http://journal.unoeste.br/index.php/ce/index
DOI: 10.5747/ce.2021.v13.n4.e378
ISSN on-line 2178-8332

Colloquium Exactarum

Submetido: 23/04/2021 Revisado: 16/12/2021 Aceito:17/03/2022

APRENDIZADO DE MÁQUINA APLICADO EM IMAGEM NDVI PARA PREVISÃO DA PRODUTIVIDADE DA CANA-DE-AÇÚCAR

Machine Learning Applied in NDVI Image for Forescasting Sugarcane Productivity

Luiz de Souza Rodrigues[1], Danilo Roberto Pereira[2]

[1]Pesquisador Independente, [2]Analytics2Go
E-mail: luiz.srodrigues@hotmail.com, dpereira@analytics2go.com

RESUMO – Este artigo apresenta uma abordagem através de modelos baseados em ML (*Machine Learning*) aplicados em Imagens NDVI (*Normalized Difference Vegetation Index*) para estimativas da produtividade na cultura da Cana-de-Açúcar. O uso de técnicas humanas baseadas em experiências cognitivas é predominante para prever a produtividade. As imagens utilizadas foram o NDVI fornecido pelo satélite Sentinel-2, sendo que os conjuntos de dados foram obtidos a partir dos pontos de georreferenciamento dos talhões e aplicados às imagens para extração e processadas. Os modelos dos algoritmos preditivos utilizados foram: (i) CNN (*Convolution Neural Network*), (ii) KNN (*K-Nearest Neighbors*), (iii) RF (*Random Forest*), (iv) SVM (*Support Vector Machie*), (v) AdaBoost (*Adaptive Boosting*). O algoritmo de RF apresentou-se o mais eficiente, de modo que os resultados para o DP (Desvio Padrão) e a fórmula para o MSE (*Mean Square Error*) obtiveram 30,71 toneladas (t) e o MAE (*Mean Absolute Error*) obteve 3,73 (t). Na relação das estimativas, a fórmula do DP para o MSE obteve 34,71 (t) e o MAE de 3,97 (t). O EM (Erro Médio) para as estimativas foi de -8,80% e o algoritmo RF de 0,012%. Os resultados mostraram-se consistentes para as estimavas da produtividade na cultura da Cana-de-Açúcar.
Palavras-chave: Aprendizagem de Máquina, Cana-de-Açúcar, NDVI.

ABSTRACT – This article presents an approach through models based on ML (Machine Learning) applied to NDVI (Normalized Difference Vegetation Index) images to estimate productivity in the sugarcane crop. The use of human techniques based on cognitive experiences is predominant to anticipate productivity. The images used were provided by the NDVI Sentinel-2 satellite, since the datasets were obtained from two georeferenced points, two plots and applied to the images for extraction and processing. Two predictive algorithms are used for the models: (i) CNN (Convolution Neural Network), (ii) KNN (K-Nearest Neighbors), (iii) RF (Random Forest), (iv) SVM (Support Vector Machie) , (v) AdaBoost (Adaptive Boost). The RF algorithm was presented or more efficient, so that the results for the DP (Standard Deviation) and the formula for the MSE (Mean Square Error) obtained 30.71 tons (t) and the MAE (Mean Absolute Error) obtained 3.73(t). Regarding the estimates, the DP formula for the MSE obtains 34.71 (t) and the MAE of 3.97 (t). The EM (Mean Error) for the estimates was -8.80% and the RF algorithm was 0.012%. The results will show consistency for the productivity estimates in the sugarcane crop.
Keywords: Machie Learning, Sugar Cane, NDVI.

1. INTRODUÇÃO

O Brasil possui uma diversidade de culturas em razão do clima que estabelece ao longo de sua extensão territorial. Dentre as culturas, destaca-se a Cana-de-Açúcar, tendo sua alavancagem a partir da política de incentivo ao PROÁLCOOL (Programa Nacional do Álcool) pelo governo, em 1975, (BRASIL, 1975), considerada fonte de energia renováveis que auxilia na diminuição dos gases do efeito estufa (LUCIANO *et al*., 2018; DO VALLE GONÇALVES *et al*., 2017). Diante dos avanços, técnicas humanas baseadas em experiências empíricas e históricos da produtividade foram empregadas a fim de prever a produtividade do ciclo da cultura da Cana-de-Açúcar. Em seu trabalho, Charoen-Ung e Mittrapiyanuruk (2018) destacam as desvantagens de usar métodos de estimativas de rendimento baseado em empirismo, já que há uma grande discrepância entre os rendimentos estimados e os reais.

Segundo Ebadi *et al*. (2019), o crescimento pela atenção aos dados como um ativo, por um lado, e as habilidades e os recursos necessários para obter vantagem comercial com os dados, por outro lado, tiveram profundas implicações na maneira como as empresas estão definindo estratégias para incorporar Inteligência Artificial (IA).

Trabalhos como os de Geetharamani e Arun Pandian (2019), Militante, Gerardo e Medina (2019), Cobeña Cevallos, Atiencia Villagomez e Andryshchenko (2019), Treebupachatsakul e Poomrittigul (2019) e Ghosal e Sarkar (2020) fizeram o uso da DL (*Deep Learning*) com intuito de automatizar o reconhecimento usando CNN (*Convolutional Neural Network*).

Em seu estudo, Khetkeeree (2020) afirma que o Sentinel-2 é um satélite de observação da Terra com Imagem multiespectral de alta resolução, em que os intervalos das imagens da mesma área ocorrem a cada cinco dias. Além disso, as informações dos serviços de satélite podem ser abertas e livremente acessíveis. O autor, Khetkeeree (2020), reforça ainda, que existem muitas utilizações destes satélites, tanto que o autor, Laneve *et al*. (2017) faz o uso Sentinel-2 em seu estudo.

No caso de imagens de satélites, há desvantagens quanto à presença de nuvens (nebulosidade) que interferem na qualidade ou de vegetação densa, em se tratando de pesquisas, por exemplo, de pragas. Nos casos das imagens capturadas por *Drones* (equipados com sensores multiespectrais), apesar de exigirem operações manuais e maior tempo para operacionalizar (capturas), há vantagens, com relação à baixa altitude na captura (excludente a nuvens) resultando em imagens mais definidas. Por outro lado, as desvantagens (*Drones*) são quanto à necessidade de capturar novas imagens durante curto espaço de tempo para áreas com grande extensão de culturas e o estado fenótipo (fato importante quanto à maturidade da planta), sendo que no caso de satélites isso não é impedimento. Os autores Vasconcellos *et al*., (2020), em seu trabalho, combinam aprendizagem de máquina e Imagens de Veículos Aéreos Não Tripulados (VANT) para monitorar animais e destacam que o aspecto negativo é interpretar visualmente o grande número de imagens de alta resolução adquiridas.

As imagens capturadas por *Drones* possuem uma boa qualidade em definições, e isso contribui para os estudos, como por exemplo, de ervas daninhas, falhas em germinação de culturas, análises das variedades de culturas, pragas e doenças, já que precisão é algo fundamental para os resultados, como destaca os autores RUBIRA CRULHAS *et al*., (2018).

Enquanto os estudos, como dos autores Charoen-Ung e Mittrapiyanuruk (2018), preveem o grau de rendimentos numa escala determinada pelos seguintes graus: “baixo rendimento, rendimento médio-volume e volume de alto rendimento”, os autores Skittou *et al.* (2020), procuram diferenciar áreas de vegetação, edificações ou outros usos da terra. Já os autores Scrivani, Zullo e Romani (2017) fazem uso das imagens NDVI com a finalidade de prever a produtividade de sete municípios do estado São Paulo.

As principais abordagens dos estudos aplicam suas pesquisas em espaços geográficos abrangentes (grandes extensões de culturas) e sem considerar estágios fenótipos específicos das culturas, com isso, a previsibilidade limita-se a abordagens restritas. Neste trabalho, o intuito é considerar os estágios fenótipos da cultura (imagens com periodicidade média de 30 dias antes da colheita da Cana-de-Açúcar), assim como os espaços georreferenciados inferiores a cinquenta hectares (como, por exemplo, talhões) em que são fundamentais. O objetivo com essa

abordagem é contribuir como medida para tomada de decisão, tal como o processamento da Cana-de-Açúcar (indústria), o direcionamento das metas, as decisões estratégicas corporativas, a previsão de vendas, a revisão de áreas improdutivas, as manutenções e as reduções de custos.

Este trabalho foi realizado em uma empresa do ramo de atividade agroindustrial. E que, por sua vez, possui em seu processo de gestão (anualmente em fevereiro) ou em eventuais intercorrências, como fenômenos da natureza (secas), a avaliação para estimar a produtividade da Cana-de-Açúcar. As estimativas empregadas não usam de procedimentos estatísticos ou técnicos específicos, mas aplica conhecimento empírico. A partir do conhecimento empírico, as equipes fazem visitas às plantações, realizando anotações, e depois compilam-se os dados, transformando-os em valores que resultam em estimativas da produtividade da cultura da Cana-de-Açúcar.

Nesse sentido, este trabalho foi conduzido através de uma abordagem utilizando metodologias que sejam capazes de predizerem a produtividade na cultura da Cana-de-Açúcar.

Sendo assim, o suporte deste trabalho, é aplicar conceitos de ML (*Machine Learning*) utilizando imagens NDVI (*Normalized Difference Vegetation Index*) para estimar a produtividade na cultura da Cana-de-Açúcar.

O trabalho está organizado da seguinte maneira. Na Seção 2 é feita uma revisão de alguns trabalhos relacionados. Na Seção 3 é apresentada uma abordagem sobre o referencial teórico, cujo objetivo consiste em solidificar o referido trabalho. Na Seção 4 é discorrido sobre os métodos empregados. Por sua vez, a Seção 5 descreve os materiais e métodos utilizados, e, na Seção 6, apresentam-se os resultados alcançados a partir dos modelos com algoritmos de ML aplicados. Por fim, na Seção 7, são feitas as considerações.

2. TRABALHOS RELACIONADOS

Nesta Seção são descritos os trabalhos que foram realizados e alinhados aos temas relacionadas à análise de imagens NDVI, de tal forma a dar consistência a essa pesquisa. A fim de responder as questões de predição da produtividade, como uma forma alternativa ao método manual (empírico), foi aplicada a utilização de imagens NDVI (após transformação) obtidas através de satélite e discutidas em trabalhos como os de (LUCIANO *et al.*, 2019a; LUCIANO *et al.*, 2018; DO VALLE GONÇALVES *et al.*, 2017).

Os autores Luciano *et al.* (2019a) e Duft e Picoli (2018) destacam que as variáveis espectrais mais utilizadas em estimativas da produtividade em cultura da Cana-de-Açúcar são as refletâncias R (*Red*) que são as regiões do espectro visível e NIR (*Near InfraRed*) que são os raios infravermelhos próximos, sendo que o indíce mais comum é o NDVI. Segundo Scrivani, Zullo e Romani (2017), o NDVI está relacionado à quantidade e à concentração de biomassa da vegetação, e é amplamente utilizada em pesquisas agrícolas. A vegetação absorve radiação no espectro visual e reflete uma grande quantidade de radiação NIR (DEN BESTEN *et al.*, 2020). Em sua pesquisa, os autores Fernandes, Ebecken e Esquerdo (2017), para prever a produção, aplicaram técnicas derivadas de séries temporais em imagens NDVI e RNAs (Redes Neurais Artificiais*)* para expressar resultados efetivamente mais precisos se comparados com dados oficiais.

Os autores Scrivani, Zullo e Romani (2017), Laneve *et al*. (2017) e Luciano *et al*. (2019b) buscaram, em seus trabalhos, validar os dados utilizando-se de processos de séries temporais em face da disponibilidade de imagens de satélites, de forma a combinarem técnicas matemáticas e estatísticas para resolverem questões relacionadas à predição na cultura da Cana-de-Açúcar. A disposição de imagens de satélite e sensores também é validada por Speranza, Antunes e Inamasu (2018) que concluíram ser potencialmente viável a utilização de imagens NDVI. Ainda nesse aspecto, Duft e Picoli (2018) fazem o uso de imagens para monitorar eventos da seca em cultura da Cana-de-Açúcar.

Em seu estudo, Khan *et al.* (2020) fazem uso de ML aplicada a imagens NDVI para estimar a cultura de Tabaco no país do Paquistão e obtiveram uma precisão de 95,81%. Para Abbas *et al.* (2018), os algoritmos de aprendizado de máquina podem efetivamente aumentar o desempenho da DL. Em seu estudo, Geetharamani e Arun Pandian (2019) destacam que as aplicações em sua maioria são propostas principalmente para análises de imagens médicas para os diagnósticos, já que na literatura as análises das doenças das plantas não têm sido tratadas com as mesmas ênfases.

Na classificação de doenças na folha do

arroz, usando CNN com DL, Ghosal e Sarkar (2020) obtiveram a acurácia de 92,46%. Em seu estudo, Treebupachatsakul e Poomrittigul (2019) propõem o desenvolvimento de uma aplicação de DL para a classificação de imagens de bactérias com o intuito de automatizar o processo de reconhecimento para reduzir o tempo de análise e aumentar a precisão.

Técnicas de modelagem para estimativa da produtividade agrícola são mais representativas do que as técnicas convencionais, em seu estudo Luciano *et al.* (2019b) usaram o algoritmo RF (*Random Forest*) para predizer a produção, comparando os modelos anuais e globais. Nesse sentido, Luciano *et al.* (2018) fizeram uso do algoritmo de classificação RF, o qual mostrou-se robusto tanto para erros nos preditores quanto em pequenos erros em mapas de referências. Para Treebupachatsakul e Poomrittigul (2019), o RF é uma combinação de preditores de árvores de decisão em que cada uma das árvores depende dos valores de um vetor aleatório de amostras, independentemente, e com a mesma distribuição para todas as árvores na floresta. Charoen-Ung e Mittrapiyanuruk (2018) reforçam em seu estudo que o método RF preditivo pode ser usado para regressão e classificação de tarefas.

Em seu estudo, Zhang *et al.* (2017) avaliam a vegetação e as propriedades do solo em aplicações de sensoriamento remoto por satélite, utilizando imagens NDVI através das bandas específicas para satélite Sentinel-2A. Em seu estudo, Wang *et al.* (2019) fazem uso do algoritmo *Polynomial*-SVM (RBF - *Radial Basis Function Kernel*) para classificação, e os sintetiza com o índice NDVI, a qual teve o melhor potencial em estimar as áreas com cultura da Cana-de-Açúcar. Em seu estudo, Kai *et al.* (2020), para distinguir as variedades da Cana-de-Açúcar, destacam os algoritmos RF e SVM, os quais obtiveram alta taxa de correlação, atingindo acurácia superiores à 80%. Em seu estudo, HU *et al.* (2020) descrevem que o algoritmo AdaBoost (*Adaptive Boosting*) é um tipo de aprendizado por reforço que pode realizar classificação de alta precisão treinando classificadores mais fracos, combinando-os, e transformando os resultados em dados mais precisos. Em seu trabalho, Hossain E., Hossain M. e Rahaman (2019) propuseram um método em que se aborda o algoritmo KNN (*K-Nearest Neighbor*) para classificar várias doenças que estão presentes nas folhas das plantas, de forma que o algoritmo forneceu precisão de 96.76%.

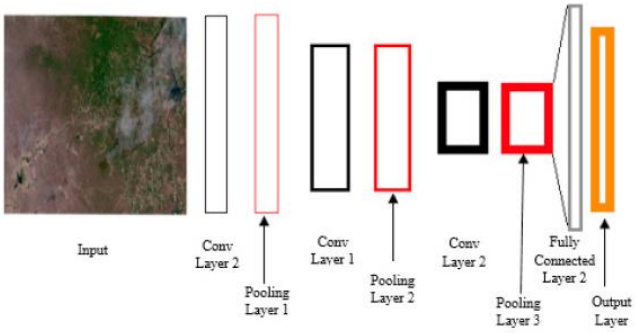
Os resultados dos estudos desses pesquisadores contribuíram para a fundamentação desse trabalho.

3. REFERENCIAL TEÓRICO

O objetivo nesta Seção será apresentar os principais trabalhos em pesquisas voltadas para o aprendizado de máquina ML (*Machine Learning*), relacionadas à DL.

Cobeña Cevallos, Atiencia Villagomez e Andryshchenko (2019) fazem o uso de CNN para reconhecimento de imagens espaciais da cultura da Cana-de-Açúcar, utilizando-se de ferramentas para melhorar as imagens como o *software ARGis*, e com uso de álgebra de mapeamento para obter o NDVI, de modo a submeter as imagens à rede CNN, conforme mostra o desenho da arquitetura na Figura 1, sendo que a mesma possui as medições fixas de 64 x 64 *pixels* como um dado de entrada, além de 32 níveis de testes de filtração, utilizando-se de 25 épocas para rede CNN com apenas duas camadas com 840 imagens, apresentando uma previsão média de 94,16%.

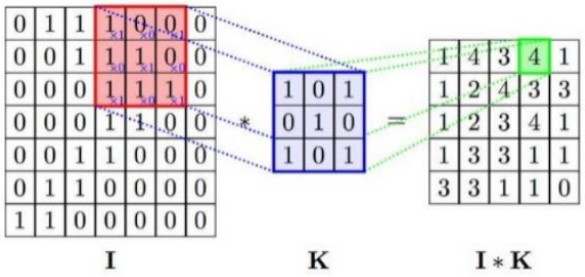
Figura 1. Modelo da Arquitetura de uma rede CNN.



Fonte: Cobeña Cevallos, Atiencia Villagomez e Andryshchenko (2019).

Segundo os autores, Militante, Gerardo e Medina (2019), uma CNN faz o uso de processamento de várias etapas que inclui uma imagem de entrada e camadas convolucionais, camadas de *pool*, camadas totalmente conectadas, funções de ativação e uma saída. A Figura 2 ilustra o fluxo de dados da convolução.

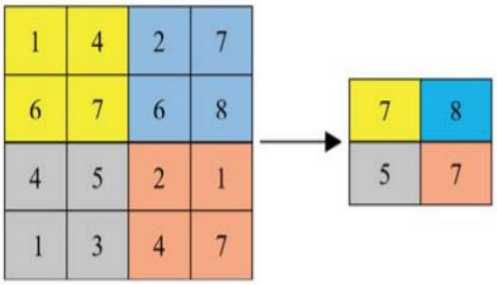
Figura 2. Fluxo de dados da convolução.



Fonte: Radu, Costea e Stan (2020).

Militante, Gerardo e Medina (2019) destacam que a camada é reduzida quanto ao excesso de ajuste e também ao tamanho do neurônio para a camada de amostragem descendente, conforme ilustra a Figura 3.

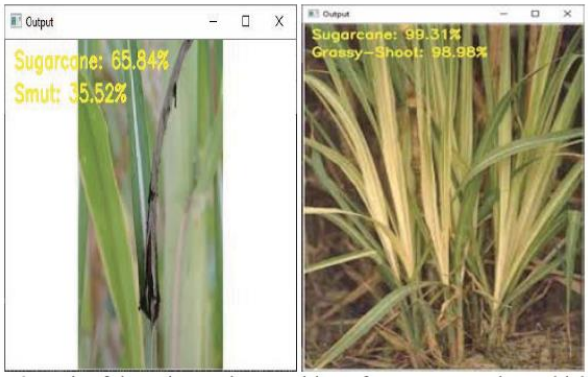
Figura 3. A redução da dimensionalidade na camada de *pooling* da CNN 2 x 2 *pixels* com o método *Max-Pooling* (considerando a região da matriz para o maior valor).



Fonte: Militante, Gerardo e Medina (2019).

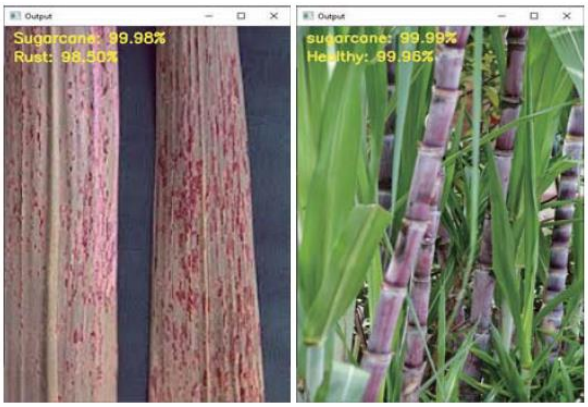
Militante, Gerardo e Medina (2019), através da rede CNN em que se analisou a folha da Cana-de-Açúcar, para classificar dentre as categorias saudável e doentes, obtiveram os seguintes resultados, conforme ilustra a Figura 4, 35,52% para a doença do "Smut" e 98,98% para a doença broto de gramíneas. Na ilustração da Figura 5, obteve-se 98,50% para a doença ferrugem e 99,96% como saudável e, na ilustração da Figura 6, obteve-se 50,23% para a doença ferrugem e 81,80% para a doença da folha amarela.

Figura 4. Resultado da detecção e reconhecimento de uma planta de Cana-de-Açúcar que mostra à esquerda 35,52% infectado com a doença do "Smut", e a imagem da direita mostra que 98,98% está infectado com a doença do broto de gramíneas.



Fonte: Militante, Gerardo e Medina (2019).

Figura 5. Mostra uma taxa de 98,50% de infecção da folha (imagem à esquerda) com a doença da ferrugem e taxa de 99,96% (imagem à direita) saudável.



Fonte: Militante, Gerardo e Medina (2019).

Figura 6. Mostra uma taxa de 50,23% de que a folha (imagem à esquerda) está infectada com ferrugem e de 81,80% de infecção (imagem à direita) com a doença da folha amarela.

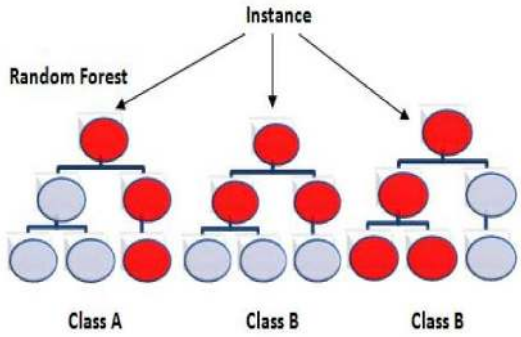


Fonte: Militante, Gerardo e Medina (2019).

Militante, Gerardo e Medina (2019) usaram técnicas para aumentar o conjunto de dados aleatoriamente, girando as imagens em 25 graus, alterando as imagens horizontalmente e verticalmente. O otimizador *"Adam"* foi incorporado, usando uma entropia cruzada categórica por meio da função de perda. O modelo foi capaz de treinar 60 épocas, usando um tamanho do lote de 32. A taxa de aprendizagem foi definida como "0,001" e reduzida por um fator de "0,3".

Segundo os autores Charoen-Ung e Mittrapiyanuruk (2018), o modelo *Random Forest* (RF) é um conjunto de árvores de decisão para cada TD (*decision tree*) treinada com subconjunto de dados aleatório de treinamento em que a amostragem é realizada com substituição. Além disso, a cada etapa, a melhor característica de divisão é escolhida de um subconjunto aleatório de características. Com isso, para tornar a previsão de uma instância de dados de teste, primeiro faz-se a previsão para cada TD e, em seguida, as previsões de todas as TDs no modelo são agregadas por votação forçada ou por livre votação, conforme a ilustração da Figura 7.

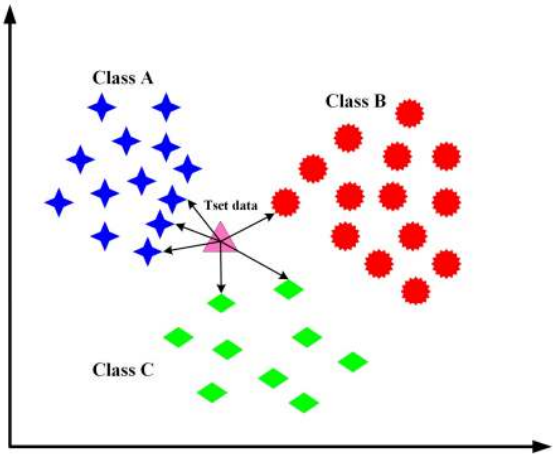
Figura 7. Ilustração gráfica *Random Forest*.



Fonte: Abbas (2018).

Segundo Nugrahaeni e Mutijarsa (2016), o algoritmo KNN (*K-Nearest Neighbor*) rotula as amostras baseadas em amostras vizinhas de um conjunto de treinamento, por exemplo, funções de distância, conforme ilustra a Figura 8. Um caso é classificado por um voto majoritário de seus vizinhos, sendo o caso atribuído à classe mais comum entre os vizinhos K mais próximos, medido por uma função de distância. Por exemplo, se K = 1, o caso é simplesmente atribuído à classe do seu vizinho mais próximo. Segundo Asadi e Pourhossein (2019), a função de distância pode ser Euclidiana, Manhattan, Minkowski ou Hamming, e estão descritas na Tabela 1.

Figura 8. Processo de classificação do KNN, usando como exemplo K = 6 e 3 classes.



Fonte: Asadi e Pourhossein (2019).

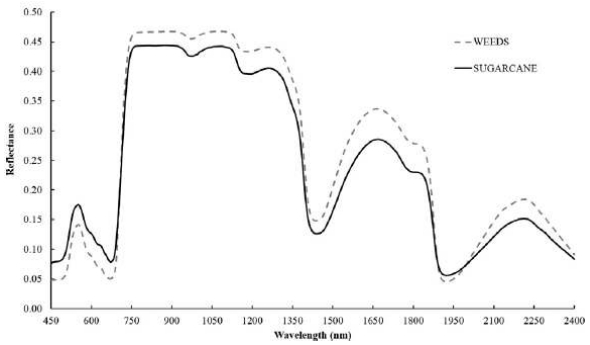
Tabela 1. Fórmulas das distâncias para as variáveis contínuas.

Métricas	Expressões Matemáticas
Euclidiana	$\sqrt{\sum_{i=1}^{k} (xi-yi)^2}$
Manhattan	$\sum_{i=1}^{k} \lvert xi - yi \rvert$
Minkowski	$\sum_{i=1}^{k} (\lvert xi - yi \rvert) q) 1^{/q}$
Hamming	$D_h = \sum_{i=1}^{k} \lvert xi - yi \rvert$

Fonte: Nugrahaeni e Mutijarsa (2016).

Em seu estudo, de Souza *et al.* (2020) concluíram que é possível diferenciar a cultura da Cana-de-Açúcar das plantas de ervas daninhas usando espectroscopia Vis-Nir (*Visible-Near Infrared*) e algoritmo de classificação para RF, conforme ilustra a Figura 9.

Figura 9. Espectros Vis-NIR médios de Cana-de-Açúcar e Ervas daninhas.



Fonte: Souza *et al.* (2020).

Os algoritmos KNN, SVM e RF foram objetos de estudos no trabalho de Skittou *et al.*

(2020) em que se fez o uso de imagens NDVI para diferenciar áreas de vegetação e áreas de construções, a qual obteve alta taxa de acurácia e o mínimo de erros. Nesse estudo, utilizou-se a proporção para o conjunto de dados, sendo 70% (treinamento) e 30% (teste), conforme demonstra a Tabela 2. Neste critério, os autores Skittou *et al.* (2020) destacam que em seus experimentos o algoritmo KNN com critérios de 15 vizinhos, conforme demonstra a Tabela 2, teve melhor desempenho para o primeiro conjunto de dados com 69% de precisão. Mas quando aumentou-se o número de casos (segundo conjunto de dados), os três algoritmos (SVM, RF e *Logistic Regression*) apresentam bons resultados, sendo que a taxa de precisão sobre o conjunto de teste foi superior a 92%, com destaque para o algoritmo SVM que teve uma precisão de 93%. E continuando, os autores Skittou *et al.* (2020) aplicaram o algoritmo de regressão logística para comparar, assim o algoritmo teve desempenho inferior no nível do primeiro conjunto de dados com precisão de 62%, de modo que foi superado pelos três algoritmos anteriores. Mas quando aplicado ao segundo conjunto de dados, seu comportamento é mais notável e sua precisão atingiu 93%.

Tabela 2. Precisão da Pontuação.

Conjunto de Dados (casos)	Treinamento/Teste divididos	Algoritmos			
		SVM	KNN	RF	*Logistic Regression*
158	110/48	0.65	0.69	0.67	0.62
340	237/102	0.93	0.92	0.92	0.93

Fonte: Skittou *et al.* (2020).
Nota: Ilustração alterada pelo autor.

Para Nugrahaeni e Mutijarsa (2016), o objetivo do SVM é produzir um modelo (com base nos dados de treinamento) que preveja os valores de destino dos dados de teste, dados apenas os atributos dos dados de teste. Dado um conjunto de treinamento de pares instância-rótulo (xi, yi), i = 1,...,l onde xi ∈ Rn e y ∈ {1, −1}l, o SVM exige a solução do seguinte problema de otimização, conforme equação (1):

$$\min_{w,b,\xi} = \frac{1}{2}w^T w + C\sum_{i=1}^{l} \dot{\xi} i \qquad (1)$$

Então $yi(w^t\ \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0.$

Os vetores de treinamento "xi" são mapeados em um espaço dimensional superior (talvez infinito) pela função φ. O SVM encontra um hiperplano de separação linear com a margem máxima neste espaço dimensional superior. C > 0 é o parâmetro de penalidade do termo de erro. Além disso, K (xi, xj) ≡ ϕ (xi) $T\phi$ (xj) é chamada de função de *kernel*, destacando os quatros *kernels* básicos, conforme descritos na Tabela 3.

Tabela 3. Fórmulas da distância para variáveis contínuas.

Métricas	Expressões Matemáticas
Linear	$K(x_i, x_j) = x^T x_j$
Polynomial	$K(x\ , x\ ) = (\gamma x^T x\ + r)^d, \gamma > 0$
*Radial basis function (*RBF*)*	$K(x\ , x\ ) = exp\ (-\gamma\|x\ - x\ \|^2), \gamma > 0$
Sigmoid	$K(x_i, x_j) = tanh(\gamma x^T x_j + r)$

Fonte: Nugrahaeni e Mutijarsa (2016).

Chen e Kégl (2010) fizeram o uso do algoritmo AdaBoost para reconhecimento de padrões de manuscritos. Nesse processo do AdaBoost, todos os pesos são inicialmente iguais, mas em cada rodada, o aprendizado fraco retorna uma hipótese, e o peso de todos os exemplos classificados erroneamente por aquelas hipóteses são aumentadas. Portanto, o aprendizado fraco concentrará nas amostras difíceis do conjunto de treinamento, e ao final é feita uma combinação das hipóteses de todas as rodadas, e as hipóteses com menor erro de classificação têm maior peso. De acordo com Chen e Kégl (2010), o algoritmo pode ser sumarizado da seguinte maneira:

Dado:

$(x_1, y_1), \ldots\ldots, (x_m, y_m)\ onde\ x_i \epsilon X, y_i \epsilon Y = \{-1, +1\}$

Inicializando $D_{1\ (i)=\frac{1}{m}}$

Laço t = 1, ...,T, repetir

1. Treina aprendizado fraco usando distribuição D_t.
2. Obtém hipótese para o mais fraco: $h_t = X \rightarrow \{-1, +1\}$ com erro $\epsilon_t = Pr_i \sim D_t[h_t(x_t)\ \neq\ y_i]$.

3. Escolha $\alpha_{t=\frac{1}{2}In(1-\varepsilon_t/\varepsilon_t)}$
4. Atualiza $D_t + 1(i) = D_t e^{-\frac{x_t y_i h_t(x_i)}{Z_t}}$

Onde z_t é a normalização do fator. Produzir a hipótese no final para a votação majoritária ponderada as hipóteses *T* fracas.

$$H(x) = sign\,(\sum_t^T = 1\; x_t x h_t(x))$$

4. METODOLOGIA

Nesta Seção são utilizados os algoritmos CNN, SVM, RF, KNN *e* AdaBoost. Para isso, a primeira aplicação no que tange ao pré-processamento foram por meio da geração das imagens NDVI a partir da obtenção dos mapas através do satélite Sentinel-2 (regiões em que concentrou o trabalho), que é dado pelo cálculo da subtração das bandas (8 e 4) divido pela soma das bandas (4 e 8) de cada *pixel*, descrita pela equação (2) (ROUSE *et al.*, 1973). Os valores NDVI variam entre –1 que representa terreno sem cobertura vegetal, e os valores próximos +1, os quais representam maior vigor da vegetação (ROUSE *et al.*, 1973). As características das bandas obtidas pelo satélite Sentinel-2 estão descritas na Tabela 4, sendo que especificamente as bandas "4" e "8", quando combinadas com RGB (*Red, Green* e *Blue*) permite a obtenção das imagens NDVI, conforme equação (2).

Tabela 4. Faixas espectrais do sensor do satélite Sentinel-2 de acordo com resolução espacial.

Resolução	Número Da Banda	Nome da Banda	Comprimento da Onda Central (nanômetro)	Combinações de Bandas
10m	B02	*Blue* (azul)	490	Cor Verdadeira
	B03	*Green* (verde)	560	*RGB* 04/03/02
	B04	*Red* (vermelho)	665	Falsa Cor 1 e 2
	B08	*NIR* (infravermelho próximo)	842	*RGB* 08/04/03 e 04/08/03

Fonte: http://www.engesat.com.br/sentinel-2/.

$$\text{NDVI} = \frac{NIR-Red}{NIR+Red} \quad (2)$$

Onde:
NDVI= *Normalize Diference Vegetativo Index;*
NIR = *Reflectance of Near InfraRed* (B08); e,
R= *Reflectance of Red* (B04).

Para estabelecer uma métrica para que melhores resultados fossem obtidos a partir dos algoritmos, foram realizadas avaliações no âmbito de cálculos matemáticos e estatísticos, sendo que o intuito foi determinar o melhor algoritmo. Sendo assim, foram utilizadas as seguintes fórmulas: MAE (*Mean Absolute Error*) que é o módulo da diferença entre os valores reais e os valores estimados ou preditos pelo algoritmo, descrita na equação (3); o MSE (*Mean Square Error*) que é a diferença média quadrática entre os valores estimados ou preditos pelo algoritmo sobre os valores reais, descrita na equação (4); a fórmula para o MAPE (*Mean Porcentage Error*) que é o percentual da diferença entre os valores estimados ou preditos pelo algoritmo sobre os valores reais, descrita na equação (5); a MA (Média Aritmética) que é obtida pela divisão da soma das observações pelo número das amostras, descrita na equação (6); e, por fim, a fórmula para o DP *(*Desvio Padrão) que indica a dispersão dos dados dentro de uma amostra com relação à média**,** ou seja, o quanto ela é uniforme, descrita na equação (7).

$$\text{MAE} = \frac{1}{2}\sum_{n=1}^{n} |\, y_j - \hat{y}_j| \quad (3)$$

Onde:
MAE = *Mean Absolute Error*;
y = são os dados;
ŷ = são os ajustes; e,
n = número de elementos.

$$\text{MSE} = \frac{1}{2}\sum_{n=1}^{n}(y-\hat{y})^2 \quad (4)$$

Onde:
MSE = *Mean Square Error;*
y = são os dados reais;
ŷ = são os ajustes; e,
n = número de elementos.

$$\text{MAPE} = \frac{1}{2}\sum_{n=1}^{n}\frac{|\,y-\hat{y}|}{n} \quad (5)$$

Onde:
MAPE = *Mean Porcentage Error;*
y = são os dados;
ŷ = são os ajustes; e,
n = número de elementos.

$$\mathrm{MA} = \frac{x1+x2\ldots+xn}{n} \quad (6)$$

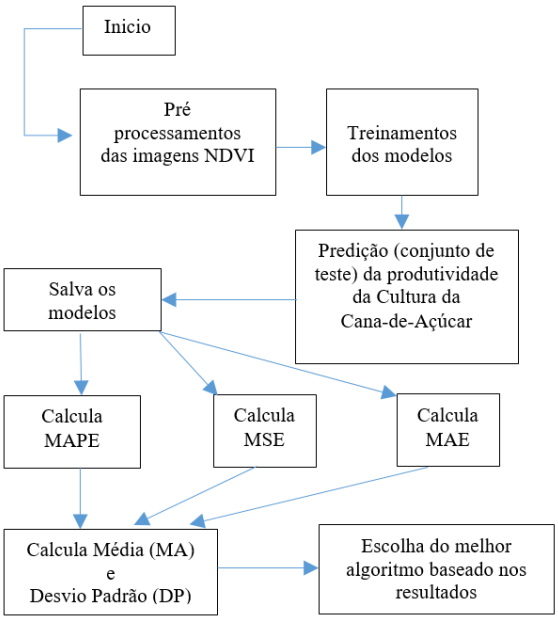
Onde:
MA = Média Aritmética;
x1+x2…xn = valores dos dados; e,
n = número de elementos.

$$\mathrm{DP} = \sqrt{\frac{\sum_{i=1}^{n}(xi - MA)^2}{n}} \quad (7)$$

Onde:
DP = é o Desvio Padrão;
xi = é um valor no conjunto de dados na posição i;
MA = é a Média Aritmética dos dados do conjunto, conforme equação (6); e,
n = é a quantidade total de elementos do conjunto.

Os procedimentos adotados, são descritos pelo fluxo do processo conforme ilustra a Figura 10, iniciando com os pré-processamentos das imagens NDVI (dada pela construção e extração dessas imagens), na sequência os modelos são treinados, posteriormente, o processo de predição é aplicado e, logo após, gravam-se os pesos, e conseguintes realizam os cálculos através das fórmulas (MAPE, MSE e MAE). Por último, calcula-se (MA e DP) de acordo com as equações (3, 4, 5, 6 e 7) respectivamente.

Figura 10. Etapas associadas aos métodos e às equações aplicadas.



Fonte: Os autores.

4.1. Configurações das arquiteturas

Os algoritmos seguiram as configurações, conforme descritos na Tabela 5, sendo que a coluna "Parâmetros Específicos" correspondem a cada uma das funções, e a coluna "Valores dos Parâmetros" é o valor de entrada para a função dos algoritmos. Foram executadas 20 repetições para cada um dos algoritmos.

Tabela 5. Configurações dos algoritmos e seus parâmetros específicos.

Algoritmos	Parâmetros Específicos	Valores dos Parâmetros
CNN	*Epochs*	300 épocas
KNN	*Neighbor* (*k1..kn*)	*Ki*
AdaBoost	*Estimators*	50
SVM	*Kernel*	*RBF*
RF	*Estimators*	50

Fonte: Os autores.

A CNN é construída sobre camadas e pelo agrupamento, de modo a extrair o mapa das características, dessa forma, ficam totalmente conectadas e, na sequência, obtêm-se os classificadores. O mapa de características da imagem é obtido pela camada de convolução através dos núcleos de convolução (*kernels*); e, nesse momento, é aplicada sobre as regiões da imagem, multiplicando os pesos dos *pixels* (conforme Figura 3). Ao finalizar, obtém-se uma nova matriz denominada mapa de características. Considerando que a proporcionalidade ainda é grande, aplica-se a camada de *Pooling* que reduz ainda mais a dimensionalidade do mapa de características. Esse processo é realizado em uma região da matriz de onde se extrai o maior valor.

Posteriormente é apresentada a configuração da arquitetura utilizada para o algoritmo da CNN, conforme a ilustração da Tabela 6, sendo que os valores para os parâmetros foram definidos, tais como: *bias* "16, 32, 128, 256, 128, 64, 32, 16 e 64", *kernels* "3, 5, 8, 10, 8, 6, 5 e 2", *maxpooling pool size* para todas as camadas "1" e "*dropout*" calibrados em "0.001" em todas as camadas. Sendo que até a penúltima camada de ativação foi utilizada "ReLU" (*Rectifying Linear Unit).* Por último, conectou-se a camada de saída "*Linear*" que possui por finalidade a predição.

Para obter o resultado "*Param#*" conforme ilustra a tabela 6, o cálculo fica da seguinte forma: conv1d_1 na coluna "*Output Shape*" (156 *input*, 156 *first layear*, 16 *bias*, *kernel*

size = 3, tal que, 156 * 156 / 3 + 16 = 8.128), conv1d_2 (16 *bias first layear*, 32 *bias second layear*, *kernel size* = 5, tal que 16 * 32 *5 + 32 = 2.592), e assim sucessivamente.

Tabela 6. Arquitetura utilizada para CNN.

Layer (type)	***Output Shape***	***Param#***
conv1d_1 (Conv1D)	(None, 156, 16)	8128
max_pooling1d_1	(None, 156, 16)	0
dropout_1	(None, 156, 16)	0
conv1d_2 (Conv1D)	(None, 152, 32)	2592
max_pooling1d_2	(None, 152, 32)	0
dropout_2 (Dropout)	(None, 152, 32)	0
conv1d_3 (Conv1D)	(None, 145, 128)	32896
max_pooling1d_3	(None, 145, 128)	0
dropout_3 (Dropout)	(None, 145, 128)	0
conv1d_4 (Conv1D)	(None, 136, 256)	327936
max_pooling1d_4	(None, 136, 256)	0
dropout_4	(None, 136, 256)	0
conv1d_5	(None, 129, 128)	262272
max_pooling1d_5	(None, 129, 128)	0
dropout_5	(None, 129, 128)	0
conv1d_6	(None, 124, 64)	49216
max_pooling1d_6	(None, 124, 64)	0
dropout_6	(None, 124, 64)	0
conv1d_7	(None, 120, 32)	10272
max_pooling1d_7	(None, 120, 32)	0
dropout_7	(None, 120, 32)	0
conv1d_8	(None, 119, 16)	1040
max_pooling1d_8	(None, 119, 16)	0
dropout_8	(None, 119, 16)	0
flatten_1	(None, 1904)	0
dense_1	(None, 64)	121920
dense_2	(None, 1)	65
Linear		
Total *params*: 816.337		

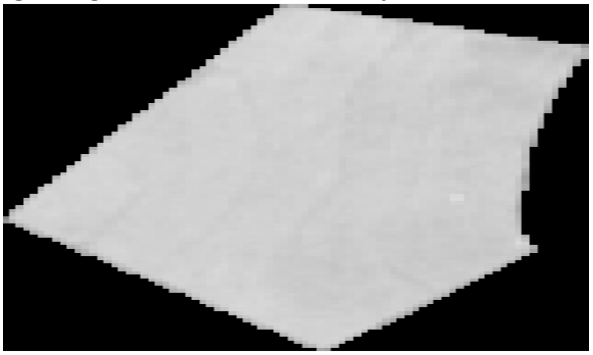
Fonte: Os autores.

4.2. Organização dos dados

Os dados para a geração das imagens foram organizados a partir dos dados dos polígonos disponibilizados pela empresa fornecedora dos dados (pontos das coordenadas) que são as informações dos talhões (georreferenciamento das unidades agrárias), conforme exemplo da Figura 11. Através do *software QGIS*, foram aplicados aos mapas a equação (2), dando origem às imagens NDVI. Foram utilizadas 5.583 imagens NDVI que compõem o conjunto de treinamento, e 1.116 imagens NDVI para o conjunto de teste, que representam 80% e 20% respectivamente. O conjunto de teste foi definido aleatoriamente.

As produtividades reais da cultura da Cana-de-Açúcar foram concedidas pela empresa fornecedora dos dados a partir da pesagem do produto (balança), assim como os valores estimados foram disponibilizados através de uma planilha eletrônica com extensão CSV (*Comma Separated Values*). A planilha possui 6.699 linhas (apenas dados utilizados) e está disponível através do conjunto de dados "*Datasets*". O *link* do conjunto de dados está descrito na Seção das Referências. Os dados reais efetivamente produzidos estão dispostos na coluna "PROD_REAL" e as estimativas estão dispostas na coluna "PROD_PLAN". Os demais campos foram utilizados para criar uma borda na imagem com as características baseadas em HASH, e estão sumarizadas na Tabela 7.

Figura 11 - Exemplo da imagem de uma unidade agrária georreferenciado do conjunto de dados.



Fonte: Os autores. (conjunto de dados).

Tabela 7. Detalhes dos campos utilizados na planilha CSV pela empresa fornecedora dos dados.

Nome	Detalhes
COD_TALHAO	Nomeclatura do número do talhão (espaço geográfico)
BALANCA	Nome da balança de pesagem
PROD_PLAN	Produtividade Estimada
PROD_REAL	Produtividade Real
DESC_VARIEDADE	Descreve as variedades da Cana-de-Açúcar
DESC_TIPOAMBIENTE	Nomenclatura própria para descrever o tipo de Solo
DESC_CATEGORIA	Define as estações do Plantio
IDADE_MESES	Idade da Cana-de-Açúcar em meses
DESC_TIPOCANA	Tipo da Cana
ESC_POTSOLO	Potencial do Solo, por exemplo 5x80 (5 cortes x 80 t)
ESPACAMENTO	Espaçamento das entrelinhas (plantações)
DESC_TIPOFERTIRRIGACAO	Tipo da fertilização do Solo
DESC_PLANTIO	Descreve o modo de Plantio
TEORK	Teor de Potássio presentes no Solo
DESC_GRPMATURACAO	Tipo do grupo de maturação
DESC_SITUACAO	A situação que encontra o talhão
DESC_TIPOSOLO	Tipo do Solo
NR_CORTE	Número de cortes que a Cana-de-Açúcar já tenha sofrido
DISTANCIA	É a distância média para transporte entre a lavoura e a área industrial (onde Realiza o processamento da Cana-de-Açúcar colhida)
MÊS	Entrada da Cana-de-Açúcar na balança para pesagem da indústria

Fonte: Os autores.

4.3. As imagens

Os mapas compreendem os meses de fevereiro a dezembro do ano de 2020, sem correlacionar ao clima ou estações do ano, observando as melhores imagens disponíveis pelo satélite Sentinel-2, de modo que foram transformadas e, através do *software QGIS,* foram extraídas, gerando as imagens NDVI.

5. MATERIAIS E MÉTODOS

O *hardware* utilizado para o conjunto de treinamentos dos modelos para o conjunto de teste foi um computador Intel® *Core™ i7*-7500*U CPU @* 2.7*Ghz,* com 16*GB* de *RAM,* placa de vídeo *NVIDIA GeForce 940MX e HD Graphics 620*. O sistema operacional utilizado foi *Windows* 10 *Pro* 64 *bits*. Para pré-processamento das imagens, foi utilizado o *software* livre de sistema de informação geográfica *QGIS* versão 3.14.16 e, para o desenvolvimento dos algoritmos, foi utilizado o *software Spyder* versão 4.1.5 do pacote *software Anaconda* versão 1.9.12, sendo ambos com licenças gratuitas. No caso do *Spyder,* as bibliotecas utilizadas foram: *Numpy, Tensorflow, Keras, Pandas,* entre outros.

As imagens foram obtidas através do site do Sentinel-2 v3.0.82, onde foram realizados os *downloads* referentes aos mapas através do endereço *https://apps.sentinel-hub.com/eo-browser*. As faixas espectrais do satélite Sentinel-2, utilizadas de acordo com resolução espacial, foram de 10 m por *pixel* com a opção de configuração pelo nível "LA2" com a correção atmosférica, e o filtro do parâmetro para cobertura de nuvens aplicado foi inferior a 10%. Para extração das imagens, foram utilizados os polígonos dos talhões (mapeamento georreferenciado das regiões específicas da cultura da Cana-de-Açúcar pela empresa fornecedora dos dados), os quais foram aplicados aos mapas mensais obtidos através de *downloads*, e através do *software QGIS* foram extraídas as imagens NDVI. Para os conjuntos de treinamento e teste, as resoluções das imagens foram padronizadas em 169 x 157 *pixels*. Sendo que as imagens utilizadas, assim como os pesos dos treinamentos e teste, foram tratadas e trabalhadas pelo próprio autor e estão disponíveis através do conjunto de dados (*Datasets*).

Na elaboração dos mapas de NDVI, foram utilizadas as bandas do vermelho (R) e infravermelho próximo (NIR), as quais estão indicadas pelas bandas "4" e "8",

respectivamente descritas na Tabela 4, e aplicada a fórmula, conforme equação (2). Foram extraídas 5.583 imagens para compor o conjunto de dados de treinamento, e 1.116 imagens para conjunto de teste, que correspondem respectivamente a 80% para treinamento e 20% para conjunto de teste, ambos selecionados aleatoriamente.

6. RESULTADOS ALCANÇADOS

Nesta Seção destacam-se os resultados obtidos através dos processos sobre o conjunto de dados para o treinamento e teste realizados pelos algoritmos.

O custo computacional para treinar é um fator importante, conforme mostra a Tabela 8, observa-se a coluna "Tempo" que descreve "horas, minutos, segundos e milissegundos", tanto para conjunto de treinamento quanto para o conjunto de teste, que representam os custos em tempos gastos para cada um dos algoritmos respectivamente. Para apresentar esses resultados, foi utilizada a fórmula MA, descrita na equação (6), considerando os custos totais (tempo) para as séries.

Tabela 8. Tempo médio (aplicando a equação 6) dos algoritmos para o conjunto de treinamento e teste (20 repetições).

	Tempo (hora/minutos/segundos/milis-segundos)
Algoritmo	Treinamento ± Teste
CNN	8:06:44.00 ± 0:00:02.00
KNN	0:00:32.00 ± 0:00:53.00
AdaBoost	0:05:22.00 ± 0:00:01.00
RF	0:09:45.00 ± 0:00:00.06
SVM	0:17:42.00 ± 0:03:28.00

Fonte: Os autores..

6.1. Resultados atingidos

Os dados estimados para a produtividade foram disponibilizados pela empresa fornecedora dos dados (*Datasets*) e tomados como base para comparar com os dados preditos pelos algoritmos. A Tabela 9 destaca os resultados em toneladas (t) para as estimativas, aplicando as fórmulas para MA e DP, conforme descrito pelas equações (6 e 7), respectivamente, resultantes das fórmulas MAE, MSE e MAPE, descritas nas equações (3, 4 e 5).

Tabela 9. Resultados das equações MA (Média Aritmética) e DP (Desvio Padrão), a partir das estimativas, baseados nos resultados das fórmulas MAPE, MSE e MAE respectivamente.

	Estimativas
Fórmulas Matemáticas	**MA (**Média Aritmética**) ± DP (**Desvio Padrão**)**
MAPE	31,15 ± 4,94
MSE	987,09 ± 34,71
MAE	21,59 ± 3,97

Fonte: Os autores.

Para demonstrar os resultados dos modelos preditivos dos algoritmos (CNN, KNN, RF, AdaBoost), a Tabela 10 apresenta os resultados. O resultado MA (Média Aritmética) ± DP (Desvio Padrão) descritas nas equações (6 e 7) são os resultados de todas as repetições (séries) de execuções dos algoritmos sobre os resultados das fórmulas MAE, MSE e MAPE, descritas nas equações (3, 4 e 5).

Tabela 10. Resultados das predições dos algoritmos sobre os resultados das fórmulas (MAE, MSE e MAPE*)* descritas nas equações (3, 4 e 5) atribuindo aos resultados para as fórmulas "MA e DP", descritas nas equações (6 e 7) respectivamente.

		Preditivos
Algoritmos	**Fórmulas Matemáticas**	**MA (**Média Aritmética**) ± DP (**Desvio Padrão**)**
CNN	MAPE	40,07 ± 5,64
	MSE	1121,26 ± 34,61
	MAE	25,11 ± 4,01
NN	MAPE	33,35 ± 5,23
	MSE	934,43 ± 33,07
	MAE	21,49 ± 3,89
RF	MAPE	**31,51 ± 5,14**
	MSE	**791,27 ± 30,71**
	MAE	**19,72 ± 3,73**
AdaBoost	MAPE	48,90 ± 6,49
	MSE	1165,36 ± 34,00
	MAE	26,78 ± 4,03
SVM	MAPE	36,76 ± 5,38
	MSE	1033,02 ± 33,91
	MAE	23,62 ± 3,95

Fonte: Os autores.

6.2. Validações

Foram aplicados resultados parciais, sendo catalogadas nove imagens aleatórias do conjunto de teste, conforme Tabela 11, a partir do algoritmo RF disposto na vigésima repetição da execução. Os resultados obtidos pelo algoritmo RF das diferenças entre os valores "Estimados" e o "Real Produzido" em (t), e entre os valores "Preditos" e o "Real Produzido" em (t), estão disponíveis na Tabela 11. Apresentando os resultados (seguindo os números das imagens), conforme descrito na Tabela 11, destacam-se os valores preditos, sendo que as três primeiras imagens (a), (b) e (c) representam uma assertividade com diferença inferior de 1,0 (t), e as três imagens (d), (e) e (f) conseguintes, estão na faixa intermediária de 83,09 (t) (MA) para a série completa (20 repetições), e por último, são três imagens (g), (h) e (i) resultantes que apresentaram diferenças para as séries discrepantes (valor absoluto), ou seja, superiores a -137,00 (t). A base para a média intermediária está descrita na Tabela 12 para o valor Predito.

$$\mathrm{DRP} = Valor - RP \quad (8)$$

Onde:
DRP = Diferença entre Real Produzido (em toneladas);
Valor = Valor em toneladas da amostra; e,
RP = Real Produzido (em toneladas).

Tabela 11. Resultados pelo algoritmo RF (*Random Forest*) das diferenças entre os valores "Estimados" e o "Real Produzido" em (t), e entre os valores "Preditos" e o "Real Produzido" em (t), de acordo equação (8).

Número da Imagem	Valores Expressos em Toneladas (t)			Resultado da Subtração com o Valor Real Produzido (t)	
	Real Produzido	Estimados	Preditos	Estimados	Preditos
A	95,94	78,00	96,04	-17,94	0,10
B	110,48	125,00	110,59	14,52	0,11
C	84,47	75,00	84,64	-9,47	0,17
D	21,88	90,00	103,86	68,12	81,98
E	6,26	80,00	90,32	73,74	84,06
F	25,11	95,00	111,00	69,89	85,89
G	264,78	60,00	78,34	-204,78	-186,44
H	253,65	82,00	99,95	-171,65	-153,70
I	223,90	100,00	86,60	-123,90	-137,30

Fonte: Os autores.

Na Tabela 12, é apresentado o resultado para EM (Erro Médio) conforme equação (9) para as estimativas e valores preditos com o algoritmo RF, com intuito de validar a variação para os termos. A MA é aplicada sobre 20 repetições (série completa), sendo assim, os resultados do algoritmo RF, para os valores preditos, obtiveram 0,012%, e para as estimativas obtiveram-se -8,80%.

$$\mathrm{EM} = \left(\frac{MA-RP}{RP}\right) * 100 \quad (9)$$

Onde:
EM = Erro Médio;
MA = Média Aritmética; e,
RP = Real Produzido.

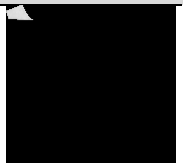
Tabela 12. Resultado do algoritmo RF para o Erro Médio (%) sobre o valor Real Produzido.

	Valores Expressos (t)	
Real Produzido (média/t)	**Estimado (média/t)**	**Predito (média/t)**
83,08	75,78	83,09
Erro Médio (%)	-8,80%	0,012%

Fonte: Os autores..

Na Figura 12, foram selecionadas as imagens aleatoriamente (georreferenciamento da unidade agrária), e a partir da predição pelo algoritmo RF, considerando a vigésima repetição e seus respectivos valores (valores estimados subtraindo dos valores reais e os preditos pelo algoritmo subtraindo dos valores reais). Sendo que para as imagens (a), (b) e (c), os resultados ficaram abaixo de 1.0 (t), em (d), (e) e (f), os valores preditos resultaram próxima da média (MA) para as séries dos resultados preditos pelo algoritmo RF que foi de 83,09 (t), e por último (g), (h) e (i) em que atingiram diferenças discrepantes para as séries da predição (predição com valores negativos), superiores a -137,00 (t).

Figura 12. Predição baseada no aprendizado de máquina pelo algoritmo RF. Os valores descritos R (Real), E (Estimado) e P (Predito).

A	B	C
		
R.: 95,94 (t) E.: 78,00 (t) P.: 96,04 (t)	R.:110,48 (t) E.:125,00 (t) P.: 110,59 (t)	R.: 84,47 (t) E.: 75,00 (t) P.: 84,64 (t)
D	**E**	**F**
		
R.: 21,88 (t) E.: 90,00 (t) P.:103,86 (t)	R.: 6,26 (t) E.: 80,00 (t) P.:90,32 (t)	R.: 25,11 (t) E.: 95,00 (t) P.:111,00 (t)
G	**G**	**I**
		
R.: 264,78 (t) E.: 60,00 (t) P.:78,34 (t)	R.:253,65 (t) E.: 82,00 (t) P.:99,95 (t)	R.:223,90 (t) E.:100,00 (t) P.:86,60 (t)

Fonte: Os autores.

As diferenças em toneladas (t) entre os valores preditos pelo algoritmo e valores reais produzidos, tiveram os seguintes resultados: (a) 0,10 (b) 0,11 (c) 0,17 (d) 81,98 (e) 84,06 (f) 85,89 (g) -186,44 (h) -153,70 (i) -137,30.

7. CONCLUSÃO

A aplicação de diferentes tecnologias na agricultura tem alcançado enormes resultados, já que tais instrumentos possibilitam a geração de dados, os quais após os tratamentos, resultam em informações relevantes que auxiliam na tomada de decisão. Deve-se considerar que tais dados ainda são tidos como desafiador diante da complexidade dos processos para atingir os resultados esperados.

A metodologia aplicada nesse projeto, delineou-se utilizando das técnicas de processamento de imagens NDVI e sustentados pelos algoritmos de ML, conforme ilustram os comparativos nas Tabelas 9 e 10, respectivamente. O uso de ML aplicada à predição da produtividade na cultura da Cana-de-Açúcar mostrou-se eficiente, se comparado às estimativas. De modo a destacar, o algoritmo RF apresentou ser mais eficiente se comparado com os demais algoritmos (CNN, KNN, AdaBoost e SVM) no que tange ao cenário atribuído às estimativas. O algoritmo RF teve os resultados para fórmula do DP de 30,71 (t) sobre o MSE, e de 3,73 (t) sobre o MAE, sendo que ao ser comparado com as estimativas, os resultados para o DP sobre o MSE foi de 34,71(t), e de 3,97 (t) sobre o MAE. Considerando a fórmula DP sobre o MAPE, teve-se um resultado superior se comparado com as estimativas; neste caso, para o DP, obteve-se 5,14 (%), sendo que as estimativas resultaram para o DP em 4,94 (%), de modo que foi descartada a fórmula MAPE.

O Erro Médio para o valor estimado foi de -8,80% se comparado ao valor médio com relação ao real produzido, e para o valor predito foi de 0,012% se comparado ao valor médio com relação ao real produzido. Fica evidenciado neste trabalho que o algoritmo RF foi mais eficiente se comparado às estimativas.

Futuros trabalhos podem ser realizados aplicando novas técnicas de DL (*Deep Learning*), no entanto, acrescentando períodos anuais com mais imagens, com o intuito de melhorar a precisão dos resultados.

REFERÊNCIAS

ABBAS, M. A. Improving deep learning performance using random forest HTM cortical learning algorithm. *In*: INTERNATIONAL WORKSHOP ON DEEP AND REPRESENTATION LEARNING (IWDRL), 1., 2018. Cairo. **Anais** […]. Cairo, Egito, 2018, p. 13-18. . https://doi.org/10.1109/IWDRL.2018.8358209

ASADI, M.; POURHOSSEIN, K. Locating Renewable Energy Generators Using K-Nearest Neighbors (KNN) Algorithm, **Iranian.** *In:* CONFERENCE ON RENEWABLE ENERGY & DISTRIBUTED GENERATION (ICREDG), 2019, [S.l.], . 2019. p. 1-6https://doi.org/10.1109/ICREDG47187.2019.190179

BRASIL. Decreto nº 76.593, de 14 de novembro de 1975. Institui o Programa Nacional do Álcool e dá outras Providências, **Diário Oficial da União**, Brasília, DF, Nov, 1975.

CHAROEN-UNG, P.; MITTRAPIYANURUK, P. Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques.

In: INTERNATIONAL JOINT CONFERENCE ON COMPUTER SCIENCE AND SOFTWARE ENGINEERING (JCSSE), 15., 2018. Nakhonpathom. THA, 2018. p. 1-6. https://doi.org/10.1109/JCSSE.2018.8457391

CHEN, G. Y.; KÉGL, B. Invariant pattern recognition using contourlets and AdaBoost, **Pattern Recognition**, v. 43, n. 3, , p. 579-583, 2010. . https://doi.org/10.1016/j.patcog.2009.08.020

COBEÑA CEVALLOS, J. P.; ATIENCIA VILLAGOMEZ, J. M.; ANDRYSHCHENKO, I. S. Convolutional Neural Network in the Recognition of Spatial Images of Sugarcane Crops in the Troncal Region of the Coast of Ecuador. *In:* INTERNATIONAL SYMPOSIUM "INTELLIGENT SYSTEMS" (INTELS'18), 13., 2019. Moscou, Russia. Anais […]. Moscou, 2019. p. 757-763https://doi.org/10.1016/j.procs.2019.02.001

DATASETS de imagens usadas no treinamento e testes das redes para predição da produtividade na cultura Cana-da-açúcar. Disponível em: https://drive.google.com/drive/folders/1sf2VSJGsQAJecCrYFVx4W-xu1XTJ3-8d?usp=sharing. Acesso em: 10 fev. 2021.

DEN BESTEN, N. I.; KASSING, R. C.; MUCHANGA, E.; EARNSHAW, C.; de JEU, R. A. M.; KARIMI, P.; van der ZAAG, P. A novel approach to the use of earth observation to estimate daily evaporation in a sugarcane plantation in Xinavane, Mozambique. **Physics and Chemistry of the Earth, Parts A/B/C**, 102940, 2020. https://doi.org/10.1016/j.pce.2020.102940

DO VALLE GONÇALVES, R. R.; ZULLO, J.; ROMANI, L. A. S.; do AMARAL, B. F.; SOUSA, E. P. M. Agricultural monitoring using clustering techniques on satellite image time series of low spatial resolution. *In*: INTERNATIONAL WORKSHOP ON THE ANALYSIS OF MULTITEMPORAL REMOTE SENSING IMAGES (MULTISTEP), 9., 2017, Bruges, BEL. **Anais** […]. Bruges, 2017, p. 1-4.https://doi.org/10.1109/Multi-Temp.2017.8035234

DUFT, D. G.; PICOLI, M. C. A. Uso de imagens do sensor modis para identificação da seca na cana-de-açúcar através de índices espectrais. **Scientia agraria**, Curitiba,v. 19, n. 1, p. 52-63, jan./mar. 2018.. https://doi.org/10.5380/rsa.v19i1.54005

EBADI, A.; GAUTHIER, Y.; TREMBLAY, S.; PAUL, P. How can Automated Machine Learning Help Business Data Science Teams? *In*: IEEE INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND APPLICATIONS (ICMLA). 18., Boca Raton, FL, **Anais** […]. Raton, FL, 2019, p. 1186-1191.https://doi.org/10.1109/ICMLA.2019.00196

FERNANDES, J. L.; EBECKEN, N. F. F.; ESQUERDO, J. C. D. M. Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. **International Journal of Remote Sensing**, v. 38, n. 16, p. 4631–4644, maio, 2017.. https://doi.org/10.1080/01431161.2017.1325531

GEETHARAMANI, G.; ARUN PANDIAN, J. Identification of plant leaf diseases using a nine-layer deep convolutional neural network. **Computers & Electrical Engineering**, p. 323–338, 2019. https://doi.org/10.1016/j.compeleceng.2019.08.010

GHOSAL, S.; SARKAR, K. Rice Leaf Diseases Classification Using CNN With Transfer Learning, **2020 IEEE Calcutta Conference (CALCON)**, Kolkata, India, p. 230-236, 2020. https://doi.org/10.1109/CALCON49167.2020.9106423

HOSSAIN, E.; HOSSAIN, M. F.; RAHAMAN, M. A. A Color and Texture Based Approach for the Detection and Classification of Plant Leaf Disease Using KNN Classifier, INTERNATIONAL CONFERENCE ON ELECTRICAL, COMPUTER AND COMMUNICATION ENGINEERING (ECCE**)**, 2019. Cox'sBazar. Anais […]. Cox'sBazar, Bangladesh, 2019 , p. 1-6. https://doi.org/10.1109/ECACE.2019.8679247

HU, G.; YIN, C.; WAN, M.; ZHANG, Y.; FANG, Y. Recognition of diseased Pinus trees in UAV images using deep learning and AdaBoost classifier. **Biosystems Engineering**, v. 194, p. 138–151, 2020. https://doi.org/10.1016/j.biosystemseng.2020.03.021

KAI, P. M.; COSTA, R. M.; OLIVEIRA, B. M.; FERNANDES, D. S. A.; FELIX, J.; SOARES, F. Discrimination of Sugarcane Varieties by Remote Sensing: A Review of Literature.**In:** IEEE ANNUAL

COMPUTERS, SOFTWARE, AND APPLICATIONS CONFERENCE (COMPSAC), 44., Madrid, Spain, p. 1212-1217, 2020. https://doi.org/10.1109/COMPSAC48688.2020.00-91

KHAN, W. *et al.* On the Performance of Temporal Stacking and Vegetation Indices for Detection and Estimation of Tobacco Crop. **IEEE Access,** v. 8, p. 103020-103033, 2020. Acesso em: 28 mar. 2021. https://doi.org/10.1109/ACCESS.2020.2998079

KHETKEEREE, S. Infrared Normalized Difference Vegetation Index for Sentinel-2A Imagery. *In*: INTERNATIONAL CONFERENCE ON ELECTRICAL ENGINEERING/ELECTRONICS, COMPUTER, TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY (ECTI-CON)*,* 2020, [S.l.], 2020. p. 405-408. https://doi.org/10.1109/ECTI-CON49241.2020.9158105

LANEVE, G.; MARZIALETTI, P.; LUCIANI, R.; FUSILLI, L.; MULIANGA, B. Sugarcane biomass estimate based on sar imagery: A radar systems comparison. *In*: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS), 2017, Fort Worth, TX. **Anais** […].Fort Worth, TX 2017, p. 5834-5837.. Disponível em: https://ieeexplore.ieee.org/document/8128335. Acesso em: 13 abr. 2020.

LUCIANO, A. C. S.; PICOLI, M. C. A.; ROCHA, J. V.; FRANCO, H. C. J.; SANCHES, G. M.; LEAL, M. R. L. V.; le MAIRE, G. Generalized space-time classifiers for monitoring sugarcane areas in Brazil. **Remote Sens. Environ**, Campinas, v. 215, p. 438–451, 2018. https://doi.org/10.1016/j.rse.2018.06.017. Acesso em: 28 mar. 2020. https://doi.org/10.1016/j.rse.2018.06.017

LUCIANO, A. C. S.; PICOLI, M. C. A.; ROCHA, J. V.; DUFT, D. G.; LAMPARELLI, R. A. C.; LEAL, M. R. L. V.; LE MAIRE, G. A generalized space-time OBIA classification scheme to map sugarcane areas at regional scale, using Landsat images time-series and the random forest algorithm. **International Journal of Applied Earth Observation and Geoinformation**, v. 80, p. 127–136, abr. 2019a. https://doi.org/10.1016/j.jag.2019.04.013

LUCIANO, A. C. dos S.; DUFT, D. G.; PICOLI, M. C. A.; ROCHA, J. V.; Le MAIRE, G. Estimativa da produtividade de cana-de-açúcar utilizando imagens landsat e random forest. *In:* SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 19., 2019, Santos, SP. **Anais** […]. Santos, 2019b. Disponível em: http://marte2.sid.inpe.br/attachment.cgi/sid.inpe.br/marte2/2019/10.23.11.40/doc/97833.pdf. Acesso em: 14 abr. 2020.

MILITANTE, S. V.; GERARDO, B. D.; MEDINA, R. P. Sugarcane Disease Recognition using Deep Learning. *In*: IEEE EURASIA CONFERENCE ON IOT, COMMUNICATION AND ENGINEERING (ECICE), 2019, Yunlin, Taiwan. **Anais** […].Yunlin, Taiwan, 2019. p. 575-578, https://doi.org/10.1109/ECICE47484.2019.8942690

NUGRAHAENI, R. A.; MUTIJARSA, K. Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification. *In*: INTERNATIONAL SEMINAR ON APPLICATION FOR TECHNOLOGY OF INFORMATION AND COMMUNICATION (ISEMANTIC), 2016, Semarang, **Anais** […]. Semarang,2016, p. 163-168https://doi.org/10.1109/ISEMANTIC.2016.7873831

RADU, M. D.; COSTEA, I. M.; STAN, V. A. Automatic Traffic Sign Recognition Artificial Inteligence - Deep Learning Algorithm. *In:* INTERNATIONAL CONFERENCE ON ELECTRONICS, COMPUTERS AND ARTIFICIAL INTELLIGENCE (ECAI)*, 12.,* 2020. **Anais** […]. [S.l.], 2020,p. 1-4. https://doi.org/10.1109/ECAI50035.2020.9223186

ROUSE Jr, J. W.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the great plains with ERTS. ~~Third~~ *In:*-ERTS-1 SYMPOSIUM, 3., 1973, Washington, D.C **Anais** […]. Washington, 1973, v. 1, p. 309-317. Disponível em: https://ntrs.nasa.gov/citations/19740022614. Acesso em: 28 mar. 2021.

RUBIRA CRULHAS, J. P.; ARTERO, A. O.; PITERI, M. A.; SILVA, F. A.; PEREIRA, D. R.; ELER, D. M.; ALBUQUERQUE, V. H. C. Blank Spots Identification on Plantations. **IEEE Latin America Transactions,** v. 16, n. 8, p. 2115-2121, Aug. 2018.. https://doi.org/10.1109/TLA.2018.8528224

SCRIVANI, R.; ZULLO, J.; ROMANI, L. A. S. SITS for estimating sugarcane production. *In:* INTERNATIONAL WORKSHOP ON THE ANALYSIS OF MULTITEMPORAL REMOTE SENSING IMAGES (MultiTemp), 2017. Brugge, Belgium,. **Anais** […]Brugge, Belgium, 2017 p. 1-4. https://doi.org/10.1109/Multi-Temp.2017.8035254

SKITTOU, M.; MADHOUM, O.; KHANNOUSS, A.; MERROUCHI, M.; GADI, T. Classification of land use areas using remote sensing data with machine learning.*In*: IEEE INTERNATIONAL CONFERENCE OF MOROCCAN GEOMATICS (MORGEO**)**, Casablanca, Morocco. **Anais** […]. Casablenca, Marocco, 2020. p. 1-5. Disponível em: https://ieeexplore.ieee.org/document/9121883. Acesso em: 15 fev. 2021. https://doi.org/10.1109/Morgeo49228.2020.9121883

SOUZA, M. F. DE; AMARAL, L. R, OLIVEIRA, S. R. M.; COUTINHO, M. A. N.; NETTO, C. F. Spectral differentiation of sugarcane from weeds, **Biosystems Engineering,** , Campinas, SP, v. 190, p. 41-46, 2020. https://doi.org/10.1016/j.biosystemseng.2019.11.023

SPERANZA, E. A.; ANTUNES, J. F. G.; INAMASU, R. Y. Uso de imagens de sensoriamento remoto para identificação de variabilidade espacial em Agricultura de Precisão. *In*: SIMPÓSIO DE GEOTECNOLOGIAS NO PANTANAL, JARDIM, MATO GROSSO DO SUL, Brasil, 7.,2018. Jardim. **Anais** [...]. Jardim, MS, 2018. Disponível em: https://www.alice.cnptia.embrapa.br/handle/doc/1099230. Acesso em: 14 abr. 2020.

TREEBUPACHATSAKUL, T.; POOMRITTIGUL, S. Bacteria Classification using Image Processing and Deep learning. INTERNATIONAL TECHNICAL CONFERENCE ON CIRCUITS/SYSTEMS, COMPUTERS AND COMMUNICATIONS (ITC-CSCC**)**, 34., 2019. JeJu, Korea (South). **Anais** […]. JeJu, Korea (South), 2019. p. 1-3. https://doi.org/10.1109/ITC-CSCC.2019.8793320

VASCONCELLOS, B. C.; TRINDADE, J. P. P.; VOLK, L. B. DA S.; DE PINHO, L. B. Method Applied To Animal MonitoringThrough VANT Images.**IEEE Latin America Transactions**, v. 18, n. 07, p. 1280-1287, jul., 2020. https://doi.org/10.1109/TLA.2020.9099770

WANG, M.; LIU, Z.; ALI, B. M. H.; WANG, Y.; Li, Y.; CHEN, Y. Mapping sugarcane in complex landscapes by integrating multi-temporal Sentinel-2 images and machine learning algorithms. **Land Use Policy**, China, v. 88, p. 1-11,2019. Disponível em: https://doi.org/10.1016/j.landusepol.2019.104190

ZHANG, T.; SU, J.; LIU, C.; CHEN, W.; LIU, H.; LIU, G. Band selection in sentinel-2 satellite for agriculture applications., *In:* INTERNATIONAL CONFERENCE ON AUTOMATION AND COMPUTING (ICAC). 23., 2017, Huddersfield, UK. **Anais** […].Huddersfield, UK, 2017. p. 1-6. https://doi.org/10.23919/IConAC.2017.8081990