Investigating Machine Learning techniques applied in cotton crop management

Investigando técnicas de Aprendizado de Máquina aplicadas no manejo da cultura de algodão

Carlos Gabriel S. Rodrigues, Carlos Rafael N. de A. Silva, Allan Vitor W. Toledo, Claudia A. Martins, Raul T. Santos

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT) – Cuiabá – MT – Brasil

carlos.rodrigues,carlos.silva12,allan.toledo{@sou.ufmt.br},
claudia,raul{@ic.ufmt.br}

Abstract. *Cotton is the raw material for several products that are used daily. Its production requires care so that productivity meets expectations. Computational techniques can assist in monitoring and production performance. Therefore, this work aims to use images obtained from cotton production and, using machine learning techniques, seek tools that can assist in the management of cotton cultivation so that the productivity of the plantation can be analyzed. In this work, several experiments were carried out with regression algorithms using R^2 as a principal metric.*

Keywords: *classification, machine learning, precision agriculture, cotton.*

Resumo. *O algodão é matéria prima para diversos produtos que são utilizados diariamente. Sua produção requer cuidados para que a produtividade atenda as expectativas. Técnicas computacionais podem auxiliar no monitoramento e desempenho da produção. Assim, este trabalho tem como objetivo utilizar imagens obtidas de produção de algodão e, a partir de técnicas de aprendizado de máquina, buscar ferramentas que possam auxiliar no manejo da cultura de algodão para que possa analisar a produtividade da plantação. Neste trabalho foram feitos vários experimentos com os algoritmos de regressão usando como métrica principal o R^2.*

Palavras-chave: *classificação, aprendizado de máquina, agricultura de precisão, algodão.*

1. Introdução

A agricultura é parte essencial da economia brasileira, segundo relatório da *United States Department of Agriculture*, o Brasil faz parte de um seleto grupo dos maiores produtores agrícolas do mundo (USDA, 2017). Com o passar dos anos a agricultura vem se especializando a fim de minimizar custos e maximizar a produtividade, com isso diversas ciências e técnicas são empregadas, de modo que a agricultura de precisão passou a ser objeto de interesse na aplicação de técnicas computacionais.

A agricultura de precisão trata-se do gerenciamento de parâmetros e variáveis em um local específico para reduzir o desperdício, aumentar a produção, alavancar os

lucros e reduzir o impacto ambiental (Adamchuk, Perk e Scherpers. 2003, p.1). Nesse contexto, é uma subárea da agricultura que visa obter, processar e gerenciar dados advindos de fazendas e plantações para melhorar a produção e reduzir o impacto ambiental.

A inserção da computação na agricultura tem crescido cada vez mais como ferramentas que possibilitam e auxiliam em uma resposta eficaz às demandas crescentes por eficiência, sustentabilidade e produtividade no setor agrícola. Por meio de aplicação de técnicas de Inteligência Artificial, é possível identificar, por exemplo, condições adversas que prejudicam o crescimento e a produtividade das plantações, como pragas, falta de nutrientes, condições climáticas e água. Alguns trabalhos relacionados mostram a aplicação dessas técnicas na previsão da produção e qualidade da plantação (Budach et al, 2022). Outros trabalhos mostram como identificar deficiências e/ou pragas, na previsão do desenvolvimento e da produtividade da plantação (Habib et al., 2020; Patrício & Rieder, 2018; Gavahale, 2014; Gulhane et al, 2014).

Por meio dessa abordagem, este trabalho consiste em analisar dados reais de imagens do manejo de uma cultura agrícola para identificar, extrair e monitorar o estado do cultivo, relacionando com a produtividade, a partir de alguns índices de vegetação, buscando detectar condições que influenciam no desenvolvimento da cultura, bem como antecipar a previsão do impacto na produtividade. Para isso, serão utilizadas técnicas de pré-processamento e algoritmos de aprendizado de máquina como um processo de tomada de decisão relacionada ao manejo da cultura do algodão.

2. Metodologia

Para o desenvolvimento deste trabalho foi obtido um conjunto de dados relacionados com o manejo da cultura do algodão. Após a obtenção dos dados, foram realizadas as etapas de pré-processamento dos dados e análise e extração de padrões nos dados, descritas a seguir.

2.1. Descrição dos Dados

As imagens da cultura do algodão utilizadas foram coletadas em 4 voos por um drone quadcoptero espectral com câmera multiespectral integrada, ocorridos em dias sem chuva e pouco nublados, entre as 10 e 14 horas. Cada voo representa uma fase (período) de desenvolvimento do cultivo de algodão. As imagens foram mapeadas em talhões das áreas divididas de forma abstrata por *grids* com 8 metros de comprimento por 8 metros de largura (8x8), obtendo, portanto, 54 blocos de área produtiva com aproximadamente 100 metros de comprimento e 100 metros de largura (100x100m).

A partir disso, as imagens receberam alguns tratamentos, por meio de composições, gerando os seguintes índices de vegetação: *Normalized Difference Vegetation Index* (NDVI), *Normalized Difference Red Edge Index* (NDRE), *Simplified Canopy Chlorophyll* (SCCCI), *Soil-adjusted Vegetation Index* (SAVI), *Modified Clorophyll Absorption in Refelectance Index* (MCARI), *Transformed Chlorophyll Absorption Radio Index* (GLI), Temperatura e RGB. Os índices de vegetação são produtos de algoritmos e modelos matemáticos que se baseiam nas características da cobertura vegetal de uma área específica, com o propósito de analisar as propriedades da plantação (Sharma et al, 2020). Cada um dos índices têm como finalidade destacar de maneira mais precisa as condições em que a cultura se encontra, permitindo uma avaliação mais eficaz do seu estado (Oliveira, 2021).

Além das imagens foram utilizados dados da produtividade alcançada por cada um dos 54 blocos, que foram coletados após a colheita.

2.2. Pré-processamento

Preparar os dados para posteriormente os submeter em algoritmos de aprendizado de máquina é uma das tarefas mais essenciais, com isso é possível corrigir as inconsistências e agregar valor aos dados. Essa fase serve para garantir que o processo de aprendizado e identificação de padrões produza resultados sólidos e assegure a qualidade dos resultados em geral.

a) Transformação dos dados

É fundamental traduzir as imagens da base de dados para suas representações matemáticas, ou seja, as converter em matrizes numéricas. Desta forma, foi calculado a média dos valores das imagens dos blocos em cada um de seus índices e em todos os voos. O resultado dessa etapa foi duas tabelas que foram submetidas aos algoritmos de regressão.

A primeira tabela foi denominada de Base-1 e utilizou todos os 8 tipos de imagens, ela é composta por 54 linhas e 32 colunas, onde cada linha representa um bloco de cultivo, e cada coluna representa o voo de uma determinada imagem, por exemplo, `mean_gli_1` é nome de uma coluna que contém as médias do índice GLI no primeiro voo.

A Base-2 foi criada partindo do princípio de que os índices de vegetação que possuem maior correlação com a produtividade seriam de certa forma os ideais para conseguir prever a produtividade de futuras plantações, com isso a base contém somente os índices GLI, MCARI, NDVI e SAVI. E contém 54 linhas, que representam os blocos de cultivo e 16 colunas que contém as médias dos índices dos blocos relacionados à produtividade.

Na planilha de produtividade foi necessário remover as linhas que não continham dados e remover os *outliers,* ou seja, valores que se diferenciam significativamente dos demais do conjunto de dados, que estão fora do padrão. O resultado disso foi a remoção de 9 das 54 linhas de blocos.

Após a separação dos dados, foi inserida a coluna dependente com as respectivas produtividades na Base-1 e na Base-2. Assim, no final as duas bases contêm a média dos blocos e a produtividade atingida por aquele bloco de plantação.

b) Normalização dos dados

Além de transformar as imagens em vetores numéricos e extrair a média das imagens, é necessário normalizar as bases de dados, ajustando as escalas para um intervalo específico, cada intervalo vai depender do algoritmo de pré-processamento que estiver utilizando, e os algoritmos utilizados foram (Scikit-Learn, 2011):

i) *Polynomial Features*: é uma técnica para gerar novas características ou atributos, essas características são obtidas utilizando combinações polinomiais dos atributos originais, ou seja, o algoritmo cria novos recursos de entrada com base nos recursos já existentes.

ii) *StandardScaler*: essa técnica consiste em padronizar os valores dos atributos, de modo que remove a média e escala a variância a uma unidade, assim, a média seria 0 e o desvio padrão 1. Com isso os dados são padronizados, tornando-os mais otimizados para o modelo.

iii) *Select K best*: é informado para função que deseja encontrar os k atributos que explicam melhor o atributo dependente. O objetivo é escolher as *K* melhores características com base em alguma métrica de relevância. Essa técnica é particularmente útil quando possui um grande conjunto de características e deseja reduzi-lo para um subconjunto mais significativo, o que pode melhorar o desempenho do modelo, reduzir a complexidade computacional e evitar problemas de sobreajuste.

Após a preparação e separação das bases, transformação das imagens e normalização dos dados, também, para melhor entendimento dos dados, foram extraídas as médias e desvio padrão dos dados. Os dados, portanto, estão prontos para serem submetidos aos algoritmos de aprendizado.

2.3. Processamento dos dados

Foram utilizados diversos algoritmos de Aprendizado de Máquina (AM), que é uma subárea da Inteligência Artificial, que consiste na aplicação de métodos de análise de dados, identificação de padrões no qual o sistema busca aprender para um processo de tomada de decisões (Oliveira, 2021).

Regressão é uma das técnicas de AM do paradigma supervisionado e, neste caso, um modelo é treinado usando um conjunto de dados rotulados, cujos exemplos de treinamento incluem as entradas (variáveis independentes) e as saídas desejadas (variáveis dependentes). Na regressão, especificamente, o objetivo é prever um modelo com valores numéricos ou contínuos, como preços, temperaturas, pontuações, ou qualquer outra variável quantitativa.

As técnicas, ou algoritmos, que foram utilizados no processamento e análise dos dados neste trabalho, implementados em Python usando a biblioteca *Sklearn* (Scikit-Learn, 2011), foram: *Linear Regression* (LR), *Linear Model - Ridge* (LM-R), *Support Vector Machine Regressor* (SVMR), *k-Nearest Neighboors Regressor* (k-NNR) e *Decision Tree Regressor* (DTR).

Assim, para cada uma das técnicas de regressão utilizadas, os dados da Base-1 e Base-2 foram submetidas para processamento duas vezes, com ajuste (otimização) de hiperparâmetros usando a biblioteca `GridSearchCV` e sem o ajuste de hiperparâmetros, visando obter um melhor desempenho.

3. Resultados e discussão

Como mencionado, após o processamento com os algoritmos selecionados e o pré-processamento, foram realizados experimentos usando duas bases de dados distintas: a base original (Base-1) com todos os atributos (índices) originais, e a base com seleção de atributos (Base-2) na qual foram selecionados apenas os índices que possuem maior correlação com a produtividade.

Para separar os dados foi utilizado o `train_test_split` da biblioteca Sklearn que os dividiu em 80% para treino e 20% para teste, sem validação cruzada. Além disso, foi utilizado um valor fixo para o parâmetro `random_stutus` que estabelece uma mesma distribuição dos dados, o que permite uma comparação entre os resultados obtidos pela Base-1 e Base-2. A biblioteca tem como função principal dividir o conjunto de dados, para isso a função tem diversos parâmetros que podem ser ajustados de acordo com a necessidade.

3.1. Análise estatística dos dados

Inicialmente, realizou-se uma análise dos dados para melhor entendimento sobre seu comportamento e integridade, para isso foram utilizados cálculos estatísticos e a biblioteca Pandas, como demonstrado na Tabela 1.

Tabela 1. Valores estatísticos da relação de cada imagem com a produção de algodão por talhão

Tipo de Imagem	Maior Valor	Média	Desvio Padrão
GLI	0,480	0,345	0,138
NDVI	0,470	0,331	0,157
SAVI	0,500	0,323	0,174
MCARI	0,510	0,301	0,177
NDRE	0,440	0,298	0,146
SCCCI	0,440	0,230	0,168
RGB	0,130	-0,127	0,221
Temperatura	0,078	-0,153	0,226

A fim de entender quais índices melhor se relacionavam com a produção e o comportamento entre eles, também foi feito um mapa de correlação de cada voo e calculado os valores estatísticos da relação de cada imagem com a produção.

3.2. Experimento com dados Base-1

Após o pré-processamento dos dados, os mesmos foram submetidos aos algoritmos de regressão citados anteriormente. Os resultados obtidos após o processamento dos dados são mostrados na Tabela 2, na qual os valores são representados no formato “xx / yy”. O valor “xx” representa os resultados dos algoritmos **sem** o ajuste de hiperparâmetros, e “yy” **com** o ajuste de hiperparâmetros, respectivamente. Os valores apresentados na tabela são os coeficientes R^2 (*R-squared*) indicando o quanto os modelos se ajustaram aos dados.

Tabela 2. R^2 dos algoritmos de regressão usando Base-1

Algoritmo	Normal	Polynomial Features	Standard Scaler	Select K Best
LR	-104.85 / -0.31	-104.85 / -0.31	-104.85 / -0.31	-104.85 / -0.31
LM-R	-0.89 /-0.87	-0.89 /-0.87	-0.89 /-0.87	-0.89 /-0.87

SVMR	-0.29 / -0.29	-0.29 / -0.29	-0.29 / -0.29	-0.29 /-0.29
k-NNR	-0.18 / -0.19	-0.18 / -0.19	-0.18 / -0.19	-0.18 / -0.19
DTR	-1.92 /-1.62	-1.74 /-1.62	-1.94 /-1.62	-4.39 /-1.62

Analisando os resultados mostrados na Tabela 2, é possível verificar que o processamento dos algoritmos que utilizam ajuste de hiperparâmetros apresenta os melhores resultados na maioria das vezes, visto a melhora do desempenho da regressão em alguns algoritmos, além disso, a escolha da técnica de normalização não tem impactos significativos nos resultados dos algoritmos de regressão.

Na Figura 1 são mostrados graficamente o desempenho dos algoritmos de regressão sem ajuste de hiperparâmetros e com o ajuste de hiperparâmetros quando aplicados a Base 1.

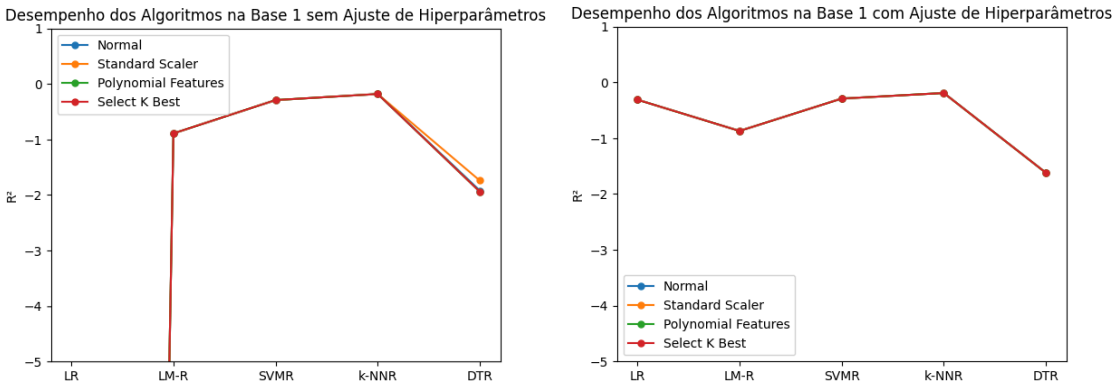

Figura 1. R² dos algoritmos sem hiperparâmetros (esquerda) e com hiperparâmetros (direita)

Analisando os resultados apresentados sem o ajuste dos hiperparâmetros é possível, portanto, verificar:

- as técnicas de normalização não tiveram impactos nos algoritmos de regressão, com exceção do DTR;
- LR apresentou um desempenho inferior em relação aos outros algoritmos, visto que sua diferença com o DTR, algoritmo de segundo pior desempenho, é de -102,93;
- LM-R, SVMR e k-NNR obtiveram valores de R² entre 0 e -1, apresentando os melhores resultados sem o ajuste dos hiperparâmetros. Além disso, se destaca k-NNR que teve melhor desempenho;
- DTR mostrou desempenho melhor com o uso do Polymonial Features e piores resultados com uso do Select K Best;

Com relação aos resultados dos algoritmos que utilizam hiperparâmetros, pode-se verificar as seguintes observações:

- as técnicas de normalização não tiveram impacto nos algoritmos de regressão, inclusive o DTR;

- LR obteve uma melhora de 104,54 no seu desempenho se tornando o algoritmo com o terceiro melhor desempenho;
- todas as técnicas tiveram melhoras no seu desempenho, com exceção da k-NNR.

Portanto, o ajuste de hiperparâmetros tende a melhorar o desempenho dos modelos de regressão em comparação com os valores padrão. No entanto, a escolha do algoritmo e da técnica de pré-processamento de dados, neste trabalho, não melhorou de forma significativa o desempenho geral do modelo.

3.2. Experimento com dados Base-2

Na Tabela 3 é mostrado os resultados da Base-2, na qual foram selecionados apenas os índices que possuem maior correlação com a produtividade. São apresentados os resultados obtidos com os algoritmos de regressão e as técnicas de pré-processamento utilizados. Nesta etapa o algoritmo Select K Best foi retirado, visto que a remoção de atributos foi feita de acordo com a correlação em relação a produtividade.

Tabela 3. Resultados dos algoritmos de regressão usando a Base-2

Algoritmo	Normal	Polynomial Features	StandardScaler
LR	-0.65 /-0.14	-0.65 /-0.14	-0.65 /-0.14
LM-R	0.01 / 0.01	0.01 / 0.01	0.01 / 0.01
SVMR	-0.30 / -0.35	-0.30 / -0.30	-0.30 / -0.45
k-NNR	-0.54 /-0.06	-0.54 /-0.06	-0.54 /-0.06
DTR	-0.88 /-1.59	-3.82 /-1.51	-0.71 /-3.00

Analisando a Tabela 3 é possível verificar que as técnicas de normalização também não geraram impactos significativos e o ajuste de hiperparâmetros não causou melhora em todos os algoritmos. Na Figura 2 é mostrado de forma gráfica o desempenho dos algoritmos. Uma possível hipótese que está sendo investigada é com relação ao tamanho do conjunto de dados e a variabilidade do conjunto de dados que, possivelmente, não está impactando em melhorias no desempenho dos algoritmos e no pré-processamento.

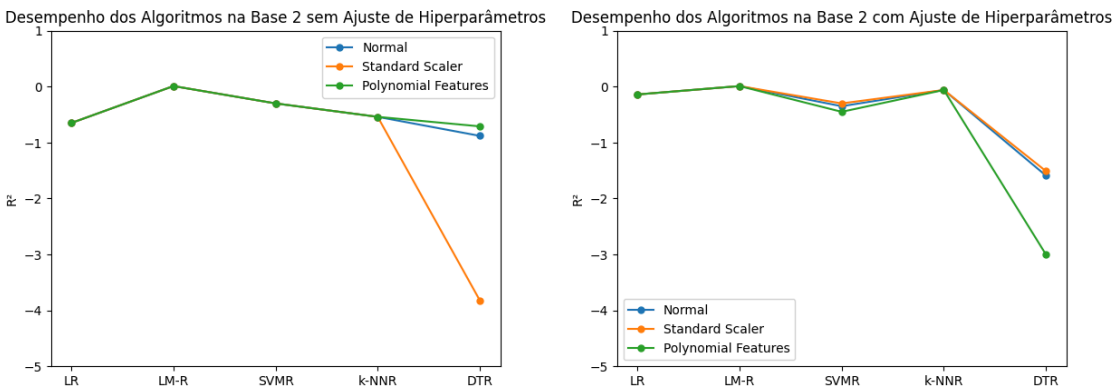

Figura 2. Desempenho dos algoritmos sem hiperparâmetros (esquerda) e com hiperparâmetros (direita)

Analisando os resultados sem ajuste de hiperparâmetros é possível, portanto, destacar as seguintes observações:

- as técnicas de normalização não tiveram impactos nos algoritmos de regressão, com exceção do DTR;
- todos os algoritmos tiveram valor de desempenho superior a -1, com exceção do DTR que obteve com o Polynomial Features um desempenho inferior a isso;
- Destaca-se o LM-R que obteve o maior valor, sendo ele maior do que 0.

Com relação aos resultados dos algoritmos que utilizam hiperparâmetros, pode-se verificar as seguintes observações:

- as técnicas de normalização não tiveram impactos significativos, mas dentre elas a que teve melhor desempenho foi a Standard Scaler;
- o uso de hiperparâmetros implicou na melhora do desempenho dos modelos de LR e k-NNR;
- LM-R não teve nenhuma mudança com uso de hiperparâmetros;
- DTR conseguiu uma piora na maioria dos seus resultados que usaram hiperparâmetro.

Portanto, o ajuste de hiperparâmetros em sua maioria nos casos analisados mostrou um desempenho igual ou melhor aos casos sem ajuste. Além disso, a normalização não apresentou impactos significativos para os resultados.

4. Conclusão e Trabalhos Futuros

Neste trabalho foi proposto o uso de técnicas de pré-processamento e aprendizado de máquina para o auxílio no manejo da cultura do algodão. Vários experimentos foram realizados na análise de melhor desempenho. Como resultado, a base de dados com os índices de maior correlação com a produção apresentou melhores resultados e homogêneos. No entanto, de forma geral, os modelos de regressão tiveram um desempenho semelhante, com exceção da técnica DTR que teve um desempenho diferente em todos os pré-processamentos.

Além disso, com o uso de hiperparâmetros houve uma pequena melhora no desempenho dos algoritmos, principalmente com o algoritmo LR. Apesar do maior resultado de R^2 ter sido de 0.01, foi um desempenho positivo considerando que a base de dados tem um tamanho muito pequeno e os algoritmos tiveram dificuldade na generalização. No entanto, é interessante observar o uso das médias das imagens para predição da produção da cultura de algodão, visto que ainda existem outros parâmetros, algoritmos e normalizações que podem ser aplicados na base dados. Novos dados estão sendo coletados e incorporados ao treinamento, assim como, novos experimentos e análise do desempenho dos algoritmos estão sendo realizados para investigar os resultados obtidos e sua relação à quantidade e representatividade de dados disponíveis para treinamento e teste.

5. Referências

Adamchuk, Viacheslav I.; Perk, Richard L.; and Schepers, James S., "EC03-702 Precision Agriculture: Applications of Remote Sensing in Site-Specific Management" (2003). **Historical Materials from University of Nebraska-Lincoln Extension**. 705.

Budach, Lukas et al. The effects of data quality on machine learning performance. **arXiv preprint arXiv:2207.14529**, 2022.

Gavahale, Kiran R. et al. Uma visão geral da pesquisa sobre detecção de doenças em folhas de plantas usando técnicas de processamento de imagem. **Iosr journal of computer engineering** (iosr-jce) , v. 16, n. 1, pág. 10-16, 2014.

Gulhane, Viraj A.; Kolekar, Maheshkumar H. Diagnóstico de doenças em folhas de algodoeiro usando classificador de análise de componentes principais. In: **Conferência Anual IEEE Índia de 2014** (INDICON). IEEE, 2014. pág. 1-5.

Habib, M. T. et al. Machine vision based papaya disease recognition. **Journal of King Saud University-Computer and Information Sciences**, v. 32, n. 3, p. 300-309, 2020.

Oliveira, Mailson Freire de. **Previsão e estimativa de variáveis de cultivo utilizando níveis e formas de sensoriamento remoto e técnicas de aprendizado de máquina**. 2021.

Patrício, D. I.; Rieder, R. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. **Computers and electronics in agriculture**, v. 153, p. 69-81, 2018.

Rodrigues, Luiz Souza; Pereira, Danilo Roberto. Aprendizado de Máquina Aplicado em Imagem NDVI para previsão da produtividade da Cana-de-Açúcar. In: **Colloquium Exactarum**. ISSN: 2178-8332. 2021. p. 82-98.

Sharma, Abhinav et al. Machine learning applications for precision agriculture: A comprehensive review. **IEEE Access**, v. 9, p. 4843-4873, 2020.

USDA (2017) “World Agricultural Production”, In: **Circular Series,** http://apps.fas.usda.gov/psdonline/circulars/production.pdf.

API deisgn for machine larning software: experiences from the scikit-learn project, Buitinck et al., 2013.

Documentação do scikit-learn. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html. Acesso em: 5 de agosto de 2023.

Scikit-Learn: Machine Learning in Python, Pedregosa et al, JMLR 12, pp 2825-2830, 2011.