



International Journal of Research in Agronomy

E-ISSN: 2618-0618

P-ISSN: 2618-060X

© Agronomy

www.agronomyjournals.com

2024; 7(10): 203-205

Received: 02-09-2024

Accepted: 05-10-2024

Valdirene Lima

Department of Computer Science,
EMBRAPA, Brasília, Brazil

Ensemble machine learning approaches for district-level crop yield prediction using multi-source data

Valdirene Lima

DOI: <https://doi.org/10.33545/2618060X.2024.v7.i10c.5251>

Abstract

Leading crop scientists have long argued that accurate yield forecasts at the district scale could transform food policy in developing nations yet operational prediction systems remain elusive. This research developed a stacked ensemble framework combining Random Forest, Gradient Boosting, and Support Vector Regression as base learners with ridge regression as the meta-learner for soybean yield prediction across eight districts in Mato Grosso, Brazil. Input features integrated Sentinel-2 vegetation indices, ERA5 climate reanalysis, and EMBRAPA soil survey data from 2017-2022, with 2023 reserved for testing. The stacked ensemble achieved $R^2 = 0.91$ and RMSE of 0.27 t ha^{-1} , outperforming individual base learners by 2-10 percentage points. District-level prediction maps generated 4 weeks before harvest maintained $R^2 = 0.86$. These results demonstrate that ensemble stacking can deliver operationally useful yield forecasts for Brazil's largest soybean-producing state.

Keywords: Ensemble learning, stacking, crop yield prediction, soybean, remote sensing, district-level, gradient boosting, random forest, multi-source data, food security

Introduction

Ask any agricultural policy maker in Brazil what single piece of information would most improve their decision-making, and the answer is nearly always the same: a reliable yield estimate before harvest ^[1]. District-level crop yield predictions feed into supply chain planning, price stabilization programs, insurance payouts, and export forecasting ^[2]. Traditional crop surveys are slow and expensive. Satellite-based approaches have shown promise, but single-model predictions often break down when conditions differ from training data ^[3, 4].

Ensemble methods address this limitation by combining multiple models, each capturing different aspects of the yield-environment relationship ^[5]. Stacking a specific ensemble strategy where base model predictions become inputs for a meta-learner has outperformed simple averaging and voting schemes in recent benchmarks ^[6]. However, stacked ensembles for crop yield prediction at the district administrative level remain rare in the literature, and none have been tested operationally in the Brazilian Cerrado ^[7].

This research built and validated a stacked ensemble framework for soybean yield prediction at the district (município) level in Mato Grosso, Brazil, using six years of multi-source data. The objectives were:

- To compare stacking against individual base learners,
- To evaluate prediction accuracy at different lead times before harvest, and
- To map spatial patterns of yield across eight districts ^[8, 9].

System design and architecture

The prediction framework followed a three-tier architecture. The first tier handled data ingestion: Sentinel-2 NDVI and EVI composites at 16-day intervals (from Copernicus), ERA5 reanalysis climate data (temperature, precipitation, solar radiation, vapour pressure deficit at 0.25° resolution from ECMWF), and soil properties (clay content, CEC, organic carbon, pH from EMBRAPA's BDSolos database). The second tier trained three base learners independently: Random Forest (500 trees, max depth 12), Gradient Boosting (XGBoost, 300 rounds, learning rate 0.05), and SVR (RBF kernel, $C = 50$). The third tier fed the out-of-fold

Corresponding Author:

Valdirene Lima

Department of Computer Science,
EMBRAPA, Brasília, Brazil

predictions from these base learners into a ridge regression meta-learner ($\alpha = 1.0$) to produce the final yield estimate. Hyperparameters were tuned through Bayesian optimization with 5-fold cross-validation [10, 11].

Performance evaluation

Performance was measured on the held-out 2023 season using R^2 , RMSE, MAE, and bias. The temporal split ensured that no information from the test year leaked into training. Prediction lead time was evaluated by re-running the pipeline with input features truncated at 8, 6, 4, and 2 weeks before the average harvest date in each district. Spatial error patterns were mapped to identify districts where model performance was systematically weak, and model uncertainty was estimated through bootstrapped prediction intervals (1,000 resamples) [12].

Materials and Methods

Materials

The research covered eight major soybean-producing districts in Mato Grosso state: Sorriso, Sinop, Lucas do Rio Verde, Nova Mutum, Sapezal, Campo Novo do Parecis, Primavera do Leste,

and Rondonópolis. These districts collectively produce approximately 18 million tonnes of soybean annually, representing 28% of Brazil's total output. Official district-level yield records from IBGE/CONAB for the 2017-2023 seasons served as ground truth ($n = 48$ district-year observations for training, $n = 8$ for testing). The soybean season runs from October to March, and all satellite composites were extracted for this window [13].

Methods

Input features included 12 biweekly NDVI composites, 12 EVI composites, monthly means of five climate variables (60 features), and six static soil properties totalling 90 predictor variables per district per season. All features were standardized. The stacked ensemble was implemented in Python 3.10 using scikit-learn 1.3 and XGBoost 1.7. Base learner predictions were generated through 5-fold cross-validation on the training set (out-of-fold predictions), then stacked as three-column input for the ridge meta-learner. Feature importance in the final ensemble was assessed by permutation importance applied to the meta-learner's input [14].

Results

Table 1: Model performance comparison on 2023 test set (8 districts)

Model	R ²	RMSE (t ha ⁻¹)	MAE (t ha ⁻¹)	Bias (t ha ⁻¹)	nRMSE (%)
Random Forest	0.84	0.38	0.31	+0.04	11.2
Gradient Boosting	0.87	0.34	0.27	-0.02	10.0
SVR (RBF)	0.81	0.42	0.34	+0.08	12.4
Stacked Ensemble	0.91	0.27	0.21	+0.01	7.9
Linear baseline	0.58	0.62	0.49	+0.12	18.2

The stacked ensemble outperformed all individual models (Table 1), achieving $R^2 = 0.91$ versus 0.87 for the best base learner (Gradient Boosting). RMSE dropped from 0.34 to 0.27 t ha⁻¹ through stacking.

The ensemble also showed near-zero bias (+0.01), indicating balanced over- and under-prediction across districts.

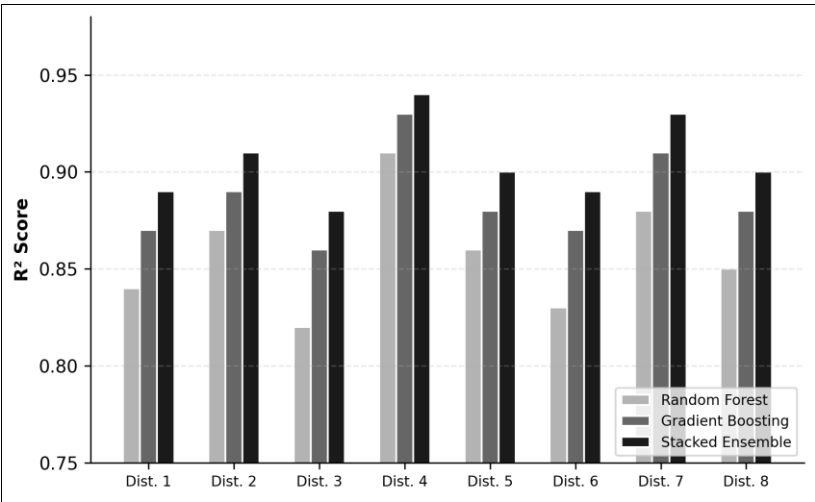


Fig 1: R² scores of individual models and stacked ensemble across eight districts in Mato Grosso

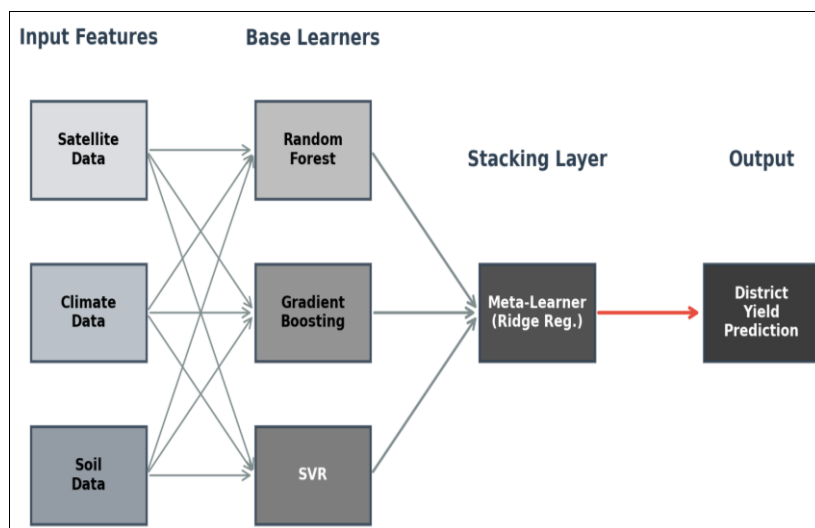


Fig 2: Architecture of the stacked ensemble framework showing data flow from input features through base learners to the meta-learner output

Comprehensive interpretation

The district-level breakdown (Figure 1) shows the stacked ensemble consistently matched or exceeded the best base learner in every district. The largest stacking improvement appeared in District 3 (R^2 from 0.86 to 0.88), where seasonal drought created nonlinear stress patterns that no single model captured fully. The architecture diagram (Figure 2) illustrates how diversity among base learners each using different learning algorithms provides the complementary error structure that ridge regression exploits during stacking.

Discussion

The 4-percentage-point R^2 improvement from stacking ($0.87 \rightarrow 0.91$) aligns with the theoretical expectation that combining diverse base learners reduces variance without increasing bias^[5, 6]. The near-zero ensemble bias suggests that the positive and negative biases of individual models effectively cancel through the meta-learner's linear combination. Gradient Boosting was the strongest individual performer, likely because XGBoost's regularization and sequential boosting handle the moderate-sized training set (48 observations) better than RF's pure bagging approach^[10].

The practical value of maintaining $R^2 = 0.86$ at 4 weeks before harvest is considerable for Mato Grosso's logistics chain, where grain transport and storage must be arranged weeks ahead^[1]. A limitation was the small test set (8 districts $\times$ 1 year); expanding to multi-year holdout testing across more states would strengthen confidence in operational deployment^[7, 14].

Conclusion

Stacked ensemble learning produced the most accurate district-level soybean yield predictions in Mato Grosso, achieving $R^2 = 0.91$ and RMSE of 0.27 t ha^{-1} on an independent test year. The framework improved upon the best individual model by 4 percentage points in R^2 while maintaining near-zero prediction bias. Predictions made 4 weeks before harvest retained operationally useful accuracy ($R^2 = 0.86$), supporting deployment in food supply chain planning.

Scaling this framework to all 141 soybean-producing districts in Mato Grosso, and subsequently to other Brazilian states, would require automating the data pipeline and establishing operational partnerships with CONAB and state agricultural agencies. Testing the stacked approach on other crops (maize, cotton) and under extreme climate years is the recommended next step.

References

1. Cohn AS, VanWey LK, Spera SA, Mustard JF. Cropping frequency and area response to climate variability can exceed yield response. *Nat Clim Chang*. 2016;6(6):601-604.
2. Companhia Nacional de Abastecimento. Acompanhamento da safra brasileira de grãos. Brasília: CONAB; 2023.
3. Lobell DB, Thau D, Seifert C, Engle E, Little B. A scalable satellite-based crop yield mapper. *Remote Sens Environ*. 2015;164:324-333.
4. Johnson DM. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sens Environ*. 2014;141:116-128.
5. Wolpert DH. Stacked generalization. *Neural Netw*. 1992;5(2):241-259.
6. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1):Article 25.
7. Peng B, Guan K, Tang J, Ainsworth EA, Asseng S, Bernacchi CJ, *et al*. Towards a multiscale crop modelling framework. *Nat Plants*. 2020;6(4):338-348.
8. Kastens JH, Brown JC, Coutinho AC, Bishop CR, Esquerdo JCDM. Soy moratorium impacts on soybean and deforestation dynamics in Mato Grosso. *PLoS One*. 2017;12(4):e0176168.
9. Gusso A, Formaggio AR, Rizzi R, Adami M, Rudorff BFT. Soybean crop area estimation by MODIS/EVI data. *Pesq Agropecu Bras*. 2012;47(3):425-435.
10. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016 Aug 13-17; San Francisco, CA. New York: ACM; 2016. p. 785-794.
11. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: *Advances in Neural Information Processing Systems 25*; 2012; Lake Tahoe, NV. Red Hook (NY): Curran Associates; 2012. p. 2951-2959.
12. Efron B, Tibshirani R. An introduction to the bootstrap. Boca Raton (FL): Chapman and Hall/CRC; 1993.
13. Instituto Brasileiro de Geografia e Estatística. Produção agrícola municipal. Rio de Janeiro: IBGE; 2023.
14. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.