleatória de parte dos dados de treinamento. O algoritmo *Random Forest* apresenta os seguintes passos: (1) construir um subconjunto de dados aleatórios a partir de dados de treinamento**;** (2) selecionar aleatoriamente um subconjunto de atributos a partir do subconjunto no passo 1; (3) criar uma árvores de decisão a partir do subconjunto; e (4) repetir o passo 1, 2 e 3 para gerar várias árvores diferentes. O número de árvores geradas é um parâmetro que pode ser determinado. A classificação final será definida para a classe que obtiver o maior número de votos. Se for regressão, o resultado poderá ser obtido pela média das predições de cada árvore. Este algoritmo apresenta grande capacidade de aprendizagem (ajuste). Dessa forma, para evitar superajuste, recomenda-se limitar o grau de liberdade do algoritmo. Para isso, deve-se definir alguns parâmetros do algoritmo, como o número de árvores geradas e a profundidade máxima da árvore. A escolha do melhor parâmetro deverá ser com base na validação cruzada.

3. APLICAÇÕES DE TÉCNCIAS DE CIÊNCIAS DE DADOS NA AGRICULTURA

Pantazi et al. (2017) utilizaram dados de imagem de refletância hiperespectral e RNA para detecção de estresse por nitrogênio (N), infecção de ferrugem amarela em plantas de trigo de inverno. As medidas das reflectâncias foram tomadas dos dosséis das plantas. O estresse de N foi detectado com 99,63% de precisão, a ferrugem amarela com 99,83% de precisão e as plantas saudáveis com 97,27% de precisão.

Grinblat et al. (2016) utilizaram DP para a identificação e classificação de três espécies de leguminosas: feijão branco, feijão vermelho e soja, via padrões de veias das folhas. Como resultados obtiveram para o feijão branco 90,2% de precisão, feijão vermelho com 98,3% de precisão e soja com 98,8% de precisão utilizando cinco camadas nas RNAs.

Ramos et al. (2020) utilizaram a técnica RF com índices de vegetação (IV) para predizer a produtividade do milho. A regressão utilizando RF proporcionou maior acurácia com coeficiente de correlação e um Erro Médio Absoluto de 0,78 e 853,11 $kg.ha^{-1}$, respectivamente. Concluíram que a estratégia baseada em classificação de IV é apropriada para prever a produtividade do milho usando ML e dados derivados de imagens multiespectrais. Essa abordagem reduz o número de IV necessários para determinar com alta precisão e baixo erro médio absoluto, e a abordagem pode contribuir para ações de tomada de decisão, resultando em um manejo preciso dos campos de milho.

4. CONCLUSÕES

Em resumo, a Ciência de Dados é uma ciência que estuda os dados, as informações, seu processo de aquisição, transformação, geração e, posteriormente, a análise desses dados agronômicos com o intuito de encontrar conhecimento relacionados a eles que colaborem para tomadas de decisões no campo. Dados agronômicos é tudo aquilo que pode ser coletado no campo para análise de dados para combater pragas e aumentar a produtividade na lavoura, desde tipo de híbrido, fertilizante de solo, manejo, maquinários etc. A Ciência de Dados, juntamente com as técnicas de Inteligência Artificial são ferramentas muito poderosas que ajudam a entender melhor as informações e auxiliam a fazer previsões precisas sobre o futuro. Essa combinação é única: ter uma boa ferramenta de monitoramento, registrar detalhadamente os dados relativos ao processo preventivo e usar inteligência que agrega valor às informações para a tomada de decisões mais assertivas.

REFERÊNCIAS BIBLIOGRÁFICAS

FAO, 2009. **The State of Food and Agriculture**. FAO, Rome, Italy. Disponível em: http://www.fao.org/3/a-i0680e.pdf. <Acessado em 10 de setembro de 2022 >.

GRINBLAT, G. L. [et al.]. Deep learning for plant identification using vein morphological patterns. **Computers and Electronics in Agriculture**, 2016.

PANTAZI, X. E. [et al.]. Detection of biotic and abiotic stresses in crops by using hierarchical self-organizing classifiers. **Precisin Agriciculture**, 2017.

RAMOS, A. P. M. [et al.]. A radom forest ranking approach to predict yeld in maize with uav-based vegetation spectral indices. **Computer and Electronics in Agriculture**, 2020.

SARAIVA, A. M. [et al.]. Dados digitais: ciclo, padronização, qualidade, compartilhamento e segurança. **In:** QUEIROZ, D. M. [et al.]. Agricultura digital. Ed. UFV: Viçosa, MG, 2020.

THE ECONOMIST. **The future of agriculture.** Technology Quarter. Jun 11th, 2016. Disponível em: http://www.economist.com/technology-quarterly. <Acessado em 12 de setembro de 2022>.

VALENTE, S. D. M. [et al]. Machine learning. **In:** QUEIROZ, D. M. [et al.]. Agricultura digital. Ed. UFV: Viçosa, MG, 2020.