


Artificial Intelligence in Agricultural Management: Use of Random Forest Models for the Prediction of Seed Production and Reservation in Brazil

 <https://doi.org/10.56238/sevened2024.023-006>

Eduardo Silva Vasconcelos¹, Leandro Aureliano da Silva², Débora Vasconcelos Melo³, Adriano Dawison de Lima⁴, Luiz Fernando Ribeiro de Paiva⁵ and Cleiton Silvano Goulart⁶

ABSTRACT

This study addresses the application of Artificial Intelligence (AI) models, more specifically random forest, for the prediction of seed production and reserve in Brazilian agriculture. The main objective is to contribute to the advancement of resource management and planning, a critical action to increase the efficiency and sustainability of the sector. The work is highlighted by the importance of understanding the role of AI in optimizing agricultural practices, providing a framework for future research at the intersection between AI technologies and agriculture. Methodologically, the study implemented a rigorous data collection and processing process provided by the Ministry of Agriculture and Livestock of Brazil, covering the harvests from 2016/2017 to 2023/2024. Data cleansing preceded the transformation of categorical variables through one-hot coding and subsequent splitting of the dataset into 80% for training and 20% for testing. Using the scikit-learn library, a random forest model was configured and evaluated, employing validation techniques such as training/test split and cross-validation, in addition to mean square error (MSE) and coefficient of determination (R^2) metrics to measure the accuracy and effectiveness of the model. The results indicate a moderate to strong positive correlation between the variables of time and number of seeds reserved for both growing periods, Safra and Safrinha. However, the analyses pointed to annual variability and differentiated confidence in predictions between periods, suggesting the influence of additional factors and the need for adaptive models. The concentration of production in a few cultures was identified as a potential risk, suggesting that diversification is key to the resilience of the sector. The generalizability of the model was evaluated, and the phenomenon of overfitting was considered a possibility given the variations in accuracy between the training and test data. This study reinforces the transformative potential that AI models, such as the random forest, possess for agricultural prediction and management, opening doors for future improvements and providing valuable subsidies for data-driven strategic decisions in the agricultural sector.

Keywords: Artificial Intelligence in Agriculture, Random Forest Models, Seed Production Prediction, Sustainable Agricultural Management, Agricultural Data Analysis.

¹ Doctor of Science - Information Systems
Federal Institute of Goiás

² Doctor of Science - Information Systems
University of Uberaba

³ Master in Management of Organizations
Federal University of Catalonia

⁴ Doctor in Agronomy
University of Uberaba

⁵ Doctor of Education
University of Uberaba

⁶ Master's Degree in Physics
University of Uberaba



INTRODUCTION

Agriculture is facing increasing challenges due to demands for efficiency and sustainability. This study addresses the importance of Artificial Intelligence (AI) in the prediction of seed production, crucial for planning and resource management in the agricultural sector. We focused on the development and evaluation of an AI model based on random forest, with the aim of improving decision-making practices and operational efficiency in agriculture.

The growing need for optimization in agricultural production and resource management has driven the adoption of advanced technologies, such as Artificial Intelligence (AI). In this context, seed production prediction emerges as a vital field of study, offering valuable insights for agricultural planning and the sustainability of the sector. The present work explores the application of AI models, specifically random forest algorithms, to forecast seed production, in order to improve resource control and management in agriculture.

The scientific importance of this study lies in its contribution to the understanding of how AI models can be applied to improve agricultural efficiency and productivity, as well as providing a framework for future research at the intersection between AI and agriculture. The integration of AI in agriculture, especially for seed production prediction, has the potential to transform resource management and decision-making in the sector. This study demonstrates the feasibility and value of using random forest models for this purpose, highlighting the importance of such approaches for future advances in sustainable and efficient agriculture.

This study has the general objective of developing and evaluating an Artificial Intelligence model based on random forest techniques, aimed at predicting seed production and reserve in the agricultural context. We aspire to contribute significantly to the improvement of resource management and planning in the sector, a prevailing need for increasing agricultural efficiency and sustainability. The study aims not only to improve strategic planning and effectiveness in resource management, but also to boost accuracy and sustainability in agriculture, providing fundamental guidelines for the development of advanced management and planning strategies in the agricultural sector.

In specific terms, the study seeks to: (1) employ a multifaceted analysis of seed production and reserve patterns in Brazilian agriculture using statistical techniques, including timeline plots, Pareto plots, scatter plots, linear regression, and analytical tables; (2) to investigate the applicability and effectiveness of the random forest algorithm, used in this work, for the modeling and prediction of seed production, to evaluate the performance of the model in terms of accuracy, efficiency and generalizability, and to examine the extent to which predictions based on Artificial Intelligence contribute to optimize planning and resource control in agriculture.



These objectives seek to synthesize the practical and theoretical relevance of Artificial Intelligence models, with emphasis on random forests, in the analysis and interpretation of seed production and reserve patterns, highlighting their applicability to elucidate trends, identify variations and discern critical factors that influence these dynamics.

THEORETICAL FRAMEWORK

The literature review contemplates a series of previous research that explores the application of Artificial Intelligence (AI) in agriculture, particularly emphasizing the use of machine learning methods, such as random forests, in the context of production prediction. This review synthesizes both the significant contributions to technological advancement in the sector and specific studies demonstrating the effectiveness of these techniques.

In the field of technological contributions and advances, Vieira Filho and Silveira (2012) discuss innovation in agriculture, underlining the complexity of the sector and the cruciality of technological knowledge and learning for its development. They argue that innovation is an essential catalyst for improving agricultural practices and for the sustainability of production. Adama Brazil (2024) and Climate FieldView (2023) expand on this discussion by addressing the fundamental importance of AI in the modernization of agriculture. These sources illuminate specific applications of AI, such as remote sensing and the analysis of large volumes of data in real time, which have transformed the ability to monitor and manage agriculture. Rehagro (2022) complements this vision, highlighting the virtues of AI in the monitoring and optimization of agriculture, emphasizing its role in promoting sustainable practices and in increasing operational efficiency in the field.

In terms of specific studies, research carried out by Mourtzinis et al. (2021), Saleem, Potgieter, and Arif (2021), and Gadotti et al. (2022) shows the applicability and efficiency of machine learning techniques in the agricultural sector. These studies provide robust evidence of the superiority of methods such as random forests and deep learning in performing specific predictive tasks, pointing to markedly superior performance compared to more traditional statistical modelling techniques. They emphasize the relevance of AI-based approaches as highly competent predictive tools, capable of handling the complexity of agricultural data and producing accurate predictions, which are critical for planning and informed decision-making in the sector.

Therefore, the literature reviewed underscores the steady progression and potential impact of AI on agriculture, indicating a promising horizon for the continued integration of these advanced technologies into contemporary agricultural practices. These advances not only corroborate the position of AI as an essential pillar in agricultural modernization, but also endorse the need for continued research to fully explore the capabilities and benefits of random forests and machine learning in agricultural prediction.



METHODOLOGY

The methodology used to create a "random forest" model aims to estimate the amount of seeds that will be produced and reserved in agriculture. It is a method that uses computing and statistics to analyze data on past agricultural production, in order to make reliable predictions about the amount of seed reserve that will be generated in the future. This process helps farmers and managers to better plan their activities, anticipating the need for seeds and optimizing resources for more efficient and sustainable production. Using data provided by the Ministry of Agriculture and Livestock of Brazil, this study implements a rigorous approach, from data collection to processing, culminating in the development of a robust model for predictive analytics.

This study takes a quantitative approach, using historical seed production data to develop a random forest model. The effectiveness of the model is evaluated using statistical metrics, including mean square error (MSE) and coefficient of determination (R^2).

STUDY DESIGN

The study was designed to apply the random forest model, known for its ability to process large data sets and avoid overfitting. The objective was to evaluate the effectiveness of this model in predicting seed production, incorporating variables such as period, region, seed species, and planted area.

DATA COLLECTION

The data was obtained from the Inspection Management System of the Ministry of Agriculture and Livestock of Brazil, covering detailed information from the 2016/2017 harvest to 2023/2024. The exclusion of data prior to 2016/2017 was necessary due to identified inconsistencies.

DATA PREPARATION

The preparation included cleaning and transforming the data, adopting techniques such as one-hot coding for categorical variables. The division of the data into training and test sets followed the standard 80/20 ratio.

DEVELOPMENT OF THE MODEL

The random forest model was implemented in Python, using the scikit-learn library. The selection of hyperparameters was based on standard practices, considering the balance between performance and compute time.



IMPLEMENTATION AND EVALUATION OF THE RANDOM FOREST MODEL

The methodology used in the implementation and evaluation of the random forest model is detailed, reflecting the various stages necessary for the prediction analysis of seed production in agriculture. This methodology is structured to allow for a detailed and judicious future analysis of the model's results.

In the training phase, critical procedures were carried out, starting with the initial preparation of the data, where coding techniques were applied to adapt the categorical variables to the numerical format required by the model:

```
from sklearn.preprocessing import OneHotEncoder
# One-Hot coding of categorical variables
encoder = OneHotEncoder(sparse=False)
X_encoded = encoder.fit_transform(X[['UF', 'Seed Species']])
```

The data were then divided into training and test sets, with 80% going to training and 20% to testing, essential for subsequent evaluation of the model's ability to generalize to new data:

```
from sklearn.model_selection import train_test_split
# Splitting the data into training and testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

With the data prepared, the construction of the random forest model was carried out using the scikit-learn library, configuring hyperparameters such as the number of trees in the model:

```
from sklearn.ensemble import RandomForestRegressor
# Random forest model construction
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

To ensure the reliability of the model's predictions, a rigorous validation strategy was implemented, using not only the training/test split but also cross-validation to provide a more robust evaluation:

```
from sklearn.model_selection import cross_val_score
# Cross-validation application
scores = cross_val_score(model, X, y, cv=5)
```

The evaluation criteria included metrics such as the Mean Square Error (MSE) and the Coefficient of Determination (R^2), fundamental to measure the effectiveness and accuracy of the model:



```
from sklearn.metrics import mean_squared_error, r2_score
# Evaluating the model in the test data
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

These methodological steps ensure that the model is well adjusted to the data, making it possible to make reliable and accurate predictions. The generalizability of the model was evaluated through the performance comparison between the training and test sets, crucial to detect overfit or underfit.

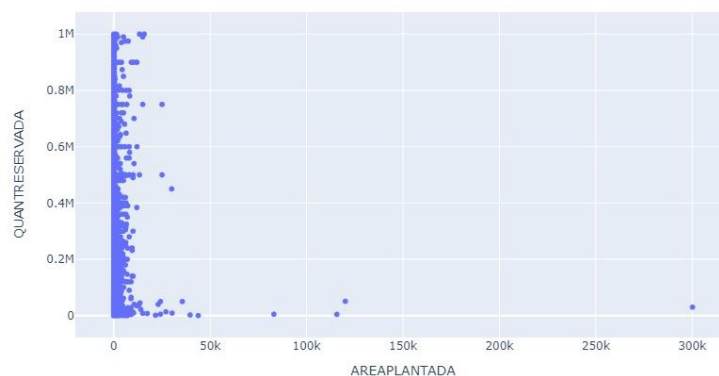
RESULTS AND DISCUSSIONS

ANALYSIS AND INTERPRETATION OF SEED PRODUCTION AND RESERVE PATTERNS

We will examine the patterns of seed production and reserve in Brazilian agriculture, using a random forest model to interpret complex and multifaceted data. The analysis of two different phases of cultivation - the Safra and the Safrinha - and the comparison of the performance of the model in each one are highlighted. The interpretation of graphs and tables provides deep insights into the dynamics of the sector, crucial for the development of effective strategies in agricultural management.

Graph 1 below shows the complex relationship between the area planted and the amount of seeds reserved, not demonstrating a clear direct correlation. Table 1 complements this analysis, detailing the distribution of reserved seeds among the main agricultural crops, indicating a significant concentration in a few species.

Graph 1 - Relationship between Planted Area and Reserved Quantity



Source: The author.

Figure 1 provides an analysis of the relationship between 'Planted Area' and 'Reserved Quantity' in the context of reserved seed production in Brazil, based on SIGEF data. Research focuses on discerning the overall trend, identifying patterns, and highlighting outliers. It is noted that



the graph does not reveal an obvious linear correlation between 'Planted Area' and 'Reserved Quantity', with most of the data grouped at the lower end of 'Planted Area', without demonstrating a proportional increase in 'Reserved Quantity' as 'Planted Area' increases. Such an arrangement suggests the absence of a direct and robust relationship between these variables.

In addition, the analysis highlights the presence of outliers, particularly where the 'Reserved Amount' remains elevated regardless of the 'Planted Area', pointing out possible exceptions that could be attributed to factors such as high productivity per area or external variables not captured by the graph. The Pearson correlation, close to 0.0973, indicates a very weak, almost negligible association, corroborating the lack of a strong connection between the variables analyzed.

This study concludes that the 'Planted Area' does not serve as a reliable predictor of the 'Reserved Quantity', inferring that other variables not considered here, such as efficiency in seed production, soil conditions, agronomic and climatic practices, may exert more significant influences. The low correlation reinforces the need for a more integrated and multifactorial approach to understanding the dynamics of seed production, implying that future analyses must incorporate a wider range of data and analytical methodologies to capture the complexity of the seed sector in Brazil.

Table 1: Percentage Analysis of Seeds Reserved for Major Agricultural Species in Brazil

| SPECIES | HOW MUCH RESERVED | % | % Accumulated |
|--------------------------------------|-------------------|---------|---------------|
| <i>Glycine max</i> (L.) Merr. (SOJA) | 3749435500,4430 | 66,2562 | 66,2562 |
| <i>Triticum aestivum</i> L. (TRIGO) | 877409998,2770 | 15,5047 | 81,7608 |
| <i>Solanum tuberosum</i> L. (BATATA) | 779113570,3830 | 13,7677 | 95,5285 |

Source: Prepared by the authors

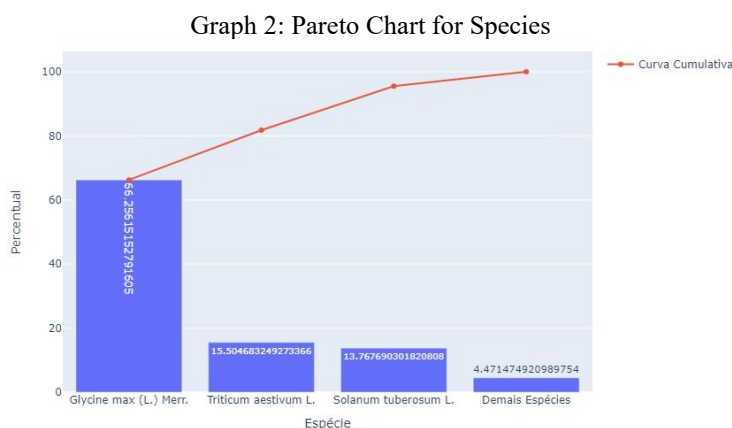
Table 1 provides a quantitative analysis of the seed reserve of three main crops in Brazilian agribusiness, highlighting the predominance of soybeans, wheat and potatoes in the national agricultural spectrum. Specifically, soybeans, *Glycine max* (L.) Merr., stand out significantly, comprising more than two-thirds of the total reserves, emphasizing its central position in Brazil's agricultural and economic structure. The relevance of wheat, *Triticum aestivum* L., and potato, *Solanum tuberosum* L., is also remarkable, underlining their essential contributions to food security and agricultural diversification.

These data point to a concentration of agricultural activity in a few cultures, which, although economically beneficial, introduce risks associated with dependence and lack of variety, emphasizing the importance of diversification strategies. The expressive productive concentration in these species, responsible for more than 95% of the total seeds reserved, shows a duality between economic strength and potential vulnerability to adversities, such as market fluctuations or adverse climatic events.

From this perspective, Table 1 goes beyond the mere presentation of data, functioning as a critical instrument for reflection on agricultural policies. The analysis underscores the urgent need to incentivize crop diversification and adopt sustainable agricultural practices, aiming not only at economic robustness but also at Brazil's resilience and food security. Thus, the study of these data becomes essential for the development of policies that balance production, sustainability and food security in the Brazilian agricultural context.

COMPARATIVE ANALYSIS: SAFRA VS. SAFRINHA

Comparing "Data_Safra" and "Data_Safrinha" reveals marked differences in model performance. "Graph 2: Pareto Graph for Species" and "Graph 3: Sum of Reserved Amount of Seeds by Period" illustrate these variations, showing how seed production responds to different factors and conditions in each crop cycle.



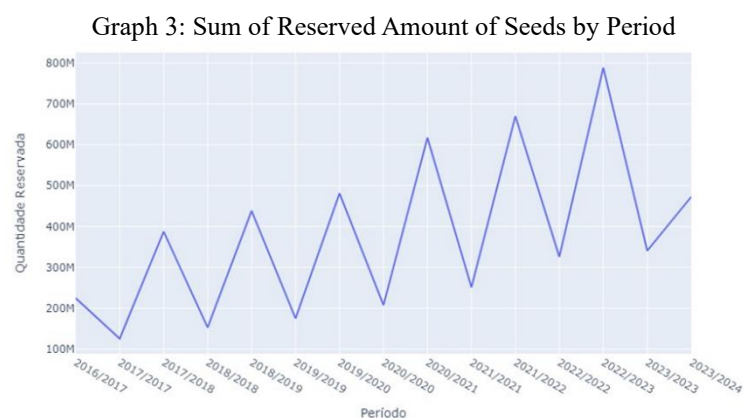
Source: The authors.

In the presentation of the results presented in Table 1, the use of the Pareto graph emerges as an outstanding analytical strategy to elucidate the distribution of seeds among the main agricultural crops in Brazil. This graph is recognized for its ability to visually represent the predominance of certain factors, in this context, specific crop varieties, in contributing significantly to a comprehensive outcome, here being the aggregate of the reserved seeds. The use of this graph is based on its ability to demonstrate the Pareto Principle, or 80/20 rule, elucidating that a smaller segment of causes tends to be responsible for most of the observed results.

The Pareto Chart, identified as Figure 2, is established as an essential visual mechanism to analyze and discuss the allocation and repercussions of the seed reserve in the Brazilian agricultural sector. This instrument facilitates the recognition of areas that demand priority attention for future interventions and research, seeking a detailed understanding of the current configuration of the seed reserve in the country's agriculture, in addition to clarifying market forces and potentials for advances in the domain.



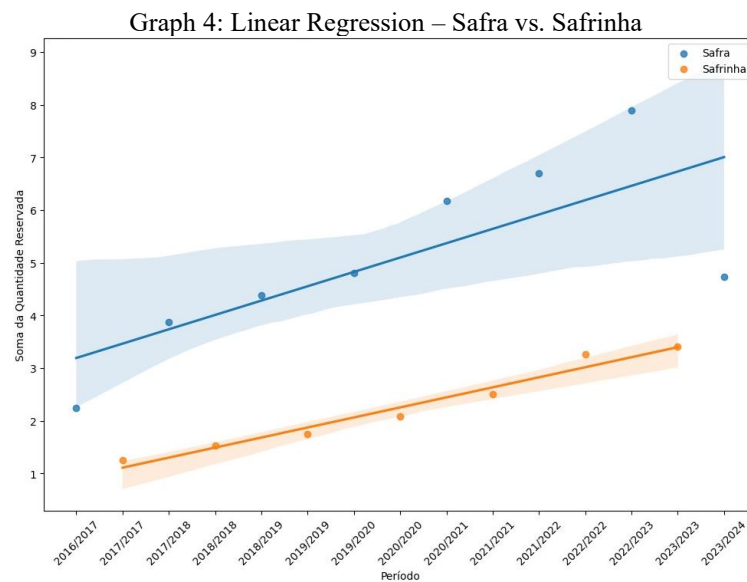
Specifically, the Pareto Chart highlights the preeminent role of crops such as soybeans, wheat, and potatoes in the Brazilian seed production landscape, bringing substantial implications for agricultural policymaking, strategy delineation, and sustainability-oriented ventures. This analysis indicates the importance of adopting a balanced approach that not only enhances the efficient and organic production of these key cultures but also encourages agronomic diversification to strengthen the resilience and adaptability of the sector. The emphasis on soybeans, which represents a significant portion of production according to the data, suggests that policy measures could be directed to expand this culture, investing in areas such as research, genetic improvement and logistics infrastructure, having a positive impact on national agricultural development.



Source: Prepared by the authors

Figure 3 provides a longitudinal analysis of seed stock in Brazilian agriculture from 2016 to 2024, revealing a pattern of significant fluctuations that reflect the dynamics of the agricultural sector. There is no uniform trend of growth or decline in reserves over the years studied; Instead, annual changes point to a response by the sector to a diversified set of influences, including environmental, economic, and market factors. Periods such as 2019/2020 and 2022/2023 are distinguished by peaks in seed stock, possibly indicating years of abundance or response to increases in demand. In contrast, the decline observed in 2023/2024 may denote adversities faced by the sector.

The study also differentiates the "Safrá" and "Safrinha" crop cycles, highlighting the influence of climatic conditions and planting strategies on production. Safrá, the main growing period, benefits from ideal climatic conditions and has high productivity, while Safrinha, despite less favourable conditions and traditionally lower productivity, represents a crucial stage for the sustainability and continuity of agricultural production. These observations underscore the adaptation capacity and resilience of the Brazilian agricultural sector to intermittent variables, providing a basis for an in-depth understanding of management and decision-making strategies in the context of domestic seed production.



Source: Prepared by the authors

Figure 4 illustrates the temporal relationship between the growing seasons, Safra and Safrinha, and the sum of the reserved quantities of seeds in the context of Brazilian agriculture. The blue line, representing the Safra, and the orange line, representing the Safrinha, are the results of a linear regression that seeks to estimate the central trend of the data for each period. The shaded areas correspond to the confidence intervals for each regression, which indicate the degree of uncertainty associated with the model's predictions.

The wider shaded area around the Safra regression line suggests greater variability and less confidence in seed reserve predictions for this period. This implies that additional factors, potentially not captured by the model, may be influencing seed stocks beyond time, such as climatic variations, agricultural policies, or economic changes. Meanwhile, the narrower shaded area around the Safrinha regression line implies greater accuracy in predictions, suggesting that the amount of seeds reserved may be more consistently predicted based on time for that period.

The presence of points above or below the regression lines indicates the variability of the data in relation to the modeled trend, and those points can be considered outliers if they are outside the confidence interval. For example, if a data point is above the shaded area of the Safra, this may indicate a year when the seed stock was exceptionally high, possibly due to a particularly bumper harvest or policies to stimulate the seed stock.

Pearson's coefficient for each dataset quantifies the strength of the linear relationship between time and seed stocks. A Pearson coefficient of about 0.755 for Safra and about 0.983 for Safrinha indicates a strong positive correlation for both periods, with Safrinha showing a near-perfect relationship. However, the interpretation of the coefficient must consider the specific context and potential confounding factors that may affect seed stocks.



These statistical analyses provide fundamental insights into the dynamics of seed production and reserve, evidencing continuous growth in both phases, but with greater predictability and stability during Safrinha. Particular attention is paid to the anomalous behavior of the data in 2023/2024, which highlights the need for additional research to understand unexpected variations and their implications for farm management. The methodology employed reveals critical patterns and provides a solid foundation for strategic planning and informed decision-making in the agricultural sector.

Therefore, the statistical analysis in Figure 4 demonstrates the usefulness of quantitative methods such as linear regression and calculation of confidence intervals to understand the dynamics of agricultural production and to inform strategic decisions in the sector. The analysis suggests that, while there is a positive trend in the volume of seeds reserved over time for both growing periods, annual variability and confidence in predictions may differ substantially between Safra and Safrinha.

MODEL PERFORMANCE ANALYSIS

This chapter is dedicated to evaluating the effectiveness of the randomized forest model developed to predict seed production in Brazilian agriculture. This analysis is structured in three distinct subchapters, each focusing on different datasets and contexts to provide a comprehensive view of model performance.

In the first subchapter, "Seed Production Prediction: A Study on Data_Safra", we examine the model's performance using the specific data of the main crop, focusing on the accuracy of the predictions and the model's ability to conform to the historical data of this critical period.

The second subchapter, "Seed Production Prediction: A Study on Data_Safrinha", discusses the model's predictions for safrinha, a secondary but equally significant growing period, highlighting the peculiarities and specific modeling challenges for this crop cycle.

Finally, in the third subchapter, "Seed Production Predictions: An Analysis between 'Data_Safra' and 'Data_Safrinha'", we made a direct comparison of the model's performance between the two growing periods. This comparison seeks to understand the variations and consistency of the model in different agricultural contexts, contributing to a deeper understanding of the effectiveness of Artificial Intelligence techniques in predicting variable dynamics in the agricultural sector.

Each subchapter contributes to a detailed understanding of how the random forest model adapts and responds to the complexities of seed production, providing valuable insights for future applications and improvements in the field of precision agriculture.



Seed Production Prediction: A Study on 'data_safra'

The analysis of the regression model using the 'data_safra' subset of data, drawn from a larger set, provides a quantitative and qualitative view of the effectiveness of seed production predictions. The metrics obtained — Mean Square Error (MSE) and Coefficient of Determination (R^2) — are essential to evaluate the random forest model adopted in this research.

The Mean Square Error (MSE) metric for the overall and test sets is significant, indicating a noticeable discrepancy between the predicted and actual values. The model's accuracy for new data is therefore lower, as illustrated by the MSE of about 2,833,804,168.5. This high number shows the limitations of the model in predicting new events, a critical aspect in predictive modeling.

On the other hand, the MSE for the training set shows a more promising performance, with a substantially lower value of 1,653,458,753.5. This difference highlights a greater accuracy of the model under the conditions under which it was trained, an expected result and indicative of an appropriate fit to the training data.

As for the Coefficient of Determination (R^2), the values for the test and training sets reveal a significant disparity. For the test set, the R^2 of 0.4149566274957277 implies that only about 41.5% of the variance is explained by the model. Although it indicates a predictive capacity, it also suggests that factors not contemplated in the model have a significant role. For the training set, the R^2 is 0.6544389740961896, revealing that the model manages to explain approximately 65.44% of the variance, demonstrating a reasonable fit to the data used for training.

This variation between the training and test sets, particularly a test success rate of 63.41%, suggests a phenomenon known as overfitting. This occurs when the model learns the particularities of the training data to such an extent that it fails to generalize its predictions to new data, limiting its practical usefulness. This finding suggests the need for a more refined balance between the fit of the model and its generalizability.

Seed Production Prediction: A Study on 'data_safrinha'

The study in question uses the 'data_safrinha' subset of a larger DataFrame to model and evaluate predictions about seed production, using a random forest model. The metrics resulting from the analysis, Mean Square Error (MSE) and Coefficient of Determination (R^2), are vital tools for the performance of the predictive model.

The Mean Square Error (MSE) arises as an indicator of the mean of the squared errors, that is, the differences between the model's predictions and the actual values observed. A high MSE, such as the one obtained for both the general and test sets, with a value around 3,113,846,766.768054, indicates a considerable magnitude of error in the predictions for the new data, pointing to an area of attention and potential improvement in the accuracy of the model. However, the MSE obtained for



the training set is substantially lower, evidencing a more accurate performance in the predictions with a value of 1,386,346,989.5445054.

When analyzing the Coefficient of Determination (R^2 Score), which reflects the proportion of variance in the dependent variable that is explained by the independent variables in the model, we notice a marked distinction between the test and training sets. An R^2 of 0.7183833224143992 for the test set indicates that approximately 71.84% of the variance is captured by the model, a considerable value that attests to the model's effectiveness in explaining the data. This percentage increases for the training set, where the R^2 achieves a robust 0.8676384019256064, suggesting that the model fits well with the training data.

However, the difference between the training and test metrics reveals that, although the model shows efficiency in learning the patterns in the training data, as indicated by a relatively high training R^2 , it does not maintain the same level of accuracy when applied to new data, a discrepancy that signals the presence of overfitting.

These results, together with the Percentage of Success of the Test in Relation to Training of 82.80%, highlight the need for further research and fine-tuning of the model to improve its generalization and predictive capacity. Despite the areas for improvement, the random forest model is presented as a robust and promising approach to address the complexity of data in agriculture, with the potential to significantly improve the accuracy and effectiveness of agricultural resource planning and management.

Seed Production Predictions: An Analysis Between 'data_safra' and 'data_safrinha'

The data reveal that the performance of the model varies between the two periods, with a Coefficient of Determination (R^2) of 0.4149566274957277 for the "data_safra" and a considerably higher R^2 of 0.7183833224143992 for the "data_safrinha". The latter suggests that the random forest model manages to explain about 71.84% of the variability of the data for the safrinha period, a notable contrast with the lower percentage of about 41.5% for the safra period.

When analyzing the Mean Square Error (MSE), we notice that although the values are high for both datasets, the MSE of the "data_safrinha" is slightly higher, which could suggest a greater discrepancy in the predictions for this period. However, it is essential to consider this data together with R^2 , which tells us a different story, indicating a better predictive capacity for safrinha.

More notable is the discrepancy between the training and test results for each dataset. The "data_safra" shows a more significant difference, an indicator that can signal an overfit of the model to the training data, reducing its effectiveness in generalizing to unseen data. On the other hand, the "data_safrinha" shows a smaller discrepancy between the training and test results, reflecting a better



generalization and a strong linear correlation between the variables, as indicated by the high Pearson coefficient of 0.9829378120964237.

This strong correlation for the "data_safrinha" reaffirms the ability of the random forest model to capture data trends, allowing accurate and reliable predictions for this period. The model not only achieves superior performance in predicting safrinha data, but also reveals the ability of AI algorithms to adapt and learn efficiently from consistent and strongly linear patterns.

The comparative data from the "data_safra" and "data_safrinha" sets illustrate that success in predictive modelling depends significantly on the nature and quality of the input data. The differences between the two periods reinforce the importance of adjusting predictive models according to the specific characteristics of each seed production cycle. The analysis underlines the need for a careful and contextualized approach in the application of modelling techniques, with the aim of maximising the effectiveness of predictions, essential for agricultural planning and management.

CONCLUSIONS OF THE STUDY

This detailed study, articulating the use of random forest models, outlines a detailed picture of the dynamics of seed production and reserve in the Brazilian agricultural context, offering a solid basis for future decision-making strategies in the sector. The research underscores the intrinsic complexity of agricultural practices in the country, illuminating the variations between distinct crop cycles and diversified regions. With an analytical and statistical perspective, the study maps the interaction between a myriad of determining factors, emphasizing their interconnectivity and mutual influence.

The implementation of the random forest model proved its analytical capabilities, efficiently managing the wide range of industry data and extracting patterns that lead to reliable predictions. Although the variations in the accuracy of the predictions between the Safra and Safrinha cycles have been significant, the application of the model highlighted the adaptability needed to accommodate these fluctuations, a key aspect for predictive modelling in agriculture.

The concentration of production in a limited number of crops emerges as a double facet, representing both an economic strength and a point of vulnerability to market fluctuations and climate changes. The study therefore highlights the critical importance of promoting diversification as a pillar to strengthen the resilience and sustainability of the sector. Agricultural management strategies based on robust data are identified as essential to respond effectively to the dynamics described, implying the need for judicious adjustments in cultivation practices and harvest planning.

The work also points to the need for more incisive future research, which can reveal even more deeply the relationships between the variables that affect agricultural production and refine the



predictive models used. The research opens up perspectives for the integration of climate, economic and agronomic data in a multidisciplinary approach, with the aim of improving the predictive capacity and management of the sector.

By providing valuable contributions to both the scientific literature and agricultural practice, the study supports farmers, managers and policymakers in making informed and evidence-supported decisions. In sum, the present study has not only clarified the complex seed production configurations in Brazil, but has also ratified the relevance and effectiveness of random forest models, representing a notable advance at the intersection between data science and agriculture.



REFERENCES

1. Adama Brasil. (2024). Inteligência artificial na agricultura: quais são as principais funcionalidades? Disponível em: <https://www.adama.com/brasil/pt/inovacao/inteligencia-artificial-na-agricultura-quais-sao-principais-funcionalidades>. Acesso em: 17 jan. 2024.
2. Climate Fieldview. (2024). Quais as aplicações da inteligência artificial na agricultura? Disponível em: <https://blog.climatefieldview.com.br/inteligencia-artificial-agricultura>. Acesso em: 17 jan. 2024.
3. Gadotti, G. I., Ascoli, C. A., Bernardy, R., Monteiro, R. C. M., & Pinheiro, R. M. (2022). Machine Learning For Soybean Seeds Lots Classification. Scientific Paper, Special Issue: Artificial Intelligence. Eng. agríc. (Online), 42(spe). DOI: [org/10.1590/1809-4430-Eng.Agric.v42nepe20210101/2022](https://doi.org/10.1590/1809-4430-Eng.Agric.v42nepe20210101/2022). Disponível em: <https://www.scielo.br/j/eagri/a/LtTLRpzgNQPWp5mw3qRMdtM/?format=pdf&lang=en>. Acesso em: 15 jan. 2024.
4. Mourtzinis, S., Esker, P. D., Specht, J. E., et al. (2021). Advancing agricultural research using machine learning algorithms. Sci Rep, 11, 17879. DOI: [org/10.1038/s41598-021-97380-7](https://doi.org/10.1038/s41598-021-97380-7). Disponível em: <file:///C:/Users/User/Downloads/s41598-021-97380-7.pdf>. Acesso em: 14 jan. 2024.
5. Rehagro. (2024). Inteligência artificial na agricultura: benefícios e aplicações. Disponível em: <https://rehagro.com.br/blog/inteligencia-artificial-na-agricultura/>. Acesso em: 17 jan. 2024.
6. Saleem, M. H., Potgieter, J., & Arif, K. M. (2021). Automation in Agriculture by Machine and Deep Learning Techniques: A Review of Recent Developments. Precision Agric, 22, 2053-2091. DOI: [org/10.1007/s11119-021-09806-x](https://doi.org/10.1007/s11119-021-09806-x). Disponível em: <https://link.springer.com/article/10.1007/s11119-021-09806-x>. Acesso em: 15 jan. 2024.
7. Vieira Filho, J. E. R., & Silveira, J. M. F. J. (2012). Mudança tecnológica na agricultura: uma revisão crítica da literatura e o papel das economias de aprendizado. Revista de Economia e Sociologia Rural, 50(4), 651-666. Disponível em: <https://www.scielo.br/j/resr/a/Pjz4mbbbKwDz8Vm4sbDY7mR/?format=pdf&lang=pt>. Acesso em: 18 jan. 2024.