 doi:10.20944/preprints202112.0138.v2

Comparison of machine learning techniques in cotton yield prediction using satellite remote sensing

Francielle Morelli-Ferreira[123*], Nayane Jaqueline Costa Maia[1], Danilo Tedesco[1], Elizabeth Haruna Kazama[1], Franciele Morlin Carneiro[2], Letícia Bernabé dos Santos[2], Getulio de Freitas Seben Junior[23], Glauco de Souza Rolim[1], Luciano Shozo Shiratsuchi[2], Rouverson Pereira da Silva[1]

[1]Department of Engineering and Mathematical Sciences, Sao Paulo State University, Jaboticabal, SP, Brazil.
[2]School of Plant Environmental and Soil Sciences, Louisiana State University AgCenter, Baton Rouge, LA.
[3]Faculty of Applied Social Sciences and Agricultural, State University of Mato Grosso, Nova Mutum, MT, Brazil.
*Corresponding author: francielle@unemat.br

Abstract: The use of machine learning techniques to predict yield based on remote sensing is a no-return path and studies conducted on farm aim to help rural producers in decision-making. Thus, commercial fields equipped with technologies in Mato Grosso, Brazil, were monitored by satellite images to predict cotton yield using supervised learning techniques. The objective of this research was to identify how early in the growing season, which vegetation indices and which machine learning algorithms are best to predict cotton yield at the farm level. For that, we went through the following steps: 1) We observed the yield in 398 ha (3 fields) and eight vegetation indices (VI) were calculated on five dates during the growing season. 2) Scenarios were created to facilitate the analysis and interpretation of results: Scenario 1: All Data (8 indices on 5 dates = 40 inputs) and Scenario 2: best variable selected by Stepwise regression (1 input). 3) In the search for the best algorithm, hyperparameter adjustments, calibrations and tests using machine learning were performed to predict yield and performances were evaluated. Scenario 1 had the best metrics in all fields of study, and the Multilayer Perceptron (MLP) and Random Forest (RF) algorithms showed the best performances with adjusted R2 of 47% and RMSE of only 0.24 t ha^{-1}, however, in this scenario all predictive inputs that were generated throughout the growing season (approx. 180 days) are needed, so we optimized the prediction and tested only the best VI in each field, and found that among the eight VIs, the Simple Ratio (SR), driven by the K-Nearest Neighbor (KNN) algorithm predicts with 0.26 and 0.28 t ha^{-1} of RMSE and 5.20% MAPE, anticipating the cotton yield with low error by ±143 days, and with important aspect of requiring less computational demand in the generation of the prediction when compared to MLP and RF, for example, enabling its use as a technique that helps predict cotton yield, resulting in time savings for planning, whether in marketing or in crop management strategies.

Keywords: Yield mapping; Vegetation index; Stepwise; Simple Ratio; Random Forest; KNN.

1. Introduction

Cotton (*Gossypium hirsutum L.*) is the main source of natural fiber and plays an important role in the economy of several countries. Although in the USA and Brazil (the world's largest fiber exporters – USDA, 2020) cotton harvesting is fully mechanized with an elevated level of technology, manual methods are still used to estimate yield before the harvesting operation on farms that do not have harvesters with a yield sensor. In this conventional agriculture, the average yield until now is determined with the help of weight scales. Despite being a reliable method, it is characterized by destructive sampling collections in the field and that takes time to perform due to the need for sampling and labor.

The complex big data scenario and the exploitation of artificial intelligence (AI) in digital agriculture allow studies on the different responses measured by monitoring and data processing technologies available for free on digital platforms (satellite images, climate/weather parameters, time series for yield forecast) processed in open source software such as QGIS and programming language environments such as Java, R, Python and Jupyter notebook environments such as Anaconda and Google Collaboratory), becoming increasingly popular in the field of digital agriculture, enabling the use of machine learning algorithms, free of charge, encouraging research and development of commercial software. However, for cotton, which is a crop with a long cycle in Brazil (up to 220 days) and with an

 doi:10.20944/preprints202112.0138.v2

indeterminate growth habit, this type of approach is incipient, especially using on farm precision experimentation, making this work a differential for the literature.

The most popular techniques used to analyze images include vegetation indices, regression analysis and correlations, where in recent years machine learning techniques (Singh et al., 2016) are being directed towards the agricultural environment due to the intensification of digital agriculture and generating large volumes of data. Machine learning (ML) algorithms are very promising for analyzing large volumes of data, acting in a faster, more efficient, and more accurate way. Some examples of ML algorithms are K-Nearest Neighbor (KNN), Multiple Linear Regression (MLR), Artificial Neural Networks – Multilayer Perceptron (MLP) and Random Forest (RF) (Badnakhe et al., 2018)

Studies evaluating machine learning algorithms with cotton begin to emerge. Researchers such as Fue et al. (2018); Xu et al. (2018) and Yeom et al. (2018) proposed methods based on deep learning to identify regions of interest in cotton images (eg, morphological features). Xu et al. (2018) developed a method using a convolutional neural network to detect and count the number of newly opened cotton flowers in aerial images. Yeom et al. (2018) created an algorithm to identify and count bolls by estimating cotton yield using aerial imagery. And recently in Brazil, Oliveira-Tedesco et al. (2020) proposed an approach that estimates cotton yield quickly and efficiently using the identification and counting of bolls obtained by smartphone camera in commercial production fields.

At the region/county level and based on climatic parameters, Aparecido et al. (2020) estimated the cotton yield in the Midwest region of Brazil, using machine learning algorithms and emphasized that it is possible to accurately predict the cotton yield for the main production regions in Brazil, concluding that the best algorithm was the Extras- tree-regressor (TREE) and the one with the lowest performance was the Multiple Linear Regression. According to the authors, the TREE algorithm was able to predict with an anticipation around ± 80 days before harvest.

The use of machine learning techniques for yield prediction based on remote sensing is constantly growing. In the search for new results that prove the prediction of cotton yield to help the cotton grower in anticipating decision-making, whether for marketing, or change in the management of the area or the harvest operation itself, we proposed research on this topic, but with a focus on commercial fields and based on vegetation indices obtained from free medium-resolution satellite images.

The objective of this research was to identify how early in the growing season, which vegetation indices and which machine learning algorithms are best to predict cotton yield at the farm level.

2. Material and Methods

2.1 Fields of Study

The research was conducted in the 2019 season in the northern mesoregion of the State of Mato Grosso, the largest cotton producer in Brazil (CONAB, 2020). The study sites (Figure 1) were three commercial cotton fields on a farm near the city of Santa Rita do Trivelato (13º59'17" S, 55º23'15" W).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

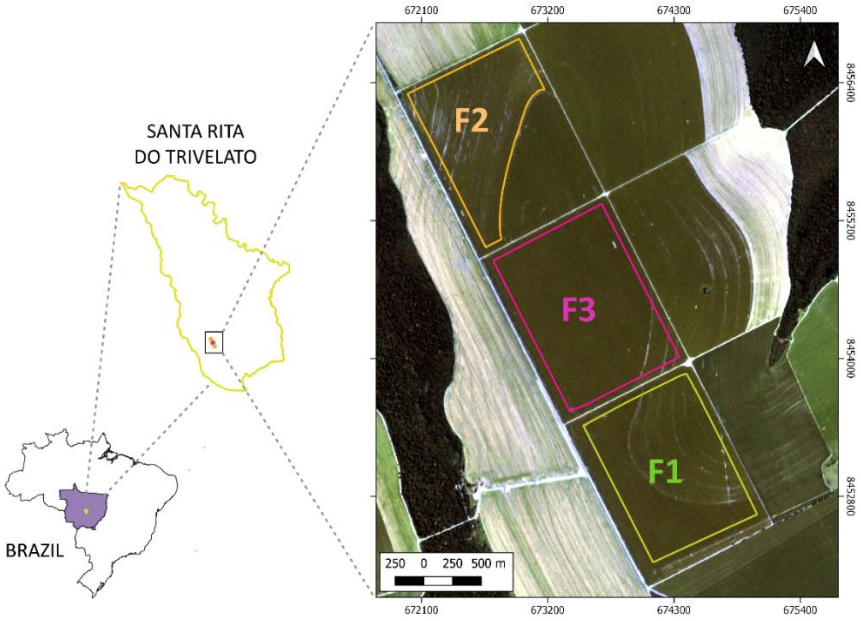

Figure 1. Location of study fields in Santa Rita do Trivelato, Mato Grosso state, Brazil.

The fields were named as: Field 1 (137 ha), Field 2 (102 ha) and Field 3 (159 ha) totaling 398 ha of study. The fields have a predominantly flat topography and the soils in the study region are classified as Dystrophic Red Yellow Oxisol (EMBRAPA, 2013), with medium texture (30 to 35% clay). The climate in the region is tropical hot and humid, with dry winter (Aw) (Köppen & Geiger, 1928), with an average annual temperature close to 25 °C and rainfall of approximately 1800 mm per year.

The crop was monitored by technicians following the best management practices of the the commercial farm to control pests and diseases, as well as all the operations necessary for the proper establishment and conduction of the crop. The cultivar used was IMA5801B2RF, with row spacing of 0.76 m, seeded in January/2019 and harvested in July/2019.

During the cotton cycle evaluated in the study fields, the average temperature was 22.5 °C with accumulated precipitation of 654 mm, being close to the 700 mm required by the culture. Meteorological data directly influence the development conditions of the cotton plant and are presented in Fig. 2.

Climatic data were obtained on a daily scale from the National Aeronautics and Space Administration/Prediction of Worldwide Energy Resource Platform - NASA/POWER (Stackhouse, 2010) and was important for interpretation of vegetation index results on study dates.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

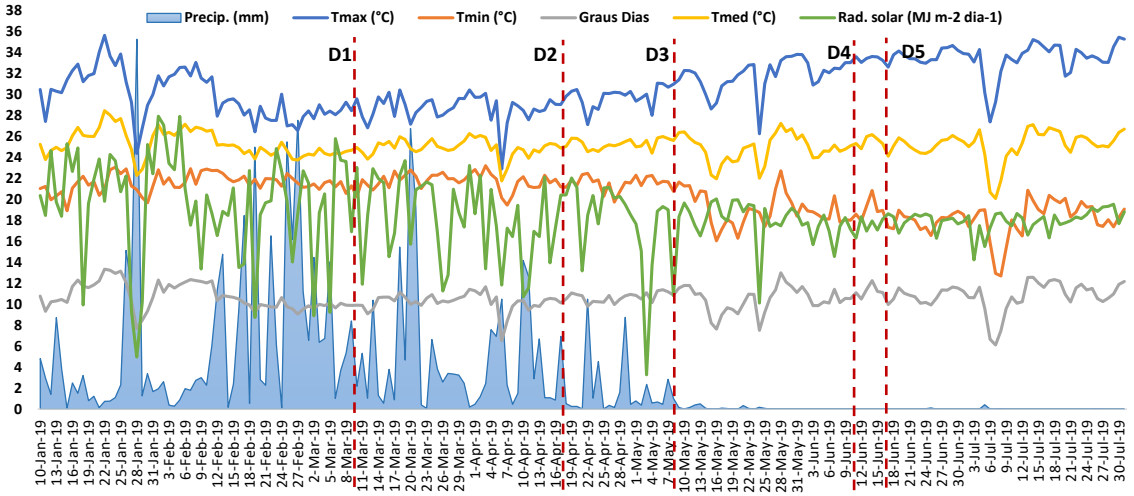

Figure 2. Precipitation (mm), maximum temperature (°C), minimum (°C), degree days, average temperature (°C) and solar radiation (MJ m^2 day^{-1}) for the study fields. D1, D2, D3, D4 and D5 are the acquisition dates of the images used in Santa Rita do Trivelato, Mato Grosso, Brazil.

2.2 Acquisition and processing of satellite imagery

During the 2019 season, Sentinel-2A images were acquired (Table 1) by the Earth Explorer portal of the USGS (United States Geological Survey) along the cotton phenological stages for monitoring and extraction of vegetation indices. The study sites were subsets of the total image, normalized for reflectance, corrected for atmospheric effects and, through the spectral values of the Green (560 nm), Blue (490 nm), Red (665 nm), Red Edge (705 nm) bands nm) and NIR (842 nm), the vegetation indices were calculated as the input predictors for the prediction of cotton yield.

Table 1. Characteristics of Sentinel-2A satellite images (10 m resolution) in the study area dates. Santa Rita do Trivelato - MT, 2019.

Acquisition date	Phenological Stadium	Mean angle of the azimuth of the sun	Mean angle of the sun's zenith	Relative humidity (%)
D 1 – 08/03/19	B (flower button)	27.85	73.30	86.95
D 2 – 17/04/19	F (flowering)	34.10	45.16	84.66
D 3 – 07/05/19	C (boll opening)	38.40	37.17	79.53
D 4 – 11/06/19	C (boll opening)	43.87	32.44	51.16
D 5 – 16/07/19	C (boll opening)	44.22	32.53	47.38

2.3 Phenology and thermal sum of cotton

Understanding the effects of heat stress on crops is extremely important, due to the large variations in temperatures, which can occur from one day to another or even between periods of the same day. The thermal sum of degrees days followed the equation proposed by Arnold (1959), using a basal temperature of 15°C for cotton culture according to Rosolem (2020).

$$ST = \sum GD = \frac{(TM - Tm)}{2} - Tb \quad (1)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

Among the environmental factors, temperature is one of the parameters that most influence cotton development. A clear example of this is the use of monitoring the crop phenology (Fig. 2) through the sum of the degree-days that was carried out in this study, as higher mean temperatures result in the advancement of the crop cycle (Reddy et al., 1996).

The optimum temperature for cotton development is in the range between 20ºC and 30ºC, however, it is common for cotton to be cultivated in regions with temperatures below 15ºC (USA - emergency period) and above 40ºC (India) (Snider & Kawakami, 2014). Rosolem (2020) describes that the cotton cycle is divided into four phases: vegetative (V), formation of flower buds (B), flower opening (F) and boll opening (C). In Figure 3, we can see in detail the phenological phases of cotton described by Aparecido et al. (2020).

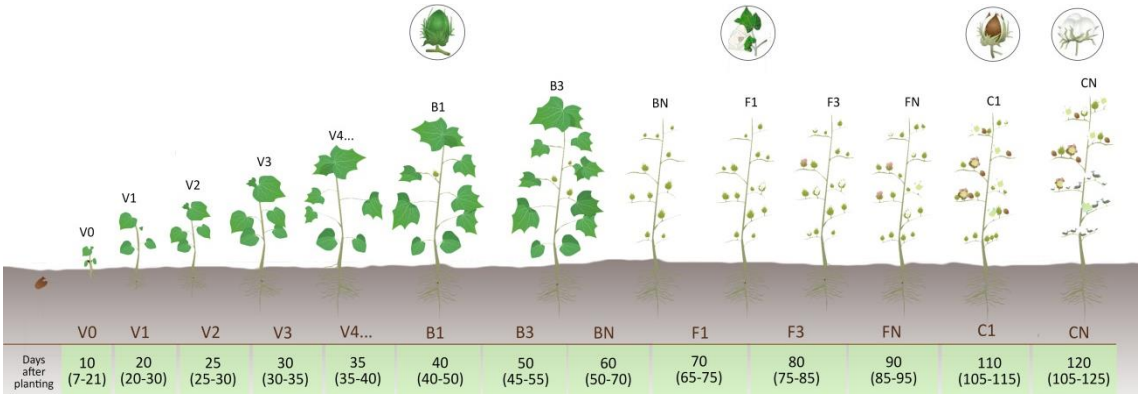

Figure 3. Phenology average for the cotton crop. Legend: V 0 = beginning of the plant emergency; V1 = V0 to the main rib of the second leaf; V2 = V1 to the main rib of the third leaf; V3 = V2 to the main rib of the fourth leaf; V4 = V3 to the main rib of the fifth leaf; B1 = first visible floral bud; B3 = first floral bud visible on the third branch; F1 = opening of the first flower in the first fruit branch; F3 = opening of the first flower in the third fruit branch and C1 = first apple of the first open branch (boll cotton). Source: Aparecido et al. (2020).

2.4 Vegetation indices - input predictors

A sample grid of 10,000 points was created for each field of study to extract the spectral values of each image band, together with the yield value at each point. After the 10,000 points files with the geolocated information of the 2, 3, 4, 5 and 8 bands (Blue, Green, Red, RedEdge1 and NIR), the vegetation indices (IV) for each area in each date of the images obtained, as shown in Fig. 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

Figure 4. Scheme for the creation of points and interpolation of yield data and extraction of vectorized pixel values from the spectral values of bands 2 (Blue), 3 (Green), 4 (Red), 5 (Red Edge), 8 (NIR) of the images of the Sentinel 2A, totaling 10.000 features per field of study.

The VI used in this study were Simple Ratio (SR, Birth and McVey, 1968), the Normalized Difference Vegetation Index (NDVI, Rouse et al., 1973), Normalized Difference Red Edge (NDRE, Barnes et al. 2000), the Red Green Blue Vegetation Index, (RGBVI, Bendig et al., 2015), the Soil Adjusted Vegetation Index (SAVI, Huete, 1988), the Infrared Percentage Vegetation Index (IPVI, Crippen, 1990), the Enhanced Vegetation Index (EVI and EVI2 , Huete et al., 1997) calculated according to Table 2, in Eq. 2 to 9.

Tabela 2. Índices de vegetação calculados no estudo como preditores de entrada do rendimento.

$$SR = \frac{NIR}{R} \quad (2)$$

$$NDVI = \frac{NIR - R}{NIR + R} \quad (3)$$

$$NDRE = \frac{NIR - RE1}{NIR + RE1} \quad (4)$$

$$RGBVI = \frac{G^2 - (B\ .\ R)}{G^2 + (B\ .\ R)} \quad (5)$$

$$SAVI = \frac{NIR - R}{(NIR + R + L)} .(1 + L) \quad (6)$$

$$IPVI = \frac{NIR}{(NIR + R)} \quad (7)$$

$$\text{EVI 1} = 2.5 * \frac{(NIR - R)}{(NIR + 6 * RED - 7.5\ * B + 1)} \quad (8)$$

$$\text{EVI 2} = 2.5 * \frac{(NIR - R)}{NIR + (2.4 * R) + 1} \quad (9)$$

Where: NIR = Near Infrared reflectance, R = Red reflectance, RE1 = Red Edge reflectance1, G = Green reflectance, B is the Blue reflectance and L is an empirical fit coefficient for soils (0.5).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

2.5 Observed yield – dependent variable

Mechanized cotton harvesting was in July 2019 using a model CP690 spindle harvester, which has internal baler module technology enabling it to harvest, press, bale and deposit the modules in the field without stopping the machine. The harvester has a yield monitor sensor and data was collected from a calibrated picker and exported using Ag Leader SMS Basic software. Data were exported without pre-processing, and it was necessary to filter yield data outliers for each field.

For this filtering process we used the methodology proposed by Menegatti and Molin (2004). Finally, we analyzed filtered and coherent data using descriptive statistics. Due to the large amount of data, for the interpolation and generation of maps, we again performed another filtering following the methodology suggested by Tukey (1977), to remove outliers, which, based on the quartiles, calculate the upper and lower limits of the seed cotton yield (t ha^{-1}).

A -50 m buffer was performed in the polygon of each field to ensure that the spectral reflectance values in each pixel of the orbital images were derived only from the canopy reflectance of the cotton plants, without influence from surrounding areas such as roads or adjacent fields.

The vector data files had from 108,000 to 162,700 features/points in each field of study, thus, these data were rasterized by interpolating them by Inverse Distance Weighting (IDW), using the QGIS software version 3.12.3 (QGIS, 2009) with a 10 x 10 m pixel grid to match the pixel size of the Sentinel-2A images.

The interpolated yield values were used to adjust prediction models during the training process (Figure 3). Thus, the yield values were extracted from the pixels along with the spectral data of each band using the Point Sampling Tool (QGIS software plugin), so that inputs and targets have corresponding geolocation. The evaluation of the normality of the yield data was carried out using the Ryan-Joiner test.

2.6 Correlation, selection, and standardization of predicting variables

Pearson's correlation between variables was used to understand the relationship between vegetation indices and cotton yield. Stepwise Regression (Forward) was used to select the most important variable for the response variable (observed yield), and it was conducted to solve a problem of multicollinearity in the regression model. Therefore, this method provides an initial screening of candidate variables when you have a large group of variables, thus selecting the variable(s) that has the highest R^2.

Standardizing a dataset is a common requirement for many machine learning estimators. Therefore, the database was standardized by transforming them into standard deviations ranging from 1 to -1 (Pedregosa et al., 2011).

2.7 Machine Learning Algorithms for Cotton Yield Prediction

Scenarios were created to facilitate the analysis and interpretation of results. Scenario 1: All Data (40 inputs) and Scenario 2: the most important variable selected by Stepwise Regression (1 input), as shown in Table 3.

Table 3. Predictive variables used in machine learning algorithms to predict cotton yield. Scenario 1 (All Data) includes all vegetation indices on all evaluated dates. Scenario 2: includes a single most important variable selected by Stepwise Regression.

	FIELD 1	FIELD 2	FIELD 3
Scenario 1 All Data	SR-D1, SR-D2, SR-D3, SR-D4, SR-D5, NDVI-D1, NDVI-D2, NDVI-D3, NDVI-D4, NDVI-D5, NDRE-D1, NDRE-D2, NDRE-D3, NDRE-D4, NDRE-D5, RGBVI-D1, RGBVI-D2, RGBVI-D3, RGBVI-D4, RGBVI-D5, SAVI-D1, SAVI-D2, SAVI-D3, SAVI-D4, SAVI-D5, IPVI-D1, IPVI-D2, IPVI-D3, IPVI-D4, IPVI-D5, EVI-D1, EVI-D2, EVI-D3, EVI-D4 , EVI-D5, EVI2-D1, EVI2-D2, EVI2-D3, EVI2-D4, EVI2-D5	SR-D1, SR-D2, SR-D3, SR-D4, SR-D5, NDVI-D1, NDVI-D2, NDVI-D3, NDVI-D4, NDVI-D5, NDRE-D1, NDRE-D2, NDRE-D3, NDRE-D4, NDRE-D5, RGBVI-D1, RGBVI-D2, RGBVI-D3, RGBVI-D4, RGBVI-D5, SAVI-D1, SAVI-D2, SAVI-D3, SAVI-D4, SAVI-D5, IPVI-D1, IPVI-D2, IPVI-D3, IPVI-D4, IPVI-D5, EVI-D1, EVI-D2, EVI-D3, EVI-D4 , EVI-D5, EVI2-D1, EVI2-D2, EVI2-D3, EVI2-D4, EVI2-D5	SR-D1, SR-D2, SR-D3, SR-D4, SR-D5, NDVI-D1, NDVI-D2, NDVI-D3, NDVI-D4, NDVI-D5, NDRE-D1, NDRE-D2, NDRE-D3, NDRE-D4, NDRE-D5, RGBVI-D1, RGBVI-D2, RGBVI-D3, RGBVI-D4, RGBVI-D5, SAVI-D1, SAVI-D2, SAVI-D3, SAVI-D4, SAVI-D5, IPVI-D1, IPVI-D2, IPVI-D3, IPVI-D4, IPVI-D5, EVI-D1, EVI-D2, EVI-D3, EVI-D4 , EVI-D5, EVI2-D1, EVI2-D2, EVI2-D3, EVI2-D4, EVI2-D5
Scenario 2 Stepwise	VI-Date	VI-Date	VI-Date

We use different algorithms to predict cotton yield. In all cases, the yield was the dependent variable. The results of eight vegetation indices on the five assessment dates were the independent variables of the All-Data scenario. The best variable in this set was the independent variable from Scenario 2.

The algorithms used to predict cotton yield were: 1) Linear Regression; 2) Multilayer Perceptron (MLP); 3) Random Forest Regression (RF); 4) K-Nearest Neighbor Regressor (KNN); and 5) AutoML: Automatic machine learning.

LR is the simplest and fastest approach used in the study. LR fits a linear model with coefficients to minimize the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation (Pedregosa et al. 2011).

KNN is a simple technique, easy to implement and very flexible, this method finds a group of k samples (training data) closest to unknown samples (test data). In samples k, the unknown samples are determined by the mean of the response variable (Aparecido et al. 2020).

RF is a widely used LM technique for crop forecasting (Everingham et al. 2016). Random Forests are a combination of tree predictors, such that each tree depends on the values of an independently sampled random vector with the same distribution for all trees in the forest (Breiman, 2001).

MLP was the artificial neural network used, in which a minimum of three layers are defined: input, hidden and output layers. The activation function used was rectified linear (ReLU). MLP training was backpropagation. Return propagation is a form of supervised learning in which the error rate is sent back across the network to change the weights to improve prediction and decrease error (Kaul et al. 2005).

For the automatic ML, the H20 platform's AutoML was used, which was designed to have as few parameters as possible, so that the user only needs to point to the data set, identify the response column and, if he chooses, define a limit in the total number of trained models. Quickly, AutoML performs a hyperparameter search on a variety of H2O algorithms to provide the best model (LeDell & Poirier, 2020).

To provide a basis for comparison, the same set of predictors was kept for all evaluated algorithms, randomly separating them into 70% of data for training and 30% for testing. For the MLP, RF and KNN algorithms, we optimized the models with hyperparameters using GridSearchCV (Pedregosa et al., 2011).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

2.8 Algorithm Performance Evaluation

Observed and predicted field yield data in all models were compared using statistical indices regarding accuracy and precision.

Accuracy indicates the proximity of an estimate of the observed value and was evaluated using the root mean square error (RMSE) given in the same unit as the predicted variable (in t ha-1), and the mean absolute percentage error (MAPE), eq . 13 and 14.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Yobs_i - Yest_i)^2}{n}} \quad (13)$$

$$\text{MAPE} = \frac{\sum_{i=1}^{n}\left(\left|\frac{Yest_i - Yobs_i}{Yobs_i}\right| * 100\right)}{n} \quad (14)$$

Where: n is the data number, Yest is the yield value estimated by the algorithm, and Yobs is the actual observed yield.

Precision is the ability of a model to repeat an estimate and was evaluated using the adjusted coefficient of determination R^2 (CORNELL; BERGER, 1987), eq. 15.

$$R^2 adjusted = 1 - \frac{(1 - R^2)\ x\ (n - 1)}{n - k - 1} \quad (15)$$

3 Results and Discussion

3.1 Observed yield x predicting variables

The average seed cotton yield observed in the study fields ranged from 3.20 to 3.65 t ha^{-1} (Table 4). Descriptive statistics are presented in Table 4, and we observed that the yield data resulted according to a normal distribution (values close to 1) by the Ryan-Joiner test in the three fields of study.

The highest average yield was observed in field 1 (3.65 t ha^{-1}) with the lowest coefficient of variation of 9.31 %, this low variation can be observed in the yield map (Fig. 3) for dealing with commercial data at farm level. Despite the low variation in the average yield between the fields evaluated, these results are below the average for the state of Mato Grosso, which was 4.29 t ha^{-1} (CONAB, 2020).

Table 4. Descriptive statistics of seed cotton yield data observed in the study fields on the farm in Santa Rita do Trivelato, Mato Grosso, Brazil, 2019.

Description	Field 1	Field 2	Field 3
Ryan-Joiner	0.99^N	0.99^N	0.98^N
Average (t ha^{-1})	3.65	3.20	3.48
Median (t ha^{-1})	3.65	3.21	3.47
Standard deviation	0.34	0.32	0.45
CV (%)	9.31	10.12	12.92
Minimum	2.16	1.54	1.46
Maximum	5.36	5.10	5.31
Quartile 1	3.45	3.01	3.27
Quartile 3	3.87	5.01	5.31
Interquartile distance	0.42	0.39	0.48

^{N}: normal distribution of data; CV: Coefficient of variation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

The farm has already adopted some precision farming techniques, including yield mapping to generate interpolated maps. The observed yield values filtered were extracted from the pixels of the interpolated maps (Fig. 5) and constituted goals that the model tried to predict during the training of the algorithms. As seen in the descriptive analysis, we also observed on the map that the lowest yields were observed in field 2, and the greatest spatial variability of yield occurred in field 3. The field with the least variability of yield values was field 1.

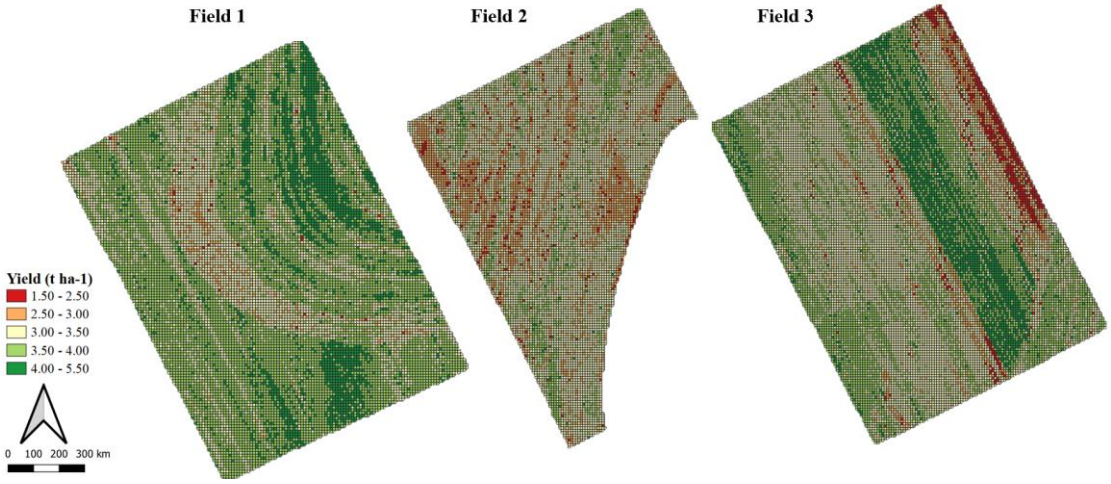

Figure 5. Representation of the yield maps observed in the fields in this study (2019 season) on farm in Santa Rita do Trivelato, Mato Grosso, Brazil.

Eight vegetation indices (SR, NDVI, NDRE, RGBVI, SAVI, IPVI, EVI and EVI2) were calculated on five dates of Sentinel-2A imaging corresponding to 03/08, 04/17, 05/07, 06/11 and 06/16 in the year of 2019 which will be referred to in the text as D1, D2, D3, D4 and D5.

In table 5 we observed the cotton development for each field of study, and noticed that cotton emergence occurred at 9, 6 and 17 days after sowing, respectively for fields 1, 2 and 3, which resulted in different accumulated growing degree days, resulting in the variation of thermal sum and phenological stage between fields, which will reflect on the variables measured in this study.

It was observed that the phenological stages for each image acquisition date varied between the formation of flower buds (B1, B6 and B9), flower opening/flowering (F5, F7, F8 and F10) and boll opening (C8, C9, C10, C12, C14), for fields 1, 2 and 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

Table 5. Information on sowing dates, emergence, phenological stage, harvest, cycle and final population as a function of image acquisition dates for the study fields on the farm in Santa Rita do Trivelato - MT.

	Field 1			Field 2			Field 3		
Sowing:	01/12/19			01/11/19			01/14/19		
Emerg.:	01/21/19			01/17/19			01/31/19		
Harvest:	07/21/19			07/31/19			07/27/19		
Cycle:	187 days			195 days			177 days		
Final Pop.:	55,476.00			53,390.33			57,390.33		
Area:	137 ha			102 ha			159 ha		
Date Images	**DAS(DAE)**	**FS**	**TS**	**DAS(DAE)**	**FS**	**TS**	**DAS(DAE)**	**FS**	**TS**
03/08/19	55(46)	B6	606	56(50)	B9	615	53(35)	B1	569
04/17/19	95(86)	F8	1015	96(90)	F7	1024	93(76)	F5	978
05/07/19	115(106)	C1	1223	116(110)	C3	1238	113(96)	F10	1191
06/11/19	150(141)	C10	1596	151(145)	C12	1605	148(131)	C8	1558
06/16/19	155(146)	C12	1652	156(150)	C14	1682	153(136)	C9	1636

Emerg.: Emergency; Final Pop.: final population; DAS: day after sowing; DAE: days after emergency; FS: phenological stage; TS: thermal sum. B: flower buds; F: flower opening; C: boll opening.

This variation caused different spectral responses in each field of study, as shown in Figure 6, being reflected in the vegetation indices in each field, as shown by the proportional Pearson correlation between yield and VI during the five analyzed dates.

All vegetation indices were positively significant at the 1% significance level and the correlations ranged from 16 to 44% (field 1), 18% to 52% (field 2), 12 to 31% (field 3), being the largest in field 2 and the smallest in field 3.

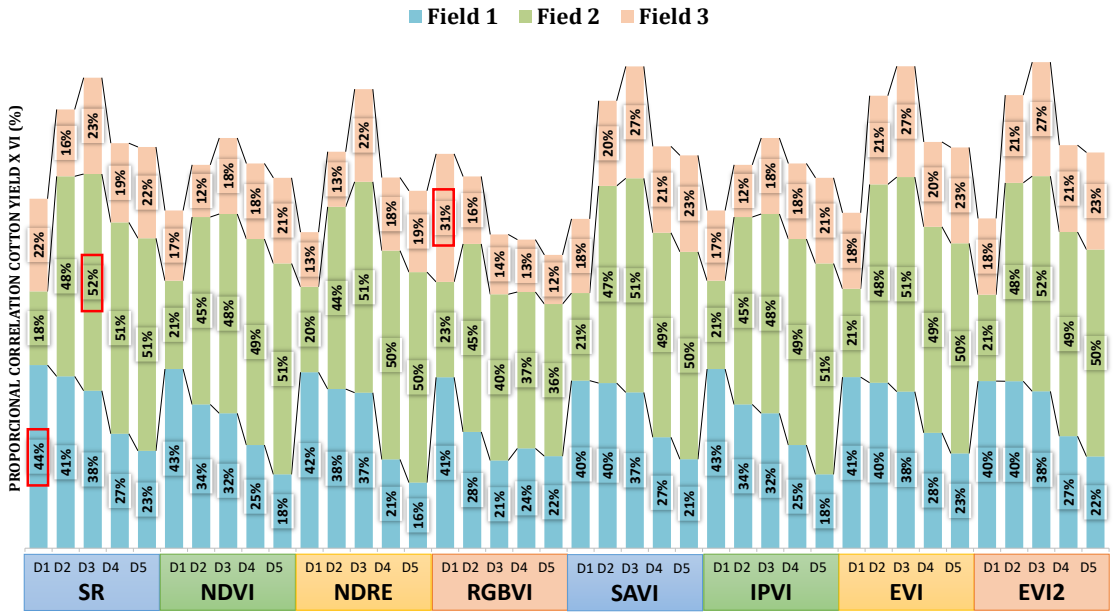

Figure 6. Pearson correlation Proportional between observed yield and vegetation index following the five image acquisition dates for fields 1, 2 and 3.

According to Cohen's classification (1988), which classifies as strong (0.50 and 1), moderate (0.30 to 0.49) and weak between (0.10 and 0.29) the relationship between two variables, thus, it is noted that only field 2 showed a strong correlation with values above 50%, and only on dates D2, D3, D4 and D5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

Through correlation, two important aspects were observed: 1) All indices had the same trend during all dates for the same field. 2) In field 1, the highest correlations were observed in D1, however, in field 2, D1 presented the smallest correlations. 3) In field 3, the correlations showed a weak correlation, and the highest values remained on D3. In other words, the values are correlated with yield, but they vary in each field for the same evaluated vegetation index.

3.2 Selection and standardization of predicting variables

The Stepwise selection method is straightforward and simple, as it starts with no candidate variable in the model until the complete analysis of the variables among themselves. The variable(s) that has the highest R-squared is selected or, at each step, the combination of candidate variables that most increases the R-squared is selected.

In field 1, the selected variable was SR in D1. In field 2, the most important variable was again SR, but in D3. And in field 3 the variable selected by stepwise was RGBVI in D1. Note that these SR and RGBVI vegetation indices were able to predict the cotton yield on dates 1, 2 and 3 with good accuracy (low RMSE).

3.4 Machine learning Algorithms and Cotton Yield Prediction

The algorithms demonstrated different accuracy and precision in predicting cotton yield in the evaluated scenarios and fields (Table 6) and the best performances are highlighted.

Scenario 1) All Data: For field 1 the best performance was MLP, with good precision and low error (R2: 0.47; RMSE: 0.24 t ha^{-1}) in predicting seed cotton yield at farm level (MLP) resulting in the lowest MAPE of all evaluated scenarios and fields, of only 5.11 %.

For fields 2 and 3, the best performance was obtained by RF, with the lowest RMSE obtained found in all scenarios and fields, of only 0.23 t ha^{-1}. Field 3 had the worst metrics, with RMSE from 0.1 to 0.2 t ha^{-1} more than the other fields. The MAPE was also higher reaching 2.82 % more when compared to fields 1 and 2, ranging from 7.54 to 9.86 %.

Table 6. Performance evaluation by RMSE and MAPE (in parentheses) for machine learning algorithms evaluated in cotton yield prediction in farm fields in Santa Rita do Trivelato, Mato Grosso, Brazil.

Field	Scenario	Input	RMSE (MAPE) - t ha^{-1} (%)				
			RLM	MLP	RF	KNN	AutoML
1	1. All Data	40	0.29 (6.06)	0.24(5.11)	0.25(5.29)	0.25(5.37)	0.28(6.09) GLM
	2. Stepwise	SR-D1	0.31(6.66)	0.30(6.45)	0.30(6.42)	0.28(6.21)	0.31(6.74) GLM
2	1. All Data	40	0.27(6.29)	0.25(5.98)	0.23(5.61)	0.24(5.70)	0.26(6.16) GLM
	2. Stepwise	SR-D3	0.28(6.72)	0.27(6.45)	0.27(6.40)	0.26(6.17)	0.28(6.51) DL
3	1. All Data	40	0.40(8.88)	0.36(7.86)	0.34(7.54)	0.36(7.78)	0.38(8.15) DL
	2. Stepwise	RGBVI-D1	0.43(9.86)	0.42(9.60)	0.42(9.59)	0.40(9.17)	0.42(9.63) DL

LR: Linear Regression, **MLP**: Multilayer Perceptron, **RF:** Random Forest Regressor, **KNN:** K-Nearest Neighbor Regressor; **AutoML**: Automatic Machine learning; **DL:** Deep Learning-1-AutoML; **GLM:** Generalized Linear Model-1-AutoML; **1 SR:** Vegetation index SR in date 1 (03/08/2019) with 606 of thermal sum, **3 SR:** SR in date 3 (05/07/2019) with 1238 of thermal sum; **1 RGBVI:** RGBVI in date 1 (03/08/2019) with 569 of thermal sum.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

Scenario 2: Variable selected stepwise. In this scenario, the best variables within the dataset with 40 inputs were the SR index on date 1 and 3, for fields 1 and 2 respectively. For field 3, the most important variable in response to yield was the RGBVI in D1.

It was noted that, regardless of the metrics, among a set of predicting variables that corresponded to five dates preceding the harvest, it is observed that the best was on date 1 and 3, that is, 141, 145 and 142 days before the harvest of fields 1, 2 and 3 respectively.

The SR index, despite being considered simple to calculate and being influenced by the atmosphere, showed the best correlations with the yield observed in fields 1 and 2 of this study on most dates. Field 3 showed different results, possibly due to the delay in emergence in relation to the other fields (11 and 8 days of delay in relation to field 1 and 2 respectively), which may have caused abiotic stress of the cotton in this field, resulting in in smaller values of thermal sum that reflected in greater variability, retroactive development, and low correlations of the vegetation indices.

The performance of the machine learning algorithms in each scenario (All Data and Stepwise selected variable) are presented in Fig. 7, corresponding to fields 1, 2 and 3 respectively.

Among the algorithms with hyperparameter adjustments, which are time-consuming adjustments analysis and that require great computational demand, comparing it with MLP and RF (more time consuming gridsearch, KNN was the one with the most satisfactory process, remaining only after the AutoML and LR.

We highlight an important result in Scenario 2 that, for training and testing using only a single input, the KNN Regressor algorithm showed the best performance in all fields of study, presenting the best performances at 48 DAE (D1) as well as 110 DAE (D3) which is equivalent to prediction preceding harvest by 143 days in D1, using a single index on a single date.

The best performances were again similar between field 1 and 2, with the best algorithm being the KNN using 1 input, predicting with 0.26 to 0.28 t ha-1 of RMSE, that is only 0.04 and 0.01 t ha-1 of RMSE. This difference is considered minimal because it is a commercial area with large-scale agricultural plots in Mato Grosso, noting that in scenario 2 we reduced 39 input variables and managed to predict with minimal difference in the RMSE. The KNN algorithm presented the best results in this study compared to the other methods tested.

According to Aparecido et al. (2020), which also evaluated this algorithm for cotton yield prediction in the Midwest region of Brazil, the positive point is that the KNN is a very simple algorithm, as it is based on the calculation of distance, that is, the algorithm considers that new data points are like training set data to do your ranking or regression. In regression, the KNN determines a value of a particular attribute from the neighboring data sample of the training set.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2021 doi:10.20944/preprints202112.0138.v2

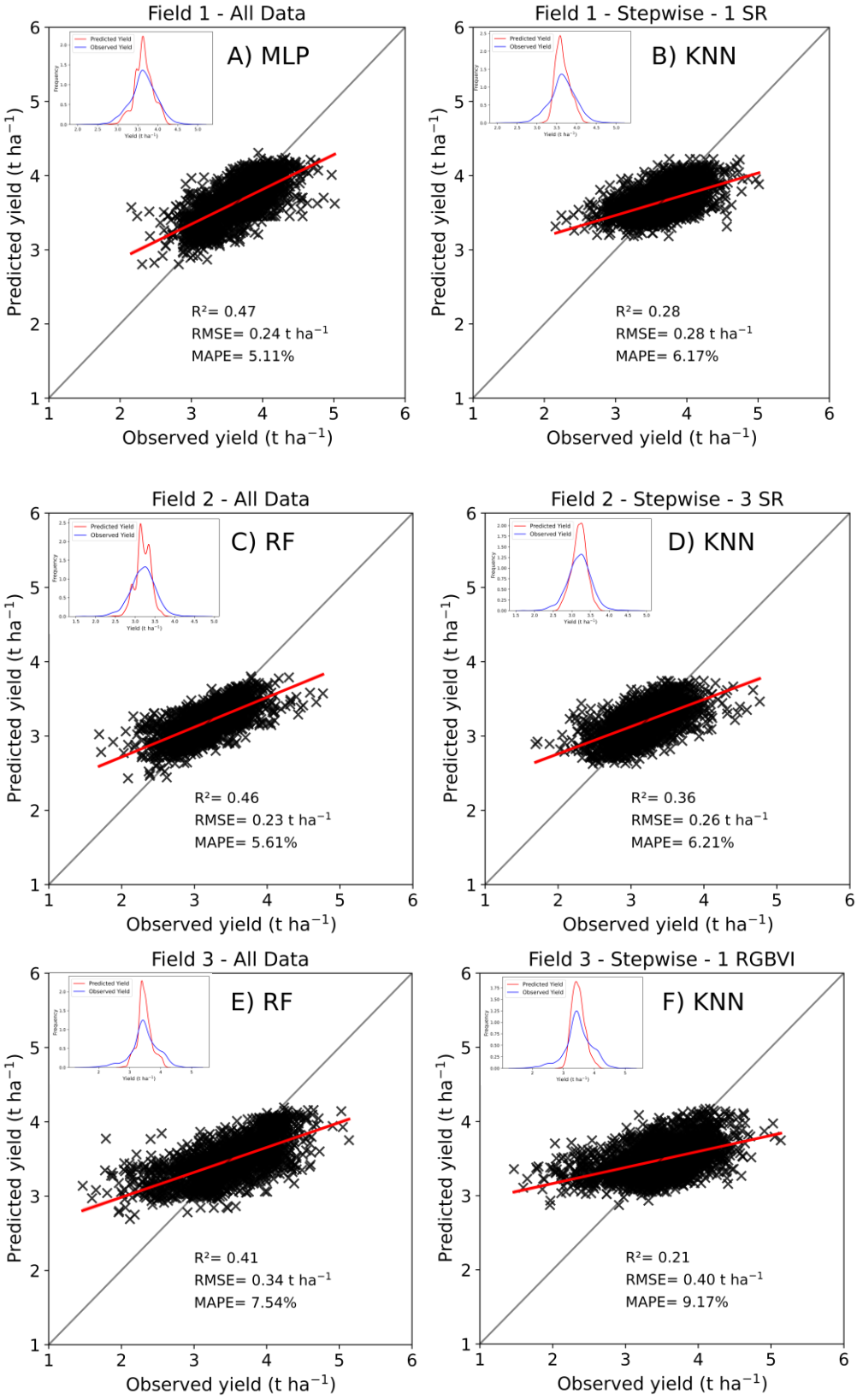

Figure 7. Relationship between observed and estimated values for cotton yield (t ha-1) data test for the best performances presented as FIELD 1: A) MLP in scenario 1: All Data; B) KNN for scenario 2: SR-D1; FIELD 2; C) RF for Scenario 1: All Data; D) KNN for Scenario 2: SR-D3; FIELD 3; E) RF for Scenario 1: All Data and D) KNN for Scenario 2: RGBVI-D1.

 doi:10.20944/preprints202112.0138.v2

Final Considerations

The cotton yield in fields 1 and 2 showed low spatial variability and the vegetation indices at all image acquisition dates showed a significant and positive correlation, with emphasis on the SR index, which had the highest correlation, and the index selected by Stepwise Regression in these two fields as the most important variable for predicting cotton yield.

Due to the difference in emergence dates, field 3 presented a different spectral response and probably different phenology that affected the VI among the evaluated fields. This fact influenced the correlation of variables with cotton yield, resulting in lower performance of the algorithms when compared to fields 1 and 2. Additional studies will be carried out with other models for other fields and study sites to confirm the results.

In addition to the best performance presented by the KNN algorithm, another important aspect observed in the study was that this algorithm required less computational demand in the adjustments of the hyperparameters and algorithm calibration, when compared to MLP and RF.

Conclusion

Scenario 1 (40 inputs) presented the best metrics in all fields of study.

The Multilayer Perceptron and Random Forest algorithms presented the best performances, noting that in this scenario we need all the predictive inputs generated throughout the entire crop cycle (approx. 180 days).

To achieve the objective of this study of identifying how early, which vegetation index and which machine learning algorithm it is possible to predict cotton yield at the farm level, we optimized the prediction and tested only the best IV in each field, and we detected that among the eight IVs, the SR (Simple Ratio) with KNN (K-Nearest Neighbor) algorithm showed the best performance, predicting with only 0.26 and 0.28 t ha^{-1} of RMSE and 5.20 % MAPE, anticipating with low error in ±143 cotton yield days.

In general, KNN had the best performance to achieve the objective in this study. The difference presented using 1 input out of 40 inputs in the prediction was minimal because it is a commercial area with large-scale agricultural plots in Mato Grosso, noting that in scenario 2 we reduced 39 input variables and predicted with minimal difference in the RMSE, enabling its use as a quick technique that will help predict cotton yield, saving time for planning, marketing and crop management decisions.

Acknowledgements

We are grateful for the support and partnership of Fazenda Mãe Margarida, and all professionals belonging to Terra Santa Agro Group for making available the fields and agronomic data for conducting this study. Authors also acknowledge the projects USDA (LAB94427) and Louisiana Cotton Board and Cotton INC for the PhD sandwich of Francielle Morelli Ferreira financial support.

References

Aparecido, L.E. de O.; de Meneses, K.C, de Souza Rolim, G.R.; Carvalho, M.J.N.; Pereira, W.B.S., da da Silva, P.A.; Santos, T. S. & Moraes, J.R.S.C.M. (2020) Algorithms for forecasting cotton yield based on

climatic parameters in Brazil, Archives of Agronomy and Soil Science, Doi: 10.1080/03650340.2020.1864821 .

Arnold, C.Y. The determination and significance of the base temperature in a linear heat unit system. Proceedings of the American Society for Horticultural Science, Alexandria, v. 74, n.1 p. 430-445, 1959.

Badnakhe, MR.; Durbha, SS, Jagarlapudi, A.; Gade RM (2018) Evaluation of Citrus Gummosis disease dynamics and predictions with weather and inversion-based leaf optical model. Comput Electron Agric 155:130–141

Barnes, E.M. et al. Coincident detection of crop wáter stress, nitrogen status and canopy density using ground-based multispectral data. In: INTERNATIONAL CONFERENCE ON PRECISION AGRICULTURE, 2000, Bloomington. Madison: ASA: CSSA: SSSA, 2000.

Bendig, J. et al. Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. International Journal of Applied Earth Observation and Geoinformation, v. 39, p. 79–87, 2015. Doi: 10.1016/j.jag.2015.02.012.

Birth, G.S., McVey, G.R., 1968. Measuring the color of growing turf with a reflectance spectrophotometer. Agron. J. 60 (6), 640–643. Doi: 10.2134/agronj1968.00021962006000060016x.
Breiman, L. Random Forests. Machine Learning 45, 5-32 (2001). Doi: 10.1023/A:1010933404324.

Cohen, Jacob. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Conab, Companhia Nacional de Abastecimento. Acomp. safra bras. grãos, v. 7 - Safra 2019/20 - n. 5 - Quinto levantamento, Brasília, p. 1-25 Feb. 2020. Access on: <https://www.conab.gov.br/info-agro/safras/graos>.

Crippen, R. E. Calculating the Vegetation Index Faster. Remote Sensing of Environment, V. 34, P. 71-73, 1990. Doi: 10.1016/0034-4257(90)90085-Z

Cornell, J. A.; Berger, R. D. Factors that Influence the Value of the Coefficient of Determination in Simple Linear and Nonlinear Regression Models. Phytopathology, v. 77, n. 1, p. 63, 1987.

Embrapa. Empresa Brasileira de Pesquisa Agropecuária. Sistema Brasileiro de Classificação de Solos. 3 ed. Brasília, 2013. 353 p.

Everingham Y, Sexton J, Skocaj D, Inman-Bamber G. 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development. 36(2):27. Doi:10.1007/s13593-016-0364-z.

Fue, K.G., Porter, W.M., Rains, G.C., 2018. Deep Learning based Real-time GPU-accelerated Tracking and Counting of Cotton Bolls under Field Conditions using a Moving Camera, In: 2018 Detroit, Michigan July 29 - August 1, 2018. Presented at the 2018 Detroit, Michigan July 29 - August 1, 2018, American Society of Agricultural and Biological Engineers. Doi: 10.13031/aim.201800831.

Huete, A. R. A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment, New York, v. 25, n. 3, p. 295-309, 1988. Doi: 10.1016/0034-4257(88)90106-X

Huete, A. R., Liu, H. Q., Batchily, K., & Leeuwen van, W. (1997). A comparison of vegetation indices over a global set of TM images for EOS-MODIS. Remote Sensing of Environment, 59, 440−451.

Kaul, M. Hill, R.L, Walthall, C. Artificial neural networks for corn and soybean yield prediction, Agricultural Systems, v. 85, n. 1, p. 1-18, 2005, ISSN 0308-521X, Doi: 10.1016/j.agsy.2004.07.009.

Köppen, W.; Geiger, R. Klimate der Erde. Gotha: Verlag Justus Perthes. 1928. Wall-map 150cmx200cm.

LeDell, E.; Poirier. S. *H2O AutoML:* Aprendizado de máquina automático escalonável. 7º ICML Workshop on Automated Machine Learning (AutoML), July 2020. URL https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf .

Menegatti, L.; Molin, J. P. Remoção de erros em mapas de produtividade via filtragem de dados brutos. Revista Brasileira de Engenharia Agrícola e Ambiental, Campina Grande, v.8, n.1, p.126-134, 2004.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weis s R, Dubourg V. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res. 12: 2825–2830. others.

QGIS, Equipe de Desenvolvimento, 2009. Sistema de Informação Geográfica QGIS. Fundação Geoespacial de Código Aberto. Access on: http://qgis.org

Reddy, V. R.; Hodges, H. F.; Mccarty, W. H.; Mckinnon, J. M. Weather and cotton growth: Present and Future. Mississippi Agr. & Forestry Exp. Sta., Mississippi State University, Starkville, MS.1996.

Rosolem, C. A. Crescimento do algodoeiro. p. 105. In: BÉLOT, J.; VILELA, P. Manual de boas práticas de manejo do algodoeiro em Mato Grosso. Cuiabá, 461 p. 4 ed. 2020. Access on: http://www.casadoalgodao.com.br/images/publicacoes/manualdeboaspraticas2020-4ed-vf-web.pdf.

Rouse Jr, J.W., Haas, R., Schell, J., Deering, D., 1973. Monitoring vegetation systems in the Great Plains with ERTS. In: NASA Special Publication 351. pp. 309–317.

Singh, A., Ganapathysubramanian, B., Singh, A.K., Sarkar, S., 2016. Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci. 21 (2), 110–124. Doi: 10.1016/j.tplants.2015.10.015

Snider, J.; Kawakami E.M. Efeito da temperatura no desenvolvimento do algodoeiro. *In:* Echer. F. R. O algodoeiro e os estresses abiótico: Temperatura, luz, água e nutrientes- Cuiabá (MT), 123 p. 2014.

Stackhouse, P. Prediction of worldwide energy resource. Hampton: NASA Langley Research Center, 2010. Access on: https://power.larc.nasa.gov/.

Tedesco-Oliveira, D., Silva, R. P., Maldonado-Jr, W., Zerbato, C. Convolutional neural networks in predicting cotton yield from images of commercial fields. Computers and Electronics in Agriculture 171 (2020) 105307. Doi: 10.1016/j.compag.2020.105307

Tukey, J.W. Exploratory data analysis. 1 ed. Reading, Massachusetts, v.1, n.3, 1997.

USDA – United State Department of Agriculture. World Agricultural Production. Circular Series WAP 2-20 February 2020. Disponível em: <https://apps.fas.usda.gov/psdonline/circulars/production.pdf >. Access on: 22 fev. 2020

Xu, R., Li, C., Paterson, A.H., Jiang, Y., Sun, S., Robertson, J.S., 2018. Aerial images and convolutional neural network for cotton bloom detection. Front. Plant Sci. 8, 1–17. Doi: 10.3389/fpls.2017.02235

Yeom, J., Jung, J., Chang, A., Maeda, M., Landivar, J., 2018. Automated open cotton bol detection for yield estimation using Unmanned Aircraft Vehicle (UAV) Data. Remote Sens. 10, 1–20. Doi: 10.3390/rs10121895