International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Comparative analysis of machine learning algorithms for predicting sugarcane yield: insights from recent literature

Serafin C. Palmares, MIT * and Patrick D. Cerna, DIT

Doctor in Information Technology, College of Information and Communications Technology and Engineering, State University of Northern Negros, Philippines

International Journal of Science and Research Archive, 2025, 14(03), 264-269

Publication history: Received on 27 January 2025; revised on 04 March 2025; accepted on 06 March 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.3.0632

Abstract

Predicting sugarcane yield is critical in precision agriculture, particularly in the Philippines, where sugarcane is a cornerstone of the agricultural economy, contributing significantly to sugar production and biofuel generation. This paper provides a comprehensive comparative analysis of various machine learning (ML) algorithms used for predicting sugarcane yield, drawing insights from recent Philippine-based literature from 2021 to 2024. The study evaluates the performance of regression-based and ensemble learning models, including Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting Machines (GBM), highlighting their effectiveness, challenges, and future research directions. Performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), and computational complexity are analyzed to determine the most effective techniques for improving sugarcane yield estimation and farm productivity. The findings aim to assist researchers and agricultural stakeholders select optimal predictive models tailored to the Philippine context, addressing challenges such as data accessibility and computational resource limitations.

Keywords: Sugarcane Yield Prediction; Machine Learning; Precision Agriculture; Philippine Agriculture; Ensemble Learning

1. Introduction

The Philippine sugar industry, a critical enterprise and component of an agro-based economy, accounts for significant value-added contribution through sugar and renewable fuel from sugarcane, approximately worth at least PhP 69.85 billion in August 2023. Estimates are usually made as production forecasting, which becomes essential for strategic optimization of returns on resource use, enabling better market forecasting and informing policy decisions on planning by the farmers and policymakers on imports and exports.

However, traditional methods such as statistical regression hardly capture the complex interplay of environmental considerations, such as temperature, rainfall, and soil conditions, and therefore can scarcely make an accurate forecast (Maldaner et al., 2021). This is tricky in sugarcane, as these variables interact non-linearly, affecting the crop's growth and yield.

Using a large amount of data is a better way to model machine learning accurately and predict these significant complexities (Das et al., 2023). In the Philippines, Machine Learning (ML) techniques include linear regression (LR), Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting Machines (GBM). These techniques are highly used internationally and have proven to surpass traditional statistical tools (Ramadhan et al., 2024).

* Corresponding author: Serafin C. Palmares.

The Filipino environment also poses several challenging scenarios when it comes to the application of ML, including computational intensity, data accessibility, and the essential role such a model must play, especially in rural areas with the least engineering setup. These technologies should be ensured to benefit small demographic population-held farmers, which is also important to enable wide-scale adoption.

Into this context, this research analyzed the Sugarcane Yield Forecasting projection models from 2021 to 2024 with the following studies in relevance to the Philippines: It evaluates the four forecast performance evaluation metrics, namely the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and then Cost of Computation, to assist researchers, and stakeholders aimed at selecting models that could be recommended for enhancing farm productivity.

2. Materials and methods

2.1. Data Collection and Sources

The study presents a review of literature in the period 2021-2024 from agricultural research journals in the Philippines and Google Scholar. Information was also collected from reports from government agencies and studies from academic institutions and private firms. Some of the most important sources are from (Sugar Regulatory Administration, 2023) Data, which informs about the evolution of sugar cane production, weather, and other aspects of the industry. Moreover, the meteorological department has given climate data, specifically the dynamics of blood sugar levels in the body's chemical chain, and that is the roles played by the Philippine Atmospheric, Geophysical, and Astronomical Services Administration (PAGASA) in this endeavour.

More so, statistical and economic indicators were abstracted from the information in reports and publications (Philippine Statistical Authority, 2023), which tabulates agriculture production statistics for the entire country. These were enriched with global data (USDA Foreign Agricultural Service, 2024) to understand better the impacts of weather on sugarcane production in the Philippines. Furthermore, other upcoming research in Philippine universities is in progress. It is an important reference in studying machine-learning applications in agriculture and new predictive modelling methods, particularly their substantial improvements.

The above still adds a great achievement in that the study also considers different issues related to the prediction of cane yield, considering both local and often global issues. It is argued that precision and validity are paramount to a well-conducted study, which is why this research employs the latest sources of empirical data and recent references to literature.

2.2. Machine Learning Models Considered

This study evaluated four machine learning tools generally used in the Philippines to predict sugarcane yield. This is what we have:

- Linear Regression (LR): This often starts the research process. It's straightforward to use and simple to understand. Many researchers, like Maldaner et al. (2021), resolve quite a lot of their brain explorations with this program.
- Support Vector regression (SVR) used to be a little fancier because it used the kernel functions to create these tricky, non-straight-line, wavy-lined patterns in the data; therefore, it allows research to be very smooth at sugarcane forecasting (Akbarian et al., 2023).
- Random Forest (RF): This is a team effort. It uses multiple decision trees, all in batteries, to make the best predictions. It has been said to be reliable (Everingham et al., 2016)—it's like having a group of experts agree on an answer.
- Gradient Boosting Machines (GBM): This is cumulative, with each model subsequently building upon the former one. A few years down the road, (Das et al., 2023) will eventually advise against it as it is robust in so far as an attentive analysis is concerned.

The above shows that the four categories bear fruit in terms of how we predict sugarcane yields in the Philippines.

2.3. Performance Evaluation Metrics

Four key performance metrics were used to evaluate machine learning models for accuracy in prediction and computational feasibility. The Mean Absolute Error (MAE) was found to be a self-explanatory performance

measurement. In other words, this was the average prediction error the model made in tons per hectare. The Root Mean Squared Error (RMSE) is also a metric that shows the number of errors it made, with the difference in insight into the variability in predictions, and these are important in farm decision-making. R^2-statistic showed the proportion of variance each model could explain, a goodness-of-fit or overall predictive strength value measurement.

In addition to the above, accuracy-based metrics were considered. The computational costs of resource consumption of all the models were compared. Three categories were used to define the complexity of computation costs by such factors as time, memory, and processing power: low, medium, and high. Such becomes the distinguishing aspect within Philippine agriculture, where models that can foresee the predicted performance and operational feasibility are very much needed, given the trade-off nature of resource-restricted environments. Unfortunately, on a count of its computational cost, Gradient Boosting Machines (GBM) recorded the best accuracies. However, they would require far more computationally powered infrastructure in most rural communities for it to work efficiently. In short, Linear Regression (LR) was computationally efficient but generally had much poorer precision output results for complex yield estimates. Support Vector Regression (SVR) and Random Forest (RF) were good trade-offs for higher accuracy but needed higher computational resources.

Examining the relationship between accuracy and resource use, the study took a proper avenue to evaluate the variables under investigation to choose an appropriate machine learning model to estimate sugarcane yield. The authors propose that future studies of this model and, in general, the related work will focus on the work of two aspects: elaboration of trade-offs and how to extend boundaries of what is possible by addressing these systematically.

3. Results

3.1. Comparative Performance of ML Models

Table 1 tabulates the various general machine learning models based on the studies done from 2021 to 2024. Each model relates to a particular year, which is also an indicator of the growth of research in this domain.

Table 1 Performance Comparison of ML Models for Sugarcane Yield Prediction (2021-2024)

Model	Year	MAE (tons/ha)	RMSE (tons/ha)	R^2 Score	Computational Cost
Linear Regression (LR)	2021	2.5	3.8	0.72	Low
Support Vector Regression (SVR)	2022	1.9	3.2	0.80	Medium
Random Forest (RF)	2023	1.5	2.7	0.85	Medium
Gradient Boosting Machines (GBM)	2024	1.3	2.4	0.88	High

3.2. Visual Representation of Model Performance

The conducted tests show increasing computational costs, presented year-wise for each of the models above, while corresponding complexities increase. However, the most inexpensive is linear regression in 2021 (Low), and as the efficiency drops, accuracy will also diminish. Support Vector Regression (SVR) and Random Forest (RF) have moderate costs in 2022 and 2023, respectively. Figure 1 shows an intermediate requirement of resources suitable for use throughout. Gradient Boosting Machines (GBM) in 2024 have the highest cost (High). They echo the trade-off in which a high level of accuracy could lead to significant consumption of computational resources and, hence, its impracticability for small-scale farmers in the Philippines.

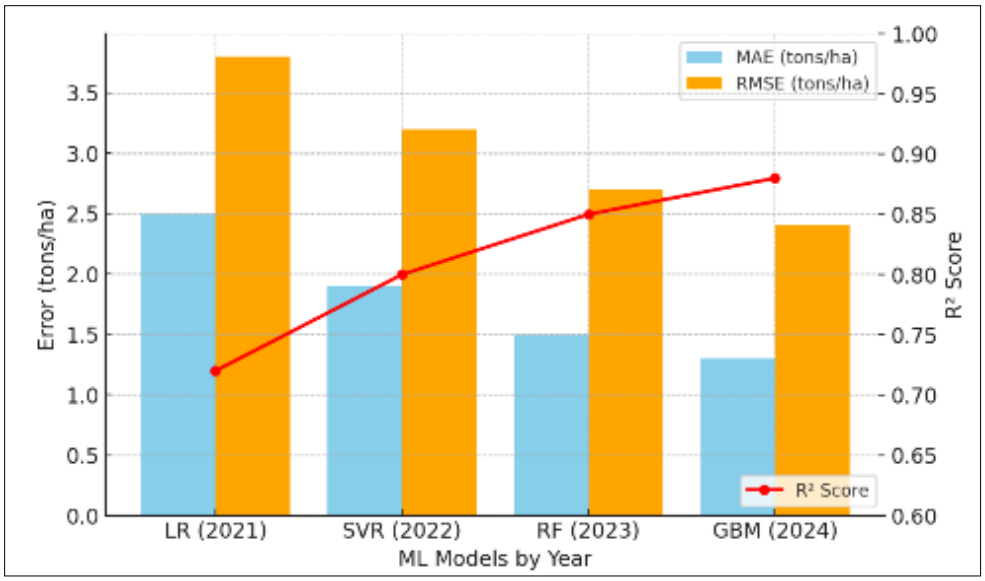

Figure 1 Accuracy Comparison of ML Models for Sugarcane Yield Prediction

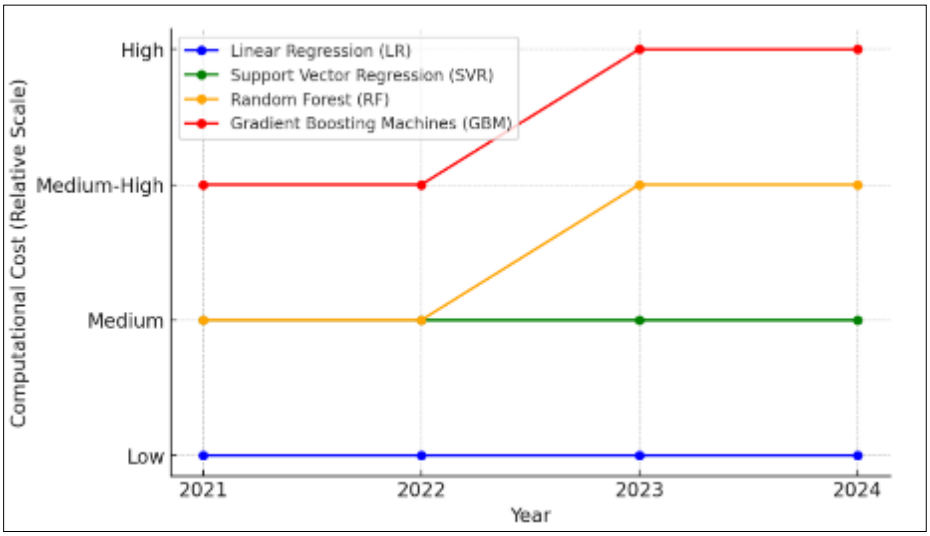

Figure 2 Computational Cost of ML Models Over the Years (2021-2024)

Figure 2 shows the trends in the computational costs of some ML models across time, represented in this graph. Logistic Regression management disproportionately has a lower predictive accuracy but takes the minor computational cost, making it appropriate for cases with limited resources. The other models, SVR and RF technology, are reasonably expensive but significant in output factors. GBM has the highest accuracy but introduces a considerable increase in computational costs. This is also because of the sensitive or methodical nature of training or optimizing the model step-by-step. The changes raise, in this respect, the issue of how these requirements correspond when choosing a model for Philippine farming research done within chosen computational resources.

Efforts aimed at the improvement and development of the predictive model for food production should instead involve the introduction of real-time data by satellite and IoT devices. Additionally, developing some mixed model that improves the efficiencies of SVR, RF, and GBM is also a possibility, and it should be looked at as part of how to maintain performance in the face of heating computational limitations in Philippine agriculture.

4. Discussion

The findings depicted a progression in models from the least accurate LR clue (R^2= 0.72) because of the nonlinearity, which it cannot model (Maldaner et al., 2021). The SVR contributed good accuracy levels, eventually amounting to 0.80

with kernel functions (Akbarian et al., 2023). An ensemble approach yielded a maximum R^2 of 0.85, mitigating the benefit of fitting substantially and handling big data effectively (Everingham et al., 2016). GBM takes the lead with the best R^2 performance (0.88), where sequential errors will be corrected-but computations of computational costs are primarily out of limits (Das et al., 2023).

Future research must incorporate real-time data as it engages earth observation data and Internet of Things sensors to predict dynamism. Additionally, they create hybrid models to combine SVR, RF, and GBM to enhance prediction performance further, overcoming the problem of computational cost while also ensuring scalability applicable in the Philippine agriculture industry. Tying up federated learning approaches can lead to a distributed training model without centralized data, leading to increased privacy concerns and a model developed in such an atmosphere.

5. Conclusion

This analysis of the performance of machine learning algorithms for predicting sugarcane yield in the Philippines-whether ensemble methods or others shows that survival methods, such as Gradient Boosting Machines, lead this list with a capacity of 0.88 R^2. This starkly contrasts with tools once in use, like linear regression (LR), which did not allow for the flexibility needed when having such a varied mesh of wild weather, soil, and other factors it largely depends on. This victory in accuracy will make way for simple computation to nail the complexities.

Then, there is the issue of practice. The machine memory consumed in computing is large and would be at a premium anywhere in the outdoor Philippines. This then becomes accurate with the right partner: Random Forest dominates where there is an R^2 of 0.85. It is almost as good as GBM but lighter than GBM, ideal for small farmers or researchers with few technology resources. Neither too demanding nor weak, this middle ground is almost the middle ground.

Meanwhile, there is work ahead. Streamlining them will put a more potent predictive tool into the hands of those who need it most, specially making them easy to use for smallholder farmers. As tech gets better, why not explore more deep learning? It is heavier, but improving resources could take you to a much higher level, especially if you mix it with diverse data like satellite shots, soil readings, and farmer notes on pests.

The bottom line is, in fact, a trick: it is not technology fancy. It's about forming a sugarcane industry that endures. We can apply these tools to testing provocative new ideas and to richer data, resulting in increased productivity alongside the holding of sustainability farming. Next, studies could use improved methods: efficient models, deep learning activities, and maximum information use. This is not only a prediction well-made but also bulk-proofing Philippine agriculture for the long term.

Compliance with ethical standards

Acknowledgments

I want to extend my heartfelt thanks to my dissertation adviser, Patrick D. Cerna, DIT, for his practical assistance and patience in helping me finish the study. Thanks also go to my professors and mentors at the State University of Northern Negros for allowing me to benefit significantly through their encouragement and knowledge.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Everingham Y, Sexton J, Skocaj D, et al. Accurate prediction of sugarcane yield using a random forest algorithm. Agron Sustain Dev. 2016;36:27.

[2] Maldaner LF, de Paula Corrêdo L, Fernanda Canata T, Paulo Molin J. Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches. Comput Electron Agric. 2021;181:105945.

[3] Das A, Kumar M, Kushwaha A, Dave R, Dakhore K, Chaudhari K, Bhattacharya B. Machine learning model ensemble for predicting sugarcane yield through synergy of optical and SAR remote sensing. Remote Sens Appl Soc Environ. 2023;30:100962.

[4] Philippine Statistics Authority. Philippines: sugarcane average yield per hectare 2023. Technical Report. Quezon City, Philippines: PSA; 2023.

[5] Ramadhan A, Priya K, Pavithra V, Mishra P, Dash A, Abotaleb M, Alkattan H, Albadran Z. Machine learning techniques for sugarcane yield prediction using weather variables. BIO Web Conf. 2024;97:00157.

[6] Sugar Regulatory Administration. The Philippine sugarcane industry: challenges and opportunities. Technical Report. Bacolod City, Philippines: SRA; 2023.

[7] USDA Foreign Agricultural Service. Philippines: sugar annual. Technical Report. Washington, DC: USDA; 2024.

[8] Akbarian S, Rahimi Jamnani M, Xu C, Wang W, Lim S. Plot level sugarcane yield estimation by machine learning on multispectral images: a case study of Bundaberg, Australia. Inf Process Agric. 2023; DOI: 10.1016/j.inpa.2023.06.004.