Supervised Learning Models for Potato Yield Prediction in Commercial Fields

Samira Luns Hatum de Almeida[1], Thiago Caio Moura Oliveira [1], Bruno Ripa Baptista[1], Vinícius Carreira dos Santos[1], Jarlyson Brunno Costa Souza [1], Rouverson Pereira da Silva[1]

[1] Department of Engineering and Mathematical Sciences, São Paulo, State University (UNESP), Jaboticabal, São Paulo, Brazil, - samira.lh.almeida@unesp.br; thiago.caio@unesp.br; bruno.ripa@unesp.br; vs.carreira@unesp.br; jarlyson.brunno@unesp.br; rouverson.silva@unesp.br

Keywords: Machine Learning, orbital sensor, vegetation index, *Solanum tuberosum* L.

ABSTRACT

The fourth most consumed food in the world, potatoes are a crop of great importance for food security. To fully explore their potential, it is essential to expand productivity studies, enabling yield to be predicted more efficiently. This study aimed to integrate remote sensing data and machine learning algorithms to develop yield prediction models for potatoes across two cultivars. For this purpose, yield samples were collected in a commercial field cultivated with two cultivars: Asterix and Markies. The sampling grid consisted of 160 points, with 80 points assigned to each cultivar. For sampling, a frame was used within which all tubers were collected and weighed. Satellite imagery from the PlanetScope platform, captured 25 days before crop desiccation, was used to calculate vegetation indices: NDVI, SAVI, and GNDVI. To develop the models, machine learning algorithms were employed: K-Nearest Neighbors, Support Vector Regression, Ridge Regression, Random Forest, XGBoost, Gradient Boosting, and Decision Tree — all of which are supervised learning methods. Given the differences between cultivars, we opted to generate separate models for each, as well as a combined model using both cultivars, in order to assess which dataset would yield more accurate predictions. The dataset was split into 80% for model training and 20% for testing. The Random Forest algorithm stood out by producing the most accurate models when using the combined cultivar dataset, with a mean absolute error (MAE) of 4.58 $t \cdot ha^{-1}$. These results highlight the potential of integrating remote sensing (RS) and machine learning algorithms — particularly Random Forest — for predicting potato yield. Further studies aimed at expanding the dataset will be essential for obtaining more precise and reliable models.

1. INTRODUCTION

The modernization of agriculture aims to make farming practices more sustainable and capable of meeting the growing demand for food. In this context, advancing research on staple crops, such as potato (*Solanum tuberosum* L.), is of great importance, especially studies related to yield prediction. Understanding crop productivity helps to analyze spatial variability within fields and to guide post-harvest and marketing strategies. When carried out in a traditional and manual manner, this process is slow and inefficient. Resorting to indirect methods of obtaining the variable of interest through digital agriculture tools is a way to make access to this type of information faster and more scalable.

Within the scope of digital agriculture, remote sensing (RS) presents itself as a decision-support science, as it enables the extraction of data from vegetated surfaces without direct contact, providing indicators of crop health, biomass quantity, among others. The expansion of studies in this field demonstrates that RS products have a direct relationship with crop yield, as is the case with soybean and maize. In crops with subterranean growth, in which the economically important products develop below the soil surface, Souza et al. (2025) demonstrated that, despite the impossibility of visually observing the fruits, the vegetation indices EVI, SAVI, and SR stood out as consistent variables for predicting peanut yield. Tesdesco et al. (2021) highlighted the use of NDVI and GNDVI indices, derived from Sentinel imagery, for predicting sweet potato yield.

In potato cultivation, studies have been developed focusing on yield prediction, but they face challenges when applied to Brazilian conditions, as observed by Gomes et al. (2019) who used Sentinel-2 satellite imagery. Although the images are free of charge, cloud cover during the summer potato season in Brazil limits image availability, compromising the applicability of the proposed model. The adoption of satellites with better temporal resolution, such as PlanetScope, may expand the adoption of the method for yield prediction.

In the agricultural context, alongside remote sensing, artificial intelligence has gained prominence due to its high capacity for pattern recognition, with wide adoption in pest and disease detection, precision irrigation, automated harvesting, and prediction tasks (Peñailillo et al., 2024). The integrated use of these technologies enables complex activities to be conducted in a more autonomous and scalable manner.

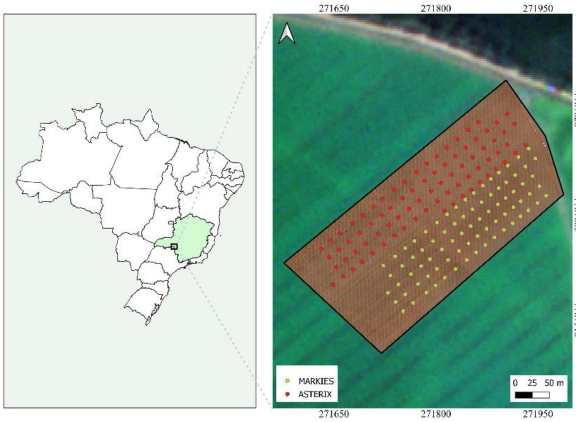
Given the above, the hypothesis is that the integration of spectral data with high temporal resolution and machine learning tools contributes to the development of more robust models for potato yield prediction. Therefore, the objective of this study was to use orbital remote sensing combined with machine learning algorithms to develop prediction models for the potato cultivars Markies and Asterix during the summer growing season and to identify which algorithm presented the best error metrics.

2. MATERIAL AND METHODS

2.1 Field Data

The experiment was conducted in a commercial field cultivated with two potato cultivars in the municipality of Sacramento, Minas Gerais, Brazil (Figure 1). Two plots were designated, one for each cultivar—Markies and Asterix—both intended for seed potato yield. The sampling grid consisted of 160 points, with 80 points allocated to each cultivar, arranged in a 25 × 25 m spacing.

In situ sampling assessments were carried out prior to harvest. A digger was used to loosen the ridges and tubers. Subsequently, a 2 m² frame was placed over the ridges, and all the tubers within the frame were collected and weighed.



Figurea 1. Study field in Sacramento – Minas Gerais State, Brazil.

2.2 Orbital Images

Monitoring of the experimental areas was carried out using satellite images from the PlanetScope platform, acquired by the SuperDove sensor. This sensor provides daily imagery with a spatial resolution of 3 meters and a radiometric resolution of 12 bits. The spectral resolution used included four bands: blue (455–515 nm), green (500–590 nm), red (590–670 nm), and near-infrared (780–860 nm).

For yield prediction, an image acquired approximately 25 days prior to crop desiccation was selected. This date was chosen due to the high vegetative vigor observed and because it corresponded to a phenological stage after flowering—a phase that could affect the accuracy of predictive models due to the distinct coloration of the flowers in the evaluated cultivars.

Using QGIS 3.16, the previously georeferenced sampling points were identified, and a 3-meter buffer was generated around each point. The mean reflectance value for each spectral band was then extracted within each buffer. Subsequently, the vegetation indices NDVI (Rouse et al., 1974), SAVI (Huete et al., 1992), and GNDVI (Gitelson et al., 1996) were calculated.

2.3 Data Analysis

As input variables for the model, we used spectral bands and the vegetation indices. Data analysis was performed using the Python programming language. Initially, the dataset was normalized using the Min-Max method. For model development, supervised machine learning algorithms were employed, including K-Nearest Neighbors, Support Vector Regression, Ridge Regression, Random Forest, XGBoost, Gradient Boosting, and Decision Tree. Hyperparameter optimization was carried out using the GridSearchCV method available in the Scikit-learn library (table 1). These values represent the configurations that consistently yielded the best performance and thus provide a reliable benchmark for future studies on potato yield prediction.

Table 1. Optimal hyperparameters (mode) selected for each algorithm using Nested Cross-Validation.

Model	Optimal Hyperparameters (mode)
Gradient Boosting	learning_rate=0.01, max_depth=3, n_estimators=100
Ridge Regression	alpha=10.0
Random Forest	bootstrap=True, max_depth=10, max_features='sqrt', min_samples_leaf=2, min_samples_split=2, n_estimators=200
XGBoost	learning_rate=0.01, max_depth=3, n_estimators=100
Support Vector Regression	C=0.1, epsilon=0.1, kernel=linear
K-Nearest Neighbors	n_neighbors=10, p=2, weights=uniform
Decision Tree	max_depth=10, min_samples_leaf=1, min_samples_split=2

Individual models were generated for each cultivar, as well as a combined model using data from both. The dataset was split into 80% for training and 20% for testing. Model performance was evaluated using the mean absolute error (MAE) and Mean Absolute Percentage Error (MAPE) metric.

3. RESULTS AND DISCUSSION

The initial results of the yield analysis for the potato cultivars are presented in Figure 2 through a boxplot, which highlights the distribution and dispersion of the sampled data. The cultivars Asterix and Markies showed similar average yields of 22.30 and 23.58 $t.ha^{-1}$, respectively. However, the Asterix cultivar exhibited greater variability in yield data, reflecting a wider dispersion among the recorded values. For the Markies cultivar, 50% of the sampled yield was concentrated between 20 and 27 $t.ha^{-1}$, indicating a more consistent distribution. This analysis of the yield variable is essential for understanding the errors obtained from the yield prediction models presented below.

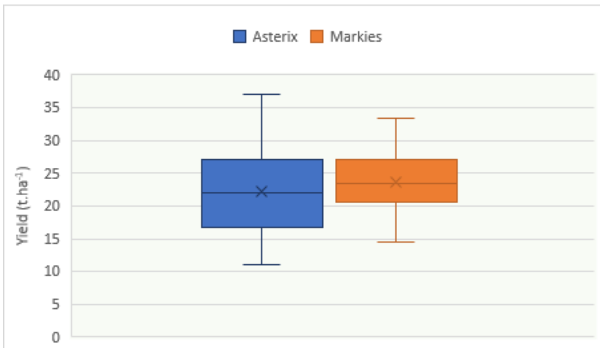


Figure 2. Boxplot representing the field-observed yield behavior of Asterix and Markies potato cultivars.

Yield predictions, as described in the methodology, were performed considering the cultivars both individually and combined. The following figures (Figures 3, 4, and 5) are bar charts presenting the accuracy metrics MAE and MAPE. In Figure 3, the lowest MAE was obtained by the model generated using the K-Nearest Neighbors (KNN) algorithm, with 4.76 $t{\cdot}ha^{-1}$ and a mean absolute percentage error of 21.4%. The Support Vector Machine, Ridge Regression, and Random Forest algorithms also stood out, as they did not exceed a MAPE of 23% and an MAE of 5 $t{\cdot}ha^{-1}$.

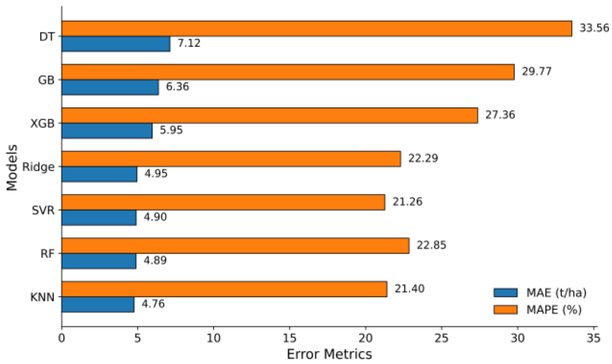


Figure 3. Error metrics for the model developed for the Asterix cultivar. DT: Decision Tree; GB: Gradient Boosting; XGB: XGBoost; Ridge: Ridge Regression; SVR: Support Vector Regression; RF: Random Forest: KNN: K-Nearest Neighbors.

For the Markies cultivar, the prediction errors were higher than those observed for the Asterix cultivar, even though Markies exhibited a more homogeneous yield behavior, as shown in Figure 1. The lowest mean absolute error (MAE) was 6.19 $t \cdot ha^{-1}$, obtained using the Gradient Boosting (GB) algorithm. Despite the higher average yield—which complicates comparisons between cultivars based on MAE—the mean percentage error demonstrated better algorithm performance in predicting the yield of the Asterix cultivar. The mean absolute percentage error (MAPE) for Markies ranged from 35.39% to 45.81%, indicating that, even with lower variability, the algorithms struggled to make highly accurate predictions.

This finding suggests that, beyond yield variability, other factors may influence model performance, such as spectral band saturation, which reduces the sensitivity of vegetation indices in dense canopies, and low variability among predictor variables, which limits the algorithms' ability to detect meaningful patterns for yield prediction. This aspect was evidenced by Morlin et al. (2019), who observed this behavior mainly in vegetation indices that use the red band, such as NDVI, which showed saturation in crops with high leaf density.

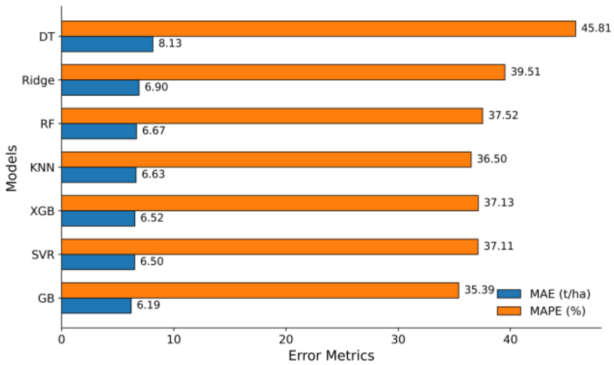


Figure 4. Error metrics for the model developed for the Markies cultivar.

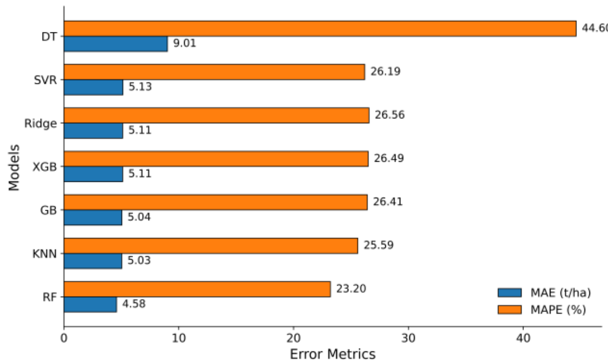


Figure 5. Error metrics for the model developed for the Markies and Asterix cultivar.

In the combined analysis of the Asterix and Markies cultivars, Random Forest (RF) stood out as the best-performing algorithm, presenting the lowest mean absolute error (MAE = 4.58 $t \cdot ha^{-1}$) and the lowest mean absolute percentage error (MAPE = 23.20%). This algorithm can be highlighted for its structure based on multiple decision trees, which enables it to capture nonlinear patterns and makes it less sensitive to outliers. These results are consistent with those reported by Imtiaz et al. (2025), who also found lower errors in potato yield prediction using the Random Forest algorithm.

Based on the analyzed results, it is possible to infer that, under the conditions of this study, the model generated from the combined data of both cultivars using the Random Forest algorithm is the most suitable for potato yield prediction. The combination of the two cultivars in a single model demonstrates the feasibility of constructing more generalist models. Furthermore, in a practical application context, models that integrate different cultivars eliminate the need for prior identification of the cultivar, thereby increasing practicality and applicability. The genotype–environment interaction is a factor that can influence the generalization potential of prediction models. This aspect was evidenced by Souza et al. (2025), who observed that the inclusion of new areas and cultivars contributed to improving both the precision and the accuracy of peanut yield prediction models.

Considering that potato harvesters are not equipped with yield monitors and that this variable is not commonly measured in the field, the accuracy metrics obtained suggest that the proposed method has strong potential for real-world application. Incorporating data from fields in different regions and with various cultivars may further enhance the precision and accuracy of the models. More consistent predictions for potato crops were observed in Canada. However, these results are directly related to the larger amount of data available, since potato harvesters in that region are equipped with yield sensors, which enabled a much broader and more detailed data collection (Imtiaz et al., 2025).

4. CONCLUSION

The results of this study indicate the potential of integrating machine learning algorithms with the spectral behavior of vegetation for predicting potato yield, even across different cultivars, with particular emphasis on the Random Forest algorithm. Although the model developed for the Asterix cultivar showed higher accuracy, adopting a model that encompasses multiple cultivars increases its applicability and reduces operational complexity. The continuation of these studies, with efforts to expand the dataset, will be essential for obtaining more precise and accurate prediction models.

Acknowledgements

To the São Paulo Research Foundation (FAPESP - grant numbers 2023/14041-2, 2023/08119-9, 2022/16084-8, and 2021/06029-7) and the National Council for Scientific and Technological Development (CNPq) for their financial support.

References

Fuentes-Peñailil lo, F., Gutter, K., Vega, R., & Silva, G. C. (2024). Transformative technologies in digital agriculture: Leveraging Internet of Things, remote sensing, and artificial intelligence for smart crop management. Journal of Sensor and Actuator Networks, 13(4), 39.

Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS (introducing GNDVI). Remote Sensing of Environment, 58, 289–298.

Gómez, D., Salvador, P., Sanz, J., & Casanova, J. L. (2019). Potato yield prediction using machine learning techniques and sentinel 2 data. Remote Sensing, 11(15), 1745.

Huete, A. R., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (1992). A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment, 25, 53–69.

Imtiaz, F., Farooque, A. A., Randhawa, G. S., Wang, X., Esau, T. J., Garmdareh, S. E. H., & Acharya, B. (2025). Optimizing potato yield mapping and prediction: Integrating satellite-based remote sensing and machine learning for sustainable agriculture. Computers and Electronics in Agriculture, 237, 110636.

Morlin Carneiro, F., Angeli Furlani, C. E., Zerbato, C., Candida de Menezes, P., da Silva Gírio, L. A., & Freire de Oliveira, M. (2020). Comparison between vegetation indices for detecting spatial and temporal variabilities in soybean crop using canopy sensors. Precision Agriculture, 21(5), 979-1007.

Rouse, J. W., Jr., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. Proceedings of the 3rd Earth Resources Technology Satellite (ERTS) Symposium, 1, 309–317.

Souza, J. B. C., de Almeida, S. L. H., de Oliveira, M. F., Carreira, V. D. S., de Filho, A. L. B., dos Santos, A. F., & da Silva, R. P. (2025). Generalization of peanut yield prediction models using artificial neural networks and vegetation indices. Smart Agricultural Technology, 11, 14.

Tedesco, D., de Oliveira, M. F., dos Santos, A. F., Silva, E. H. C., Rolim, G. de S., & da Silva, R. P. (2021). Use of remote sensing to characterize the phenological development and to predict sweet potato yield in two growing seasons. European Journal of Agronomy, 129, 126337.