Machine learning and causal analysis to support improved crop decision-making

Marcelo Chan Fu Wei[1], José Paulo Molin[2], André Freitas Colaço[3], Louis Longchamps[4]

[1]marcelochan@alumni.usp.br (independent researcher, Brazil), [2]jpmolin@usp.br (professor, University of São Paulo, Brazil), [3]andre.colaco@usp.br (professor, University of São Paulo, Brazil), [4]ll928@cornell.edu (professor, Cornell University, United States of America)

A paper from the Proceedings of the
17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture
13-15 July 2026
Porto Alegre, Brazil

Abstract.

While machine learning (ML) models, particularly Extreme Gradient Boosting (XGBoost) and Random Forest (RF), have demonstrated potential in generating accurate crop yield predictions, their practical adoption for on-farm decision support remains limited. A key challenge lies in their fundamentally associative nature, which, without additional tools, can reduce interpretability and diminish practitioner confidence. Explainable Artificial Intelligence (XAI) techniques like SHAP values address one facet of this issue by providing visibility into variable contributions. Separately, causal discovery methods offer another powerful lens by identifying potential cause-effect pathways. However, even with these tools, there is no consensus on a single "most proper" model to use, and relying on any one approach—predictive or explanatory—can lead to an incomplete or biased understanding. To overcome this, our study posits that a more robust foundation is achieved through the convergence of evidence across multiple analytical paradigms. Thus, the aim of this study is to demonstrate an integrated analytical framework that combines associative and causal discovery methods to robustly identify key drivers of crop yield. This study applied both XGBoost and RF models to a comprehensive, long-term (2012–2016) dataset from a 250-ha citrus field. This dataset is composed of yield, soil chemical and physical properties, rootstock and graft varieties, yield stability zones (YSZ) and rainfall variables, originally from an experiment with fixed and variable fertilizer treatments. Variable importance was assessed using both native model metrics (e.g., Gini importance or Mean Decrease in Impurity) and SHAP values to ensure analytical robustness. Subsequently, a causal discovery algorithm was applied to the same dataset to independently construct a causal network and identify potential cause-effect pathways. Relying on a single analytical approach may lead to over- or under-representation of variable importance. A multi-method framework offers a more comprehensive perspective by providing richer insight into the underlying structure of the data and the factors influencing yield outcomes. XGBoost identified year, YSZ,

February rainfall, rootstock type, and March rainfall as the five most important variables, while SHAP analysis highlighted year, YSZ, February rainfall, graft, and treatment. In comparison, RF variable importance emphasized YSZ, treatment, rootstock, graft, and sand content, whereas SHAP analysis pointed to YSZ, April, June, and February rainfall, and year as the most influential features. These results illustrate that different modeling approaches can result in divergent rankings of variable importance, which may pose challenges for the consistent interpretation and application of ML in agricultural contexts. However, by examining convergence across methods, certain variables—such as YSZ and year—consistently emerged as influential, regardless of the analytical technique employed. Causal analysis further reinforced these findings, identifying year, YSZ, rootstock type, and February rainfall as factors directly linked to yield outcomes. The alignment of results across predictive models and causal structures helps distinguish stable and plausibly influential drivers from spurious correlations. This methodological synthesis contributes to agricultural data science by moving beyond reliance on any single modeling approach, offering a more rigorous pathway toward extracting reliable, interpretable, and causally informed insights to support decision-making in agricultural systems.

Keywords.
Explainable artificial intelligence, Extreme Gradient Boosting, Random Forest, Yield

Introduction

While machine learning (ML) models, particularly Extreme Gradient Boosting (XGBoost) and Random Forest (RF), have demonstrated potential in generating accurate crop yield predictions (Alsaber et al. 2025; Jahan et al. 2026; Torgbor et al. 2023), their practical adoption for on-farm decision support remains limited. A key challenge lies in their fundamentally associative nature, which, without additional interpretive tools, can reduce transparency and diminish practitioner confidence in the resulting recommendations as most of the models suffer from low interpretability (Abdullah et al. 2021; Breiman 2001). Explainable Artificial Intelligence (XAI) techniques, such as SHapley Additive exPlanations (SHAP), address one facet of this issue by providing visibility into variable contributions and enabling users to assess which features drive model predictions. Separately, causal discovery methods offer another powerful lens by identifying potential cause-effect pathways within agricultural systems, moving beyond mere correlation toward mechanistic understanding.

However, even with these interpretive advances, a fundamental challenge persists: there is no consensus on a single "most proper" model for crop yield analysis, and relying on any one approach—whether purely predictive or explanatory—can lead to an incomplete or potentially biased understanding. Predictive performance does not necessarily align with interpretability, and in most cases, models optimized for prediction may obscure the underlying relationships of interest to domain experts. This disconnect extends even to predictive performance itself, where different modeling approaches applied to identical datasets can yield divergent assessments of variable importance (Wei et al. 2026b), leaving practitioners uncertain about which insights to trust.

This lack of methodological consensus is particularly consequential in agricultural contexts, where decisions based on variable importance rankings directly influence management practices, resource allocation, and long-term planning. When different models emphasize different drivers—for instance, one model highlighting soil properties while another prioritizes climatic variables—practitioners face ambiguity in determining where to focus interventions. Instead of searching for a single best model, a more reasonable strategy lies in examining the convergence of evidence across multiple analytical processes. Overlapping findings that persist across predictive algorithms and causal discovery methods are more likely to represent stable, biologically meaningful relationships rather than artifacts of any particular modeling choice.

Despite the growing availability of ML-based yield prediction studies, the explicit integration of associative models with causal discovery remains rare in agricultural research. Most studies select a single "best-performing" model and report its feature importance as definitive, without examining whether those findings are robust across algorithms or whether associative importance reflects genuine causal influence (Sánchez et al. 2025). This practice risks overconfidence in model-specific patterns that may not generalize. Furthermore, while SHAP has been applied in agriculture, e.g., for yield prediction in sugarcane (Zhu et al. 2023), causal discovery methods—such as Bayesian networks—have been used primarily in ecological or climate science contexts and remain underexploited for on-farm decision support. Causal inference in agriculture is recognized as important because it can provide more robust and transparent information (Tsoumas et al. 2025). Nonetheless, few studies have applied this type of analysis in agriculture. The present study addresses this gap by intentionally deploying multiple algorithms and explanatory techniques alongside an independent causal network, using convergence across paradigms—not predictive performance alone—as the criterion for identifying trustworthy yield drivers.

Thus, the aim of this study is to demonstrate an integrated analytical framework that combines associative machine learning with causal discovery methods to identify key drivers of crop yield. A multi-analytical approach was applied to a comprehensive, long-term (2012–2016) dataset from

a 250-ha citrus field, encompassing yield measurements, soil chemical and physical properties, rootstock and graft varieties, yield stability zones (YSZ), and rainfall variables. YSZ are regions classified according to their temporal yield patterns into four categories: high-yielding and stable, high-yielding and unstable, low-yielding and stable, and low-yielding and unstable (Marcaida III et al. 2022). By comparing variable importance rankings derived from XGBoost and RF—using both native importance metrics and SHAP values—alongside an independently constructed causal network, the study identified variables that consistently emerged as influential across methodological boundaries and are thus more likely to represent actionable agronomic knowledge. This framework offers a rigorous pathway to move beyond the question of "which model is best" toward the more actionable question of "which variables are most trustworthy," providing a foundation for more reliable, interpretable, and causally informed decision support in agricultural systems.

Material and Methods

Study Site and Experimental Design

This study utilized a database from a 5-year citrus experiment (2012–2016) conducted in a 250-ha commercial field, sourced from the Precision Agriculture Laboratory (LAP) at the University of São Paulo – "Luiz de Queiroz" College of Agriculture. The experimental area comprised 10 groves of approximately 25 ha each, with five groves assigned to fixed rate (FR) fertilization and the remaining five to variable rate (VR) fertilization for nitrogen (N), phosphorus (P), potassium (K), and lime. FR treatments followed the standardized methodology of Van Raij et al. (1997), while VR applications were implemented using the spatially adaptive framework proposed by Colaço and Molin (2017). While assessing the effectiveness of VR fertilization was not the primary objective of this study, the two treatments served as a source of management variation to interpret whether such variation affected yield stability. Yield stability zones were incorporated as described in Wei et al. (2026a); the present study extends this foundation by integrating machine learning, explainable artificial intelligence, and causal discovery methods to enhance information extraction and identify key yield drivers.

Data Collection and Integration

Yield mapping followed the methodology described by Colaço et al. (2020). Big bags used during manual harvest were georeferenced, and yield values were calculated based on individual fruit mass and representative area. These point data were subsequently interpolated into a continuous surface using ordinary kriging with a 10 m × 10 m pixel grid.

Soil samples were collected annually on a 1-ha grid (0.00–0.20 m depth), totaling 250 samples per year. Each composite sample comprised six subsamples (three per side of the tree canopy projection) and was analyzed for physical properties (clay, silt, and sand content) and chemical properties: pH ($CaCl_2$), organic matter (OM), available P, exchangeable K, Ca, Mg, H+Al, sum of bases (SB), cation exchange capacity (CEC), and base saturation (V%). Monthly rainfall data were recorded via a local rain gauge.

To merge the high-density yield maps with the low-density soil sampling grids, a spatial fusion protocol was implemented. Low-density soil sampling points served as reference locations, with a 20 m radius buffer (1,256.64 m², equivalent to two citrus tree rows) applied to extract co-located yield values from the interpolated yield surfaces. This approach ensured spatial congruence between datasets while accounting for field-scale variability.

Machine Learning and Explainable Artificial Intelligence

Two ensemble machine learning algorithms (RF and XGBoost) were implemented to model citrus yield as a function of soil properties, management practices, climatic variables, and crop characteristics: Both models were trained using an 80/20 train-test split, with the training partition

further subjected to 5-fold cross-validation for hyperparameter tuning. All analyses were conducted in R (R Core Team, 2019).

For XGBoost, categorical variables (treatment, rootstock, graft type, and YSZ class) were transformed using one-hot encoding, converting each factor level into binary indicator columns. RF handled categorical variables natively without encoding. The target variable (yield) was treated as continuous in both models.

XGBoost was fined tuned over the following parameter space: learning rate (eta: 0.05, 0.1), maximum tree depth (3, 6), subsample ratio (0.8), column sampling ratio (0.8), and regularization parameters (gamma = 0, lambda = 1, alpha = 0). Model performance was evaluated using 5-fold cross-validation on the training set with root mean squared error (RMSE) as the optimization metric. Early stopping was applied after 10 consecutive rounds without improvement. The best parameter combination was selected based on minimum cross-validation RMSE, and the final model was trained with the optimal number of boosting rounds. RF hyperparameter tuning focused on mtry (the number of variables randomly sampled as candidates at each split), with candidate values set at the square root of the total number of predictors and one-third of the total predictors. Model selection employed 5-fold cross-validation with RMSE as the optimization metric. The final model was trained using 500 trees with a minimum node size of 5 observations.

Predictive performance was assessed on the held-out test set using four metrics: coefficient of determination (R^2), RMSE, mean absolute error (MAE), and Lin's concordance correlation coefficient (CCC). These metrics were calculated for both training and test partitions to evaluate potential overfitting. Variable importance was assessed through two complementary approaches. (A) native model-specific metrics were extracted: Gain (average improvement in accuracy brought by a feature across all splits) for XGBoost, and percent increase in mean squared error (%IncMSE) for RF. For XGBoost, importance scores from one-hot encoded features were aggregated to their original variables to enable direct comparison with RF results. (B) SHAP values were computed using 30 Monte Carlo repetitions on a random subset of 500 training observations. SHAP importance was quantified as both the mean absolute SHAP value (magnitude of influence) and the mean signed SHAP value (direction of influence), and the top 15 features were visualized using bar plots and bee swarm plots.

Causal Discovery

A causal discovery analysis was conducted independently of the machine learning pipeline to identify potential cause-effect relationships among variables. A Bayesian network was learned from the complete dataset with the Bayesian Information Criterion for mixed continuous-discrete data as the optimization score. To respect the directional constraint that yield is an outcome rather than a cause, a blacklist was specified preventing yield from serving as a parent node to any other variable. The learned network structure was visualized with edge colors and labels indicating the direction and sign of relationships where calculable: positive correlations were shown in blue (+), negative correlations in orange (−), and correlations involving categorical variables—where direction cannot be unambiguously determined—were shown in gray with a question mark (?). This causal network provided an independent perspective on variable relationships, complementing the associative importance rankings derived from the machine learning models.

Methodological Rationale

This multi-method framework was designed to address the inherent limitations of relying on any single analytical approach. While XGBoost and RF capture potentially complex nonlinear relationships between predictors and yield, their fundamentally associative nature means that high importance scores do not necessarily imply causal influence. Conversely, the Bayesian network provides a graphical representation of conditional dependencies but does not match the predictive flexibility of tree-based ensembles. SHAP values bridge these paradigms by decomposing individual predictions into feature contributions, offering both global importance rankings and local interpretability. By triangulating evidence across these methods, variables that

consistently emerge as influential—regardless of the analytical lens applied—can be identified with greater confidence as robust drivers of citrus yield.

Results and Discussion

The application of the proposed multi-method analytical framework—combining XGBoost, RF, SHAP, and causal discovery—revealed variation in variable importance rankings depending on the modeling approach, while also identifying a subset of yield drivers that converged across methods. These findings directly address the central challenge outlined in the introduction: that reliance on any single analytical paradigm may produce incomplete or biased conclusions.

Predictive Performance and Feature Importance from Machine Learning Models

Both XGBoost and RF achieved satisfactory predictive performance on the held-out test set (Table 1) as their predictive power is well reported in the literature regardless of crop type, dataset, spatial and temporal cross-validation (Alsaber et al. 2025; Jahan et al. 2026; Torgbor et al. 2023), consistent with previous studies demonstrating the suitability of ensemble tree-based methods for crop yield modeling. However, the primary value of the current analysis lay not in prediction accuracy but in comparing variable importance rankings across models and explanation techniques.

Table 1. Performance metrics of Random Forest and Extreme Gradient Boosting for citrus database.

Model	Dataset	R^2	RMSE (Mg ha^{-1})	MAE (Mg ha^{-1})	CCC
Random Forest	Training (80%)	0.97	2.63	1.91	0.98
	Test (20%)	0.86	5.72	4.28	0.92
Extreme Gradient Boosting	Training (80%)	0.97	2.84	2.04	0.98
	Test (20%)	0.85	5.75	4.21	0.92

R^2: coefficient of determination, RMSE: root mean squared error, MAE: mean absolute error, and CCC: Lin's concordance correlation coefficient (CCC)

Machine Learning and Feature Importance

XGBoost (Figure 1) identified year, YSZ, February rainfall, graft type, and March rainfall as the five most important features based on GAIN importance. In contrast, SHAP analysis applied to the same XGBoost model highlighted year, YSZ, February rainfall, rootstock, and fertilizer treatment as the top five drivers. While year and YSZ remained consistently important across both metrics, the ranking of other variables diverged markedly: graft type ranked higher in GAIN importance, whereas rootstock ranked higher in SHAP importance. This divergence reflects a fundamental distinction between the two metrics—GAIN measures the average improvement in predictive accuracy contributed by a feature across all splits, while SHAP quantifies the average marginal effect of a feature on the predicted outcome. Thus, a feature may have high predictive utility (GAIN) but low marginal impact (SHAP), or vice versa, depending on its correlation structure and interaction patterns with other variables. Random Forest (Figure 2) produced yet another distinct ranking. Native variable importance (%IncMSE) emphasized YSZ, treatment, graft, rootstock, and sand content. SHAP analysis for RF pointed to YSZ, April rainfall, June rainfall, February rainfall, and year as the most influential features. Notably, rainfall in different months (April and June) emerged prominently in the RF-SHAP ranking but were less prominent in XGBoost-based rankings, suggesting that RF may be more sensitive to late-season rainfall patterns or that these effects are non-linear and better captured by the random forest algorithm.

These discrepancies illustrate a key finding: variable importance is not model-invariant. A practitioner relying solely on XGBoost GAIN might conclude that graft type is critical while rootstock is negligible, whereas XGBoost SHAP would suggest the opposite pattern. Such divergence poses a genuine risk for decision-making if a single method is treated as authoritative.

The beeswarm plots (Figures 1D and 2D) help resolve these apparent contradictions by revealing not just the magnitude but also the direction and consistency of each feature's effect on yield. For year, both XGBoost and RF beeswarm plots show a clear gradient: yellow points (higher years, e.g., 2015–2016) are consistently associated with positive SHAP contributions, while purple points (lower years, e.g., 2012–2013) show negative SHAP values. This indicates that more recent growing seasons produced higher yields—a pattern that RF's native %IncMSE metric completely failed to capture, as it did not rank year among the top 15 features. This omission underscores a critical limitation of native importance metrics: they measure predictive contribution within the model's specific structure but do not necessarily reflect the true marginal effect of a feature on the outcome.

Similarly, February rainfall shows a consistent positive directional effect in both beeswarm plots: yellow points (high rainfall) predominantly appear on the positive SHAP side, while purple points (low rainfall) cluster on the negative side. This alignment across XGBoost and RF, despite differences in their native importance rankings, further demonstrates the value of SHAP-based visualization for identifying robust drivers. Together, these observations highlight the central methodological recommendation of this study: no single metric or model should be trusted in isolation. Convergence across multiple analytical lenses—including different algorithms (XGBoost vs. RF), different importance metrics (GAIN/%IncMSE vs. SHAP), and different visualization approaches (bar plots vs. beeswarm)—provides a more reliable foundation for identifying influential yield drivers than any individual approach alone.

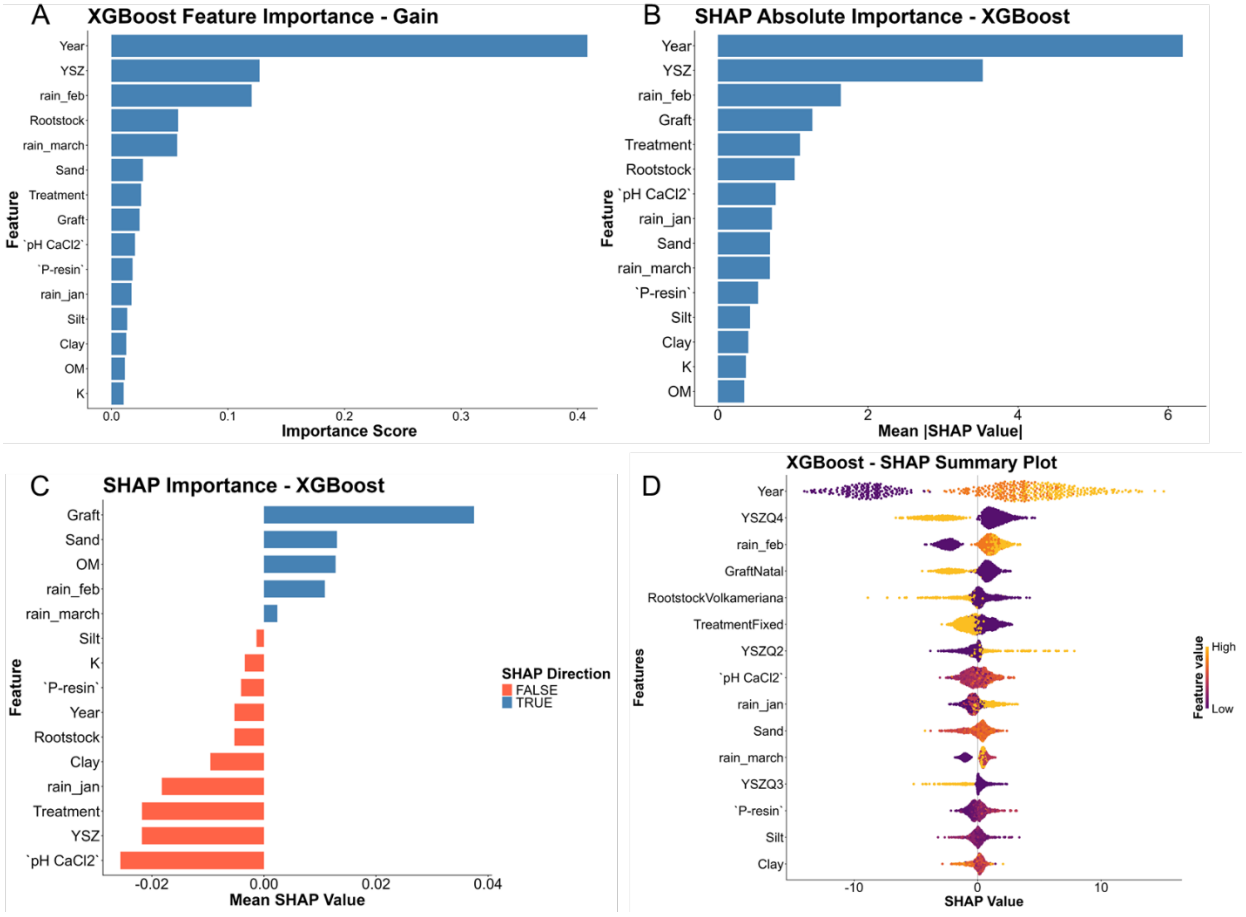

Figure 1. Extreme Gradient Boosting feature importance by (A) gain metric, (B) SHapley Additive Explanation (SHAP) mean absolute importance, (C) mean signed SHAP and (D) beeswarm plot.

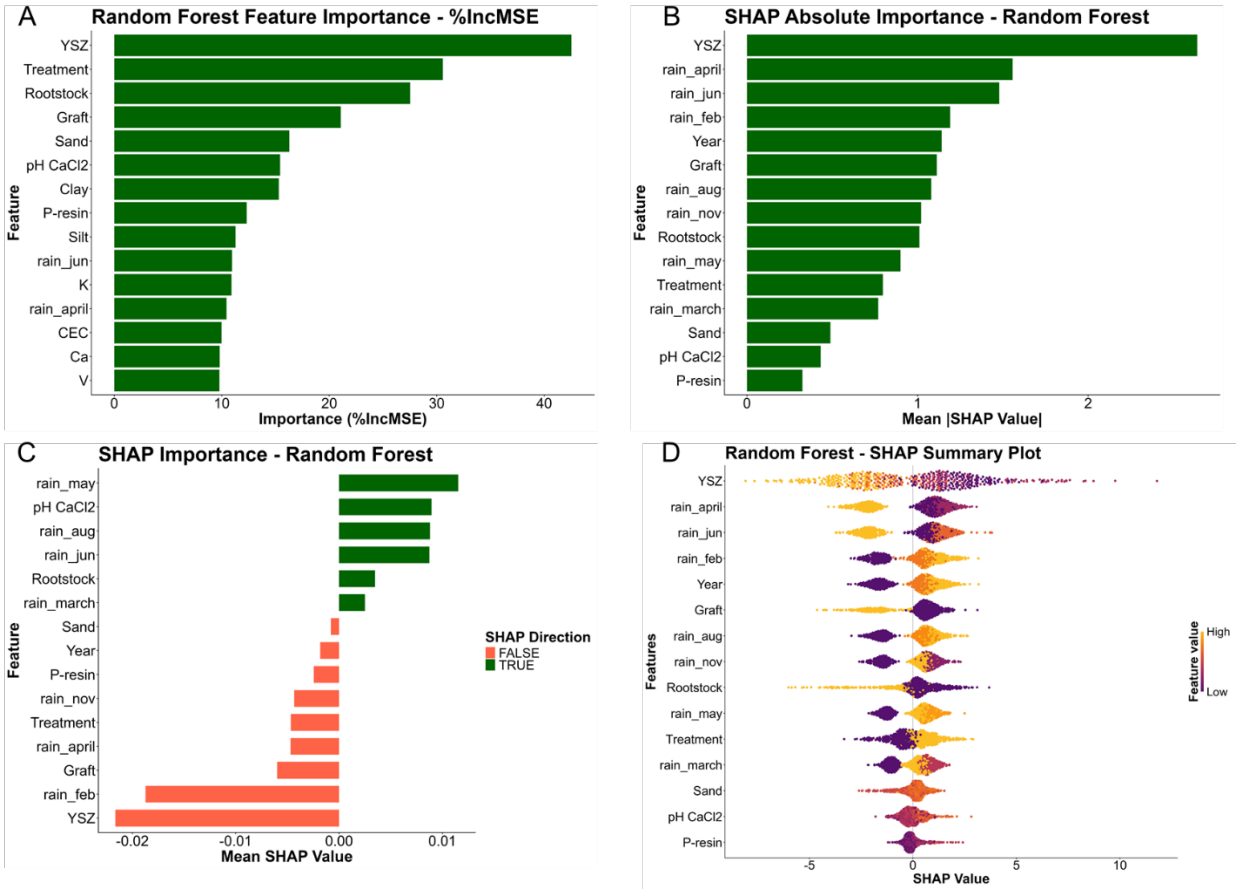

Fig 2. Random Forest feature importance by (A) percent increase in mean squared error (%IncMSE), (B) SHapley Additive Explanation (SHAP) mean absolute importance, (C) mean signed SHAP and (D) beeswarm plot.

Convergence Across Methods: Identifying Robust Drivers

Despite the observed methodological variability, certain variables consistently ranked highly across all four analytical lenses (XGBoost GAIN, XGBoost SHAP, RF %IncMSE, RF SHAP). Yield stability zone and year were among the top five in every model and explanation combination. This convergence is not trivial; it suggests that these two variables exert a consistently strong associative influence on yield regardless of algorithmic choices or importance metrics. February rainfall also appeared prominently in three of the four rankings (all except RF %IncMSE), while graft type and rootstock each appeared in two rankings, but rarely together in the same top-five set. This pattern indicates that while some agronomic factors (graft/rootstock) are sensitive to modeling choices, YSZ and year represent stable, robust predictors results that were already shown in a simple decision tree analysis regardless of its predictive performance (Wei et al. 2026a).

The robustness of YSZ as a predictor aligns with previous findings from Wei et al. (2026a), who demonstrated that stability zones derived from historical yield maps can enhance decision-making process as it provide an extra layer of data. The present study extends this result by showing that YSZ remains influential across multiple ML algorithms (XGBoost, RF) and importance metrics (GAIN, %IncMSE, SHAP), and—crucially—that it has a direct causal edge to yield in the Bayesian network. This convergence elevates YSZ from a useful predictor to a trustworthy management lever. In contrast, variables such as rootstock and treatment, which showed high importance in some metrics but lacked causal edges, illustrate the risk of relying on associative importance alone—a concern previously raised in the broader ML literature (Rudin 2019)

Causal Discovery Provides Mechanistic Corroboration

The causal discovery analysis (Figure 3) was conducted independently of the machine learning pipeline and without access to importance rankings. The resulting Bayesian network provided a directed graphical model of conditional dependencies among variables. Importantly, the causal analysis identified year, YSZ, rootstock type, and February rainfall as factors with direct directed edges to yield. No direct edge from graft to yield was found, nor from treatment (fertilizer strategy) to yield in the final causal graph. This potential causal network aligns with well reported importance of rootstock choice in citrus production as a key driver of yield (Mafrica et al. 2026). Beyond confirming expected relationships, causal discovery analysis can reveal non-intuitive associations that generate new hypotheses for future research. For example, the network shows an association between graft and May rainfall that was not anticipated a priori. While the mechanism underlying this relationship is not immediately obvious, it may reflect phenological interactions or unmeasured confounding variables. Such unexpected findings illustrate the value of causal discovery not only for corroboration but also for hypothesis generation—a contribution that purely associative ML models cannot provide. Nonetheless, causal networks learned from observational data are not definitive; they should be interpreted alongside domain knowledge and experimental validation, as some associations may be non-causal.

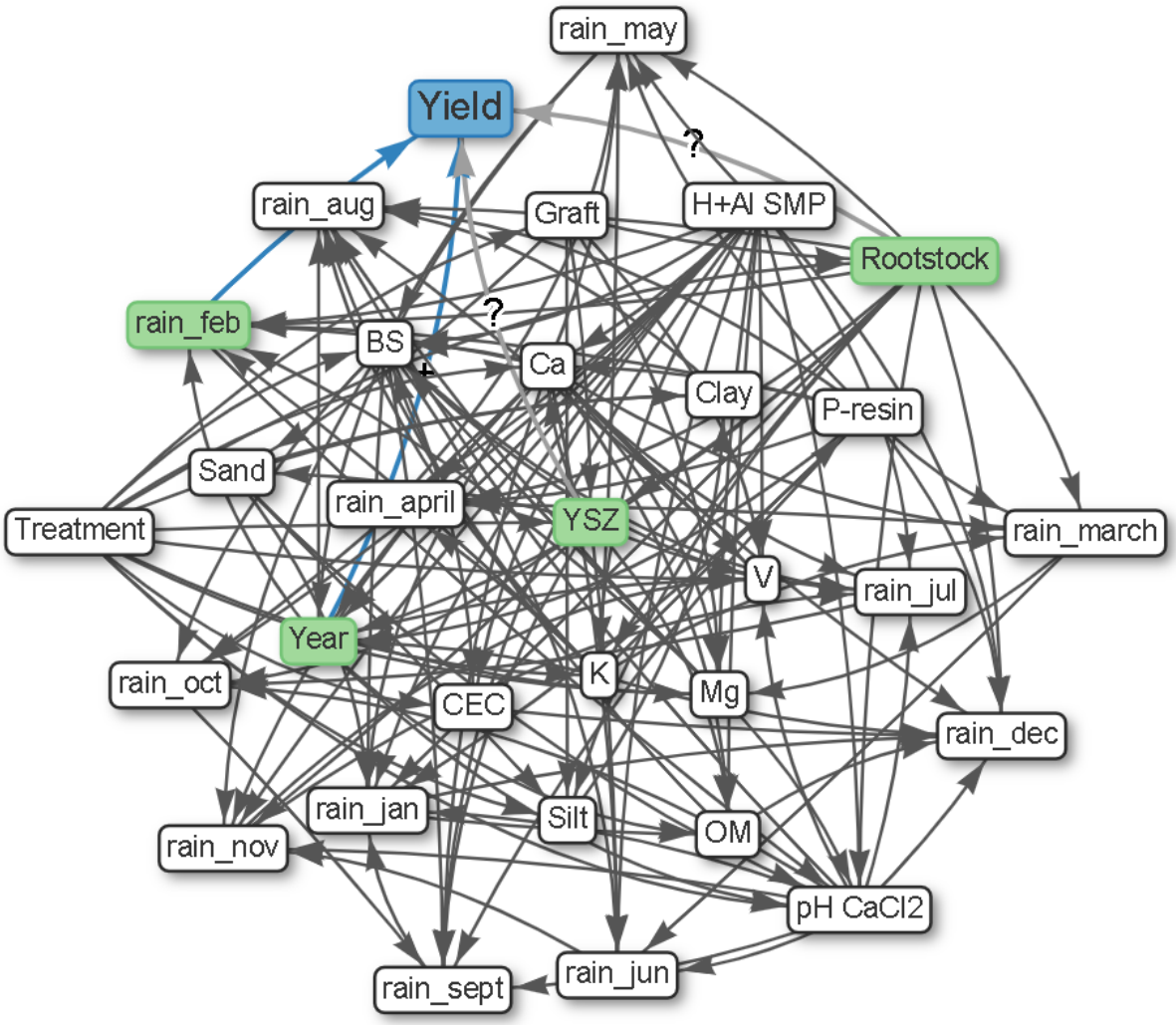

Fig 3. Causal network generated for the citrus database. Edge colors and signs indicate the nature of relationships where calculable: blue (+) = positive correlation, orange (−) = negative correlation, gray (?) = association involving categorical variables where direction cannot be unambiguously determined.

The alignment between causal findings and the convergent associative results is striking. Year, YSZ, and February rainfall emerged as influential in both the predictive (XGBoost and RF) and causal frameworks. Rootstock type, while not universally top-ranked in all ML importance lists, was directly linked to yield in the causal network, suggesting it plays a genuine mechanistic role that may be partially masked in purely associative rankings due to collinearity or interaction effects.

Conversely, treatment (FR vs. VR fertilization) appeared in some SHAP importance rankings but had no direct causal edge to yield in the Bayesian network. This raises the possibility that treatment effects observed in the associative models may be partially mediated through other variables (e.g., soil properties or YSZ), rather than representing a direct, unmediated influence. Such a distinction would be difficult to make using only predictive models and SHAP values.

Synthesis and Implications for Decision-Making

The integrated framework demonstrates that triangulating evidence across predictive and causal methods helps distinguish stable, plausibly causal drivers from spurious or model-dependent correlations. In this study, YSZ, year, and February rainfall consistently withstood cross-methodological scrutiny, whereas treatment effects and the relative importance of rootstock versus graft type varied considerably.

From a practical decision-making perspective, this suggests that growers and agronomists using this dataset should prioritize management actions targeting yield stability zones (e.g., zone-specific practices) and consider February rainfall as a critical seasonal window, while interpreting rootstock and graft rankings with caution unless confirmed by additional domain-specific experiments. More generally, the framework proposed here offers a pathway to move from "which model is best" to "which variables are most trustworthy" by requiring convergence across multiple analytical paradigms.

Limitations

While the multi-method framework presented here demonstrates the value of triangulating evidence across associative and causal methods, some limitations should be acknowledged. First, the primary objective of this study was to illustrate how different analytical paradigms—XGBoost, Random Forest, SHAP, and causal discovery—can yield divergent yet complementary insights when applied to the same dataset. As such, model evaluation focused on standard hold-out validation (80/20 train-test split) rather than spatial or temporal cross-validation. However, agricultural datasets often exhibit both spatial autocorrelation (nearby points tend to be more similar) and temporal dependence (yields in one year may influence or correlate with yields in subsequent years). Standard random splitting may overestimate model performance by allowing the model to implicitly learn from spatially or temporally adjacent points that appear in both training and test sets. Future studies applying this framework should incorporate spatial block cross-validation (e.g., clustering by field or grove) and temporal forward-chaining cross-validation (training on earlier years to predict later years) to provide more realistic estimates of model generalizability to unseen locations and future seasons.

Second, while this study employed two widely used tree-based ensemble methods (XGBoost and Random Forest), the framework is not limited to these algorithms. The principle of convergence across methods would be strengthened by including additional algorithm families. Future research should extend this approach to include, e.g., (a) neural networks (e.g., multilayer perceptrons or deep learning architectures), which can capture highly complex non-linear interactions but require careful regularization to avoid overfitting, (b) generalized linear models (e.g., logistic regression or GLMs with regularization), which offer greater interpretability and can serve as a baseline for assessing the added value of non-linear methods, and (c) support vector machines or gaussian process regression, which may capture different types of underlying data structures. The inclusion of diverse algorithm families would allow for stronger claims about convergence: variables that consistently emerge as important across tree-based, neural, and linear methods would be even

more likely to represent stable relationships rather than artifacts of a particular modeling family.

Third, the causal discovery method employed provided valuable independent evidence, causal discovery from observational data even with domain constraints, it cannot definitively establish causation without intervention or experimental validation. The directed edges identified in Figure 3 should be interpreted as plausible causal pathways consistent with the data, not as proven causal relationships.

Conclusion

This study demonstrated that different analytical processes produce divergent variable importance rankings, confirming that reliance on any single model or metric—whether XGBoost GAIN, XGBoost SHAP, RF %IncMSE, or RF SHAP—can lead to incomplete or biased conclusions about crop yield drivers. However, by examining convergence across methods, consistent drivers emerged: yield stability zone, February rainfall, and year. The independently constructed causal network provided corroboration, identifying year, YSZ, rootstock type, and February rainfall as factors with direct directed edges to yield. The use of multiple methods thus offers a rigorous pathway to move beyond "which model is best" toward "which variables are most trustworthy". For decision-making, growers should prioritize management actions targeting yield stability zones, select rootstock varieties as a direct causal lever, and consider February rainfall as a critical water stress window.

Acknowledgments

This study was supported by multiple Brazilian agencies: the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Finance Code 001); the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Process 306177/2021–8, with a Research Productivity grant); and the São Paulo Research Foundation (FAPESP, Grant #2019/07665–4).

References

Abdullah, T. A. A., Zahid, M. S. M., & Ali, W. (2021). A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. *Symmetry*, *13*(12), 2439. https://doi.org/10.3390/sym13122439

Alsaber, A., Satpathi, A., Alsabah, M., & Setiya, P. (2025). Optimizing potato yield predictions in Uttar Pradesh, India: a comparative analysis of machine learning models. *Scientific Reports*, *15*(1), 26897. https://doi.org/10.1038/s41598-025-12719-8

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Colaço, A. F., & Molin, J. P. (2017). Variable rate fertilization in citrus: a long term study. *Precision Agriculture*, *18*(2), 169–191. https://doi.org/10.1007/s11119-016-9454-9

Colaço, A. F., Trevisan, R. G., Karp, F. H. S., & Molin, J. P. (2020). Yield mapping methods for manually harvested crops. *Computers and Electronics in Agriculture*, *177*(August). https://doi.org/10.1016/j.compag.2020.105693

Jahan, M., Nassiri-Mahallati, M., Soltani, M., & Yaghoubi, F. (2026). Machine learning-driven strategies for wheat yield prediction and climate adaptation in Arid regions: a comparative analysis of random forest and XGBoost. *Theoretical and Applied Climatology*, *157*(5), 315. https://doi.org/10.1007/s00704-026-06248-1

Mafrica, R., Gattuso, A., Mafrica, D., De Bruno, A., & Poiana, M. (2026). Influence of Rootstock on Growth, Yield, and Fruit Quality of the 'Femminello' Bergamot (Citrus bergamia Risso & Poit.). *Agriculture*, *16*(4), 405. https://doi.org/10.3390/agriculture16040405

Marcaida III, Manuel, Cho, Jason, Shajahan, Sunoj, Contessa, Mike, Ketterings, Q. (2022). Yield Stability Zones. http://nmsp.cals.cornell.edu/publications/factsheets/factsheet123.pdf

R Core Team. R: A Language and Environment for Statistical Computing. (2019). Rstudio. Vienna: R Foundation for Statistical Computing. http://www.rstudio.com/

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable

models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sánchez, J. C. M., Mesa, H. G. A., Espinosa, A. T., Castilla, S. R., & Lamont, F. G. (2025). Improving wheat yield prediction through variable selection using Support Vector Regression, Random Forest, and Extreme Gradient Boosting. *Smart Agricultural Technology*, *10*, 100791. https://doi.org/10.1016/j.atech.2025.100791

Torgbor, B. A., Rahman, M. M., Brinkhoff, J., Sinha, P., & Robson, A. (2023). Integrating Remote Sensing and Weather Variables for Mango Yield Prediction Using a Machine Learning Approach. *Remote Sensing*, *15*(12), 3075. https://doi.org/10.3390/rs15123075

Tsoumas, I., Sitokonstantinou, V., Giannarakis, G., Lampiri, E., Athanassiou, C., Camps-Valls, G., et al. (2025). Leveraging causality and explainability in digital agriculture. *Environmental Data Science*, *4*, e23. https://doi.org/10.1017/eds.2025.14

Van Raij, B., Cantarella, H., Guaggio, J. A., & Furlani, A. M. C. (1997). *Recomendações de adubação e calagem para o Estado de São Paulo* (Bulletin 100: Fertilizer and liming recommendations to the State of São Paulo) (2nd ed.). Campinas: IAC

Wei, M. C. F., Longchamps, L., Colaço, A. F., & Molin, J. P. (2026a). Integrating stability zones and machine learning for enhanced crop management. *Precision Agriculture*, *27*(2), 38. https://doi.org/10.1007/s11119-026-10318-9

Wei, M. C. F., Molin, J. P., & Longchamps, L. (2026b). Predictive power vs interpretability: Machine learning approaches to unravel sugarcane yield drivers. *Computers and Electronics in Agriculture*, *243*, 111353. https://doi.org/10.1016/j.compag.2025.111353

Zhu, L., Liu, X., Wang, Z., & Tian, L. (2023). High-precision sugarcane yield prediction by integrating 10-m Sentinel-1 VOD and Sentinel-2 GRVI indexes. *European Journal of Agronomy*, *149*, 126889. https://doi.org/10.1016/j.eja.2023.126889