

# XGBoost and the Budgetary Practice of MAPA Influence on the HDI and on Agribusiness GDP

Jorge Castro

ORCID: [0000-0003-0612-4576](https://orcid.org/0000-0003-0612-4576)

Available at: [https://github.com/Jotta-se/MBA\\_USP](https://github.com/Jotta-se/MBA_USP)

---

## ABSTRACT

The efficient allocation of public resources represents a structural challenge for governments worldwide, particularly in strategic sectors such as agribusiness. The budgetary planning of Brazil's Ministry of Agriculture, Livestock and Food Supply (MAPA) lacks predictive tools capable of identifying which budgetary, socioeconomic, and regional variables most influence human development and agribusiness performance. This study applies the XGBoost algorithm, with hyperparameter optimization via GridSearchCV, to a dataset of 465 observations drawn from federal government open data, the 2017 Agricultural Census (IBGE), and CEPEA/USP, covering the period from 2018 to 2022. Two target variables are modeled, the percentage variation of the Agribusiness GDP (%GDP-AGRO) and the Human Development Index (HDI) of Brazilian states. Feature selection employed Recursive Feature Elimination (RFE) and multicollinearity analysis. The %GDP-AGRO model achieved a Coefficient of Determination ( $R^2$ ) of 0.8764 and a Root Mean Squared Error (RMSE) of 0.00956, while the HDI model obtained an  $R^2$  of 0.9632 and an RMSE of 0.00712. In both targets, the percentage of rural properties with access to technical assistance emerged as the most important feature (28.18% and 52.11%, respectively), followed, in the GDP model, by investments in Projects (analogous to CAPEX, 21.59%), and, in the HDI model, by geographic location (Southeast region, 31.89%). The results provide empirical evidence for data-driven public policy formulation, demonstrating that rural extension services and capital investments are critical determinants of Brazilian agribusiness competitiveness and rural populations' well-being.

**Keywords:** Public budget; Machine learning; Extreme Gradient Boosting; XGBoost; Agricultural sector; Human development; HDI; GDP Agribusiness; MAPA; Public policy.

## 1. INTRODUCTION

The efficient management of public resources is a global challenge that intensifies as governments must reconcile growing fiscal pressures with increasingly complex social demands [1]. In the Brazilian context, the agricultural sector occupies a strategic position. In 2022, agribusiness accounted for approximately 24.8% of the national Gross Domestic Product, consolidating Brazil as one of the world's largest agricultural producers and exporters [2]. Despite this economic prominence, the translation of public investment into measurable gains in productivity, competitiveness, and human development remains insufficiently understood in the national literature.

The Ministry of Agriculture, Livestock and Food Supply (MAPA) is the main institutional vector of these policies, coordinating entities such as EMBRAPA, INCRA, CONAB, and SFB. Between 2018 and 2022, the ministry's total authorized budget ranged from BRL 4.04 billion to BRL 7.23 billion per year, yet the average rate of conversion of commitments into effective payments remained at 47.9% [3]. This chronic gap between what is planned and what is executed reveals not only operational inefficiencies but also the absence of analytical tools capable of identifying which expenditure categories generate the greatest impact on intended outcomes.

In this context, the advancement of machine learning techniques, particularly Extreme Gradient Boosting (XGBoost), offers a concrete opportunity to fill this gap. XGBoost stands out for its robustness to multicollinearity, its capacity to capture non-linear relationships, and its computational efficiency on datasets of moderate size [4, 5]. Its application to budgetary and socioeconomic data makes it possible to identify, in an objective and reproducible manner, which variables most influence the outcomes of public policy.

This article presents the following original contributions:

- (i) Application of XGBoost to the predictive modeling of two distinct targets, the percentage of GDP-AGRO and the HDI, from MAPA open budget data (2018-2022), integrating variables from the 2017 Agricultural Census and the CEPEA/CNA historical series;
- (ii) Quantitative mapping of the relative importance of budgetary, socioeconomic, and regional features, revealing the centrality of rural technical assistance as a determinant of both targets;
- (iii) Empirical evidence for the formulation of data-driven public policy in the agricultural sector, with direct implications for the prioritization of the federal budget;
- (iv) Public release of the source code and datasets in a GitHub repository [6], ensuring transparency and replicability.

The Research Questions (RQs) that guide the study are:

RQ1. Which budgetary, socioeconomic, and regional variables most influence the percentage of Agribusiness GDP in Brazil between 2018 and 2022?

RQ2. To what extent do these same variables explain the variation in the HDI of the Brazilian states?

RQ3. Does the application of XGBoost produce sufficient explanatory power to support budget allocation decisions in the agricultural sector?

The article is organized as follows. Section 2 presents the theoretical framework and the state of the art. Section 3 describes the data and the methodology. Section 4 reports the results. Section 5 discusses the implications. Section 6 addresses threats to validity. Section 7 concludes the work.

## **2. THEORETICAL FRAMEWORK AND STATE OF THE ART**

### ***2.1 Machine learning in public budgeting and financial management***

The application of artificial intelligence techniques to public resource management has been increasingly documented in the international literature. Haytham and Elmousalami [7] compare different AI techniques for conceptual cost estimation in projects, highlighting the potential of machine learning methods as an alternative to traditional parametric approaches. Puhach [8] applies machine learning econometrics to the simulation of factors influencing local budget expenditures, showing that gradient boosting captures non-linear relationships that escape classical regression models. Capone et al. [4] confirm the potential of XGBoost for forecasting government expenditures, emphasizing its capacity to process large volumes of heterogeneous data. Zainab et al. [9] develop an integrated data intelligence model for cost estimation in construction, demonstrating that the combination of classical statistical techniques with advanced predictive models is superior to any approach in isolation.

### ***2.2 Public investment, agricultural productivity, and economic growth***

The relationship between public investment in the agricultural sector and economic growth is well established in the literature. Torres [10] demonstrates that investments in research and development are vital for raising agricultural productivity, with returns higher than those of other categories of public spending. Diallo and Wouterse [11] document, for African economies, that investments in rural infrastructure and technical extension are the strongest catalysts of agricultural GDP and rural poverty reduction. In Brazil, Corcioli, Medina, and Arrais [12] reveal that 75% of government operating credit is directed to soybean, corn, and cattle producers, with 76.1% of the soybean production chain controlled by foreign multinationals, suggesting that the current allocation model may not maximize the social return on invested resources.

### ***2.3 Agribusiness, HDI, and regional disparities***

The relationship between agricultural development and the Human Development Index is multidimensional and geographically heterogeneous. Piras, Wesz, and Ghinoi [13] observe that soybean expansion is associated with improvements in the HDI of Brazilian municipalities, although the benefits are more pronounced in the income dimension than in education and longevity. Soleman et al. [14] demonstrate, for the Indonesian context, that the HDI exerts a positive and significant impact on GDP, suggesting a reciprocal relationship between human development and economic growth. Ondetti [15] points out that historical land concentration in Brazil produces long-term negative effects on the HDI, particularly in the dimensions of longevity and income. Tahu et al. [16] emphasize the role of capital expenditure as an intervening variable between regional revenues and economic growth.

### ***2.4 Positioning of the present study***

Table 1 summarizes the positioning of the present study in relation to the three lines of research identified. The first line comprises studies on machine learning (ML) applied to the public budget [4, 7, 8, 9], in which predictive techniques are applied to budgetary data without, however, connecting their features to socioeconomic impact variables such as the HDI or

sectoral GDP. The second line gathers works on agricultural investment and economic growth [10, 11, 12], whose focus is on the effects of credit and public spending policies on rural productivity, predominantly through classical econometrics and without the use of ML predictive models. The third line concentrates studies on the HDI, land distribution, and regional disparities [13, 14, 15], which examine the determinants of human development in rural contexts, also without resorting to ML applied to budgetary data. The present article integrates the three perspectives by using XGBoost on federal budget data combined with socioeconomic and Agricultural Census variables, simultaneously modeling the percentage of GDP-AGRO and the HDI as target variables.

**Table 1 - Positioning of the present study within the literature**

Thematic line	Representatives	Predominant method	Distinction of this study
ML in public budgeting	[4, 7, 8, 9]	XGBoost, GBM, neural networks	Connects budgetary features to HDI and GDP-AGRO
Agricultural investment and growth	[10, 11, 12]	Classical econometrics	Feature importance quantifies transmission channels
HDI, land distribution, and region	[13, 14, 15]	Panel, multiple regression	Integrates 2017 Agricultural Census data into the model

### 3. MATERIAL AND METHODS

#### 3.1 Data sources and analysis period

The dataset was built from four publicly accessible open sources:

- (i) Physical-Financial Monitoring of Budgetary Actions and the Fiscal and Social Security Budget, available on the Federal Government Open Data Portal [3, 17];
- (ii) the 2017 Agricultural Census of IBGE [18], the most recent available, which provided data on land structure and access to technical assistance;
- (iii) the historical series of Agribusiness GDP (CEPEA/ESALQ/USP-CNA) [2]; and
- (iv) U.S. dollar (USD) exchange rates for the conversion of current values (CEPEA/ESALQ/USP) [2]. The analyzed period comprises the fiscal years from 2018 to 2022.

The unit of analysis is the tuple (fiscal year x region/state x budgetary unit x subfunction x action type), resulting in 465 valid observations after the cleaning and integration process. The final dataset presented no missing values (null or NA), dispensing with imputation.

#### 3.2 Variables and preprocessing

The dataset incorporates five categorical variables, namely region, state, budgetary unit (MAPA, EMBRAPA, INCRA, CONAB, SFB), subfunction (22 categories), and action type

(Activity/OPEX, Project/CAPEX, Special Operation). In addition, numerical variables were included covering the budgetary values effectively available in the public bases (authorized and paid, in Brazilian reais and in U.S. dollars), macroeconomic indicators (GDP and sectoral GDP-AGRO), and variables extracted from the Agricultural Census (total number of properties, technical assistance rate, and population distribution per property). The categorical variables were encoded via one-hot encoding, and the numerical variables were standardized through the StandardScaler, a technique that transforms each variable by subtracting its mean and dividing by the standard deviation, resulting in zero mean and unit standard deviation, so that no variable exerts disproportionate influence on the model because of its original scale. Additionally, the logarithmic transformation  $\text{Log\_Pop\_Por\_Propriedade}$  was created to capture non-linear relationships between population distribution and land structure.

Feature selection adopted a theoretical, hypothesis-driven approach, in which the variables were defined a priori from a conceptual model of the determinants of agricultural development, rather than through automatic mining of correlations. Each feature corresponds to an explanatory dimension, namely investment capacity (Pago\_USD and the Project/CAPEX action type), regional wealth (GDP), demographic pressure on the land (logarithm of the population per property), access to technical assistance (percentage of properties receiving technical guidance), and structural controls for region and budgetary unit. This strategy was chosen for three reasons. First, theoretical selection is methodologically more conservative than automatic selection, as it ensures that each variable has a prior conceptual basis, preventing the algorithm from treating spurious correlations as explanatory power. Second, a fixed set of features ensures interpretability and comparability between the two models of the study (HDI and percentage of GDP-AGRO), allowing the importance of the same variables to be compared side by side, which is central to the proposed reading. Third, given a dataset of moderate size (465 observations), restricting the degrees of freedom of the selection process reduces the risk of overfitting. The final set comprises 16 features for the GDP-AGRO percentage model and 14 features for the HDI model. Recursive Feature Elimination (RFE) with linear regression remained chained to the pipeline for mechanical consistency but, under the fixed theoretical set, did not act as an additional reducer. As a complementary diagnostic, multicollinearity was inspected using the criterion  $|r| > 0.70$  (Section 6).

### ***3.3 Modeling with XGBoost and hyperparameter optimization***

The XGBoost Regressor was selected as the main algorithm because of its robustness to residual multicollinearity, its efficiency with datasets of moderate size, and its capacity to capture non-linear interactions among features [4, 5]. Hyperparameter optimization employed GridSearchCV with k-fold cross-validation ( $k=5$ ), sweeping a grid that included  $n\_estimators$  in  $\{100, 200\}$ ,  $max\_depth$  in  $\{3, 5, 7\}$ ,  $learning\_rate$  in  $\{0.05, 0.1, 0.2\}$ ,  $reg\_alpha$  in  $\{0, 0.1\}$ , and  $reg\_lambda$  in  $\{0, 0.1\}$ . The data were split into training (80%) and test (20%) sets.

Table 2 summarizes the selected hyperparameters and their justification. For the HDI model, the absence of additional regularization ( $\alpha = \lambda = 0$ ) is justified by the smaller number of features and the lower dimensionality of the problem. For the GDP-AGRO percentage

model, L2 regularization with  $\lambda = 0.1$  (Ridge) contributed to the stability of the estimates given the larger number of variables.

**Table 2 - Hyperparameters selected by GridSearchCV for each model**

Hyperparameter	GDP-AGRO model	HDI model	Justification
n_estimators	200	100	Predictive capacity vs. computational cost
max_depth	5	5	Overfitting control
learning_rate	0.1	0.2	Stable convergence
reg_alpha (L1)	0	0	No forced sparsity
reg_lambda (L2)	0.1	0	Moderate Ridge; robust to multicollinearity

### 3.4 Model evaluation and feature importance analysis

Model performance was evaluated using the metrics  $R^2$  (Coefficient of Determination), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error). The feature importance analysis used the native XGBoost criterion (average split gain weighted by coverage), allowing the identification of which variables contributed most to reducing the prediction error. A complementary geographic and thematic analysis examined the budgetary distribution by region and subfunction over the period.

## 4. RESULTS AND DISCUSSION

### 4.1 Model performance

Table 3 summarizes the performance metrics of the two models on the test sets. Both exhibit high explanatory power, validating the suitability of XGBoost for this domain and answering RQ3 affirmatively.

**Table 3 - Performance metrics of the XGBoost models (test set, 20%)**

Model	$R^2$	RMSE	MAE	MSE
XGBoost - % GDP-AGRO	0.8764	0.00956	0.00358	$9.14 \times 10^{-5}$
XGBoost - HDI	0.9632	0.00712	0.00374	$5.07 \times 10^{-5}$

The GDP-AGRO percentage model achieves an  $R^2$  of 0.8764, indicating that approximately 87.6% of the variability in the agricultural GDP percentage is explained by the selected features. The RMSE of 0.00956 is notable given the target variable's range between 0.20 and 0.27. The HDI model surpasses this result, with an  $R^2$  of 0.9632, indicating that 96.3% of the variability in the HDI is captured by the features, with an RMSE of 0.00712 over a scale of 0.676 to 0.859. The low MAE values in both models (0.00358 and 0.00374, respectively)

corroborate that the predictions are very close to the observed values, without systematic distortions attributable to outliers. To assess the sensitivity of these results to the selection strategy, Section 4.4 reports an automatic selection experiment applied to the HDI model.

#### 4.2 Relative importance of the features

Table 4 presents the feature importance ranking for the two models, answering RQ1 and RQ2.

**Table 4 - Relative importance of the features for the XGBoost models**

Panel A - XGBoost model for % GDP-AGRO ( $R^2 = 0.8764$ )		
Rank	Feature	Importance
1	% Receiving Technical Assistance	0.2818
2	Project Action Type (CAPEX)	0.2159
3	HDI	0.1175
4	Log. Population per Property	0.1059
5	General GDP	0.0891
6	Budgetary Unit - EMBRAPA	0.0511
7	Southeast Region	0.0400
8	Northeast Region	0.0389
Panel B - XGBoost model for HDI ( $R^2 = 0.9632$ )		
Rank	Feature	Importance
1	% Receiving Technical Assistance	0.5211
2	Southeast Region	0.3189
3	Center-West Region	0.1067
4	Log. Population per Property	0.0179
5	General GDP	0.0110
6	BR Region (national)	0.0109
7	Project Action Type (CAPEX)	0.0048
8	North Region	0.0036

#### - Rural technical assistance (RQ1 and RQ2)

The variable %\_Receber\_Orientacao\_Tecnica emerges as the most important factor in both models, accounting for 28.18% in the GDP-AGRO percentage model and 52.11% in the HDI model. This result is especially relevant in light of the reality revealed by the 2017 Agricultural Census [18], according to which only 20.2% of Brazilian rural properties receive any type of technical guidance, with marked regional disparities of 48.6% in the South, 10.4% in the North, and 8.2% in the Northeast. The convergence of both models around this variable suggests that rural extension operates as the main transmission channel between public investment and socioeconomic outcomes in the countryside, corroborating Torres [10], Diallo and Wouterse [11], and Piras et al. [13].

#### - Investments in Projects/CAPEX (RQ1)

The Project action type, equated with Capital Expenditure (CAPEX) in the private sector, is the second most relevant factor for the GDP-AGRO percentage, with a relative importance of 21.59%, reinforcing the role of capital investments as a driver of the sector's performance. This finding is particularly significant when contrasted with the budget execution scenario described in Table 5, in which the deterioration of the conversion rate of commitments into payments reveals a growing difficulty in carrying out precisely the type of investment that the model identifies as most decisive for agribusiness. There is, therefore, a contrast between what the data indicate as a priority and what the budget actually executes.

- **Geographic disparities (RQ2)**

For the HDI model, geographic location occupies the second and third places, with the Southeast region accounting for 31.89% and the Center-West region for 10.67%. This distribution reflects the historical asymmetries documented by Ondetti [15] and Yuliana et al. [19] and highlights the need for regionally differentiated policies. The traditional economic variables, GDP and payments in USD, showed relatively low importance, of 1.10% and 0.35%, respectively, challenging the premise that aggregate economic growth is the main driver of human development.

### **4.3 Analysis of MAPA budget execution**

The analysis of the 2018-2022 budget series, presented in Table 5, reveals a structural mismatch between the authorization and the delivery of resources. Although the authorized budget grew from BRL 4.04 billion in 2018 to BRL 7.22 billion in 2021, the amounts effectively paid remained in the range of BRL 1.6 to 2.7 billion throughout the period. The execution rate, measured as the ratio between the paid and the authorized amount of each segment, fluctuated within a low and stable band, from 39.7% in 2018, declining to 36.5% in 2021 and recovering to 43.5% in 2022, with an average of 39.2% over the five-year period. The regional reading reinforces the diagnosis, since, while the aggregate national execution reaches 43.5%, the regional units present substantially lower rates, from 11.3% in the Center-West to 2.6% in the North. By executing agency, EMBRAPA reaches 56.9% execution and MAPA 46.6%, whereas CONAB executes only 16.7% of the authorized amount. This pattern indicates that the bottleneck does not lie in budget allocation but in the operational capacity for execution, especially in the regional units and in specific agencies, in line with the combination of economic, political, and strategic factors identified by Monteiro et al. [20].

**Table 5 - MAPA budget execution by fiscal year (2018-2022). Execution = ratio between the paid and the authorized amount.**

Year	Authorized (BRL bi)	Paid (BRL bi)	Execution (Paid/Authorized)
2018	4.04	1.60	39.7%
2019	5.96	2.22	37.2%
2020	6.27	2.46	39.2%
2021	7.22	2.64	36.5%
2022	6.16	2.68	43.5%
Average	5.93	2.32	39.2%



#### ***4.4 Robustness analysis through automatic feature selection***

To assess the sensitivity of the results to the feature selection strategy, a two-stage automatic selection procedure, free of information leakage, was applied to the HDI model. In the first stage, multicollinearity was computed exclusively on the training set, discarding, in each pair with  $|r| > 0.70$ , the variable with the highest average redundancy. In the second stage, RFE was coupled to the pipeline and the number of features to retain was swept by the GridSearchCV cross-validation, jointly with the hyperparameters, with rescaling within each fold. Starting from 42 candidate features, multicollinearity pruning removed 24 variables, leaving 18, of which RFE selected 14.

The resulting model reaches an  $R^2$  of 0.9959 on the test set, with an RMSE of 0.00239. This performance should not be read as evidence of a superior model, but as a symptom of the dominance of the technical assistance variable discussed in Section 5.3. By removing the theoretical anchor, the procedure concentrates importance even further on this variable, which rises from 52.1% in the theoretical model to 66.2% under automatic selection. The relevant finding is one of consistency, since the importance hierarchy remains essentially the same under both selection philosophies, with technical assistance and the Southeast region occupying the top positions in both cases. The study's central conclusion, therefore, survives the change in selection strategy. The experiment was conducted only for the HDI model, in an illustrative capacity, since the strategy adopted in the study is fixed theoretical selection; replicating it for the GDP-AGRO percentage model would add no methodological evidence and would compromise the comparability between the two models that motivates the fixed feature set.

### **5. DISCUSSION**

#### ***5.1 Implications for public policy***

The results converge on three public policy recommendations, each anchored in the importance percentages produced by the models, and reinforce the need to consolidate a data-driven culture in the budget management of the public sector:

- (i) The expansion of access to rural technical assistance should be treated as a first-order budgetary priority, not as a residual line. With only 20.2% of properties covered [18] and a relative importance of 28% to 52% in the models, the marginal return of each unit of currency allocated to this channel exceeds that of virtually any other expenditure category analyzed.
- (ii) Investments in Projects/CAPEX, which by definition encompass capital goods aimed at expansion and operational improvement, should be protected from budget freezes, given their relative importance of 21.59% in the GDP-AGRO model.
- (iii) Development policies should incorporate an explicit regional dimension, given that geographic location explains 42.6% of the variation in the HDI in the model, evidence that universal transfers and programs have limited impact on the reduction of structural regional inequalities.

### ***5.2 Interpretation in light of the theoretical framework***

The centrality of technical assistance in both models is consistent with the argument of Soleman et al. [14] regarding the reciprocal relationship between human development and economic growth, in that, by raising the HDI through rural extension, the agricultural sector creates the conditions for its own future growth. The low importance of traditional economic variables, such as GDP and payments in USD, in the HDI models confirms the warning of Piras et al. [13] that agribusiness growth, without the mediation of distributive policies, does not automatically translate into improved well-being. The result aligns with Corcioli et al. [12], who denounce the allocation bias in favor of large producers, suggesting that the fragmentation of benefits explains part of the dissociation between GDP-AGRO growth and the regional HDI.

### ***5.3 Limitations of the study***

Four limitations deserve explicit acknowledgment. The first is that the dataset covers 465 observations across five fiscal years, a size determined by the very availability of the official federal government public bases, which cannot be expanded artificially. It is worth noting that there was full use of the available observations, without discarding valid records, and that the theoretical selection of features, by restricting the degrees of freedom of the modeling process, mitigates the risks associated with a dataset of moderate size. Even so, changes of government, climatic shocks, and revisions of agricultural policy occurring after 2022 may alter the importance hierarchy of the features, which recommends caution in the temporal generalization of the models. The second is that the model does not incorporate direct health and education variables, dimensions constitutive of the HDI, owing to their absence from the collection scope, so that their inclusion in future work could refine the diagnosis. The third is that the Agricultural Census data are from 2017, which introduces a relevant temporal lag in a sector of rapid dynamics, since the next census, under preparation by IBGE for the 2026/2027 cycle, is not yet available. The fourth, and most sensitive from an interpretive standpoint, is the strong dominance of the technical assistance variable, which accounts for 52.1% of the importance in the HDI model and 28.2% in the GDP-AGRO percentage model. The multicollinearity analysis revealed that the variables derived from the same technical assistance source (number of properties that receive and that do not receive guidance, in absolute and percentage values) are strongly collinear with one another, and that the technical assistance variable itself shows a correlation of 0.814 with the HDI in the GDP-AGRO percentage model. This pattern suggests that part of the high explanatory power may stem from an associative origin between technical assistance coverage and the structural development level of the states, rather than from a direct causal relationship. The high  $R^2$  should therefore be read as evidence of a robust association, not of causation, which motivates the causal investigation proposed in the future agenda.

## **6. THREATS TO VALIDITY**

### **- Internal validity**

The main threat lies in the possibility of omitted variables that correlate simultaneously with the features and with the target variables. To mitigate this risk, the study design adopted three complementary strategies.

- (i) First, multicollinearity was assessed already at the conception stage of the variable set, with the exclusion of redundant variables derived from the same source. The correlation matrix diagnosis ( $|r| > 0.70$ ), conducted separately for each model, confirmed the absence of collinear pairs in the HDI model set and identified a single pair in the GDP-AGRO percentage model set, between the HDI and the technical assistance variable, whose interpretation is addressed in construct validity and in the dominance discussed in Section 5.3.
- (ii) Second, the theoretical, hypothesis-driven selection restricted the set to variables with a prior conceptual basis, preventing the modeling process from capturing spurious correlations as explanatory power.
- (iii) Finally, L2 regularization was employed in the GDP-AGRO percentage model to confer stability on the estimates given the larger number of predictors.

#### - **Construct validity**

The equating of Projects with CAPEX and of Activities with OPEX, although grounded in Tahu et al. [16] and Puhach [8], constitutes an approximation, since the public budget follows logics distinct from those governing the corporate balance sheet. The conclusions about the importance of CAPEX should therefore be interpreted within this analogical context. It is worth noting, however, that this caveat is specific to the public data used in this study. Applied to a base of corporate origin, in which CAPEX and OPEX constitute native accounting categories rather than approximations, the same pipeline would preserve its analytical coherence, which shows that the limitation resides in the nature of the data, not in the method employed.

#### - **External validity**

The models were trained and tested with data from a single institution, MAPA, in a single country. Comparative analyses with the agriculture ministries of other countries with a similar agricultural profile, such as Argentina, the United States, and India, could broaden the generality of the conclusions. It is also worth highlighting that the pipeline architecture is not specific to the agricultural context, so that the same method can be extended to the budgetary analysis of other ministries, provided that the target variables and the feature set are duly adjusted to the domain of each portfolio. This transfer potential also reaches the private sector, as discussed in the construct validity analysis, where the accounting categories are natively defined. Taken together, these possibilities reinforce the potential of the method as a general data-driven decision-support tool, applicable both to public management and to corporate budget management.

#### - **Conclusion validity**

The models achieve high coefficients of determination, with an  $R^2$  of 0.8764 for the GDP-AGRO percentage and 0.9632 for the HDI, which confers statistical robustness on the conclusions about the relative importance of the features. A caveat applies regarding the dominance of the technical assistance variable (Section 5.3), since part of

the explanatory power may stem from an associative origin between technical assistance coverage and the structural development level, so that the coefficients of determination support claims of association, not of causation. The robustness analysis through automatic selection (Section 4.4) corroborates the stability of the importance hierarchy under a distinct selection philosophy. Beyond its immediate value, the study paves the way for the investigation of the causal relationships underlying the identified associations and the replication of the method in other institutional and sectoral contexts.

## 7. CONCLUSION

The results of this study showed that XGBoost, applied to MAPA open budget data integrated with variables from the 2017 Agricultural Census and the CEPEA/CNA series, produces predictive models with high explanatory power for the GDP-AGRO percentage ( $R^2 = 0.8764$ ) and for the HDI of the Brazilian states ( $R^2 = 0.9632$ ), answering the three proposed RQs affirmatively.

The main empirical contributions are:

- (i) rural technical assistance is the most influential factor for both targets, surpassing in importance traditional economic variables such as GDP and the budget amounts paid;
- (ii) investments in Projects/CAPEX are the second determinant of the GDP-AGRO percentage, with a relative importance of 21.59%, reinforcing their strategic role in resource allocation;
- (iii) geographic disparities explain 42.6% of the variation in the HDI, evidencing the insufficiency of uniform policies for reducing regional inequalities.

The limitations of the study, namely the five-year time window, the absence of direct health and education variables, and the lag of the Agricultural Census, define the agenda for future research:

- (i) the incorporation of climatic and environmental variables, whose relevance grows with the pressures of global warming on agricultural productivity;
- (ii) international comparisons with countries of a similar agricultural profile and the extension of the method to other ministries and to the private sector;
- (iii) the continuous updating of the models with annual data, ensuring their relevance for future budget cycles;
- (iv) the investigation of the causal relationships underlying the identified associations, confirming to what extent the expansion of technical assistance coverage produces measurable gains in HDI and GDP-AGRO.

In summary, the integration of machine learning techniques into federal budget planning represents a concrete methodological evolution toward more transparent, efficient, and data-driven public management. The results presented here offer a solid empirical basis for

managers and policymakers to prioritize rural extension and capital investments as fundamental levers of the sustainable development of Brazilian agribusiness.

## ACKNOWLEDGMENTS

The author thanks the University of Sao Paulo / ESALQ for its support of the 2024 MBA Program in Data Science & Analytics, as well as the entities that made the data available in open format: the Government of Brazil (Open Data Portal), IBGE, CEPEA, and the National Confederation of Agriculture and Livestock (CNA).

## REFERENCES

- [1] MINENHLE, H. AI in Finance: Challenges, Techniques, and Opportunities. ACM Computing Surveys, 2022. DOI: 10.1145/3502289.
- [2] CEPEA/ESALQ/USP; CNA. PIB do agronegócio brasileiro de 1996 a 2024. Available at: <https://www.cepea.esalq.usp.br>. Accessed on: 15 Jul. 2024.
- [3] GOVERNO DO BRASIL. Dados do Orçamento Federal – Orçamento Fiscal e da Seguridade Social (2018–2022). Available at: <https://dados.gov.br>. Accessed on: 15 Feb. 2024.
- [4] CAPONE, C. et al. Artificial Intelligence Models for Predicting Budget Expenditures. Eurasian Journal of Economic and Business Studies, 2024.
- [5] DARMAWAN, H.; YULIANA, M.; HADI, M.Z.S. GRU and XGBoost Performance with Hyperparameter Tuning Using GridSearchCV and Bayesian Optimization on an IoT-Based Weather Prediction System. International Journal on Advanced Science, Engineering and Information Technology, 2023.
- [6] CASTRO, J. MBA\_USP – Repositório público de código e dados. GitHub, 2024. Available at: [https://github.com/Jotta-se/MBA\\_USP](https://github.com/Jotta-se/MBA_USP).
- [7] HAYTHAM, H.; ELMOUSALAMI. Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction. arXiv: Learning, 2019.
- [8] PUHACH, Y.V. Modern econometric approaches of machine learning in the simulation of factors influencing local budget expenditures. Actual Problems of Economics, 2023.
- [9] ZAINAB, H.A. et al. Developing an Integrative Data Intelligence Model for Construction Cost Estimation. Complexity, 2022. DOI: 10.1155/2022/4285328.
- [10] TORRES, D.A.P. Investimento em pesquisa e a produtividade agropecuária: uma revisão da literatura recente. Revista de Política Agrícola, v. 32, p. 106-120, 2023.
- [11] DIALLO, M.; WOUTERSE, F. Agricultural development promises more growth and less poverty in Africa. Development Policy Review, 2022. DOI: 10.1111/dpr.12669.
- [12] CORCIOLI, G.; MEDINA, G.S.; ARRAIS, C.A. Missing the Target: Brazil's Agricultural Policy Indirectly Subsidizes Foreign Investments. Frontiers in Sustainable Food Systems, 2022.
- [13] PIRAS, S.; WESZ, V.J.; GHINOI, S. Soy Expansion, Environment, and Human Development: An Analysis across Brazilian Municipalities. Sustainability, v. 13, n. 13, 2021. DOI: 10.3390/su13137246.
- [14] SOLEMAN, R. et al. The Effect of HDI and Macroeconomic Variables on Economic Growth in Indonesia 2015–2020. Jurnal Ilmu Ekonomi Terapan, 2022.
- [15] ONDETTI, G. Ideational Bases of Land Reform in Brazil: 1910 to the Present. 2020.

- [16] TAHU, G.P. et al. The Influence of Regional Original Income and General Allocation Funds on Economic Growth with Capital Expenditures as an Intervening Variable. *Revista de Gestão Social e Ambiental*, v. 18, 2024.
- [17] GOVERNO DO BRASIL. Acompanhamento Físico-Financeiro das Ações Orçamentárias. Available at: <https://dados.gov.br>. Accessed on: 15 Feb. 2024.
- [18] IBGE. Censo Agropecuário 2017: Resultados definitivos. Rio de Janeiro: IBGE, 2019.
- [19] YULIANA; PRIHATININGSIH, A.; ISPA, A.Y. Dynamics Of Economic Development Disparities: A New Perspective In Regional Analysis. 2024.
- [20] MONTEIRO, J.J. et al. Reflexos da COVID-19 no orçamento público do governo federal. *Revista de Gestão e Organizações*, v. 14, 2021.
- [21] SILVA, L.A.P. et al. Vetores de degradação das pastagens no Cerrado norte mineiro. *Raega – O Espaço Geográfico em Análise*, v. 57, p. 66-80, 2023.
- [22] REIS, C.C. et al. Sistemas integrados de lavoura-pecuária-floresta no Cerrado. 2020.