Technologique: A Global Journal on Technological Developments and Scientific Innovations
2025, Vol. 4, No.1, 51 – 56 | ISSN Online: 3028-1415 | ISSN Print: 3028-1407
DOI: https://doi.org/10.62718/vmca.tech-gjtdsi.4.1.SC-0425-009

Predicting Irrigators' Association Membership in Bohol Using Decision Trees and Random Forests

Article History:
Received: 23 April 2025
Accepted: 14 May 2025
Published: 21 May 2025

Max Angelo D. Perin[1], ORCID No. 0000-0002-2746-7220
Larmie S. Feliscuzo[1], ORCID No. 0000-0001-8155-3843
Chris Jordan G. Aliac[1], ORCID No. 0000-0002-3501-4539
Nelia Q. Catayas[2], ORCID No. 0000-0002-1235-2746

[1]College of Computer Studies, Cebu Institute of Technology-University, Cebu City, Philippines
[2]College of Technology, Bohol Island State University-Bilar Campus, Bohol, Philippines

Abstract

Irrigators' Associations (IAs) are critical in managing water resources and influencing agricultural productivity and rural sustainability. Predicting IA membership is complex due to socio-economic, organizational, and environmental factors. This study applies machine learning techniques—Decision Trees and Random Forests—to model IA membership and identify the key predictive variables. Using a dataset of 234 IA records, the models were evaluated based on Mean Absolute Error (MAE), with the Random Forest model achieving a lower MAE of 19.49, compared to 24.92 for the Decision Tree. Key predictors include the number of farm beneficiaries, service area, years of operation, and leadership structure. The results demonstrate the potential of machine learning in supporting data-driven planning for IA engagement, offering valuable insights that can enhance resource allocation and inform policy development in the agricultural sector, ultimately contributing to more efficient water management and improved rural livelihoods.

Keywords: Irrigators' Associations, Machine Learning, Decision Trees, Random Forests, Mean Absolute Error

INTRODUCTION

Effective irrigation management plays a pivotal role in the sustainability of agricultural communities, particularly in developing nations like the Philippines, where a significant proportion of the population relies on farming for livelihood. Central to this management are Irrigators' Associations (IAs), which function as grassroots organizations that facilitate equitable water distribution, infrastructure maintenance, and coordination among farmers (Gragasin et al., 2005). Predicting IA membership can offer valuable insights into farmer participation, engagement levels, and potential gaps in service delivery, ultimately contributing to improved agricultural productivity and policy planning.

The advent of machine learning (ML) has opened new frontiers in agricultural research by enabling data-driven decision-making. Decision Trees (DT) and Random Forests (RF) are among the most effective algorithms. These methods are known for their interpretability, robustness to noise, and ability to handle both categorical and continuous variables (Kotsiantis, 2011; Rokach, 2016). Their versatility has led to successful applications across diverse domains, including education (Beaulac & Rosenthal, 2018), healthcare (Shaikhina et al., 2017), and environmental studies.

In this study, we proposed a supervised learning approach using Decision Trees and Random Forests to predict IA membership based on various demographic and socio-economic features collected from farmers. The objective is to evaluate and compare the predictive performance of these algorithms and identify key variables associated with IA participation. This approach enhances the interpretability of results for local government units and agricultural policymakers and provides a

foundation for developing digital tools supporting precision agriculture.

Machine learning models, especially Decision Trees and Random Forests, have gained popularity recently due to their adaptability and predictive power. A Decision Tree is a non-parametric supervised learning method for classification and regression. It builds a tree-like model of decisions by recursively splitting data based on input variables. These models are valued for their simplicity, fast computation, and ease of interpretation (Kotsiantis, 2011).

Random Forests, an ensemble learning method, overcomes the overfitting tendency of single decision trees by constructing multiple trees and aggregating their predictions. This improves accuracy and generalization across datasets (Rokach, 2016). Studies have shown that Random Forests perform well even with noisy and incomplete datasets, making them ideal for real-world agricultural data, which often suffers from inconsistencies (Tsang et al., 2011). Schonlau and Zou (2020) offer a statistical perspective on using Random Forests, highlighting their utility in exploratory data analysis and predictive modeling. Salman et al. (2024) review Random Forests in the context of agriculture, emphasizing their effectiveness in crop yield prediction and resource allocation. Similarly, Beaulac and Rosenthal (2018) applied Random Forests to predict students' academic success and major selection, illustrating the model's applicability in social science domains.

In the healthcare sector, Shaikhina et al. (2017) demonstrated the superiority of Random Forests and Decision Trees over traditional statistical methods in predicting medical outcomes, further validating their performance in high-stakes prediction tasks. Klusowski and Tian (2021) expand on this by examining the scalability of decision trees for large-scale applications, a critical factor when dealing with nationwide or regional agricultural datasets such as those used in irrigation management.

Locally, Gragasin et al. (2005) emphasized the importance of IAs in enhancing farm productivity in the Philippines. Their comparative study of two irrigation systems revealed significant differences in yield and farmer satisfaction, attributed to the presence or absence of organized water user groups. The current study builds on this foundational work by introducing a machine learning dimension to predict IA membership, offering a proactive rather than reactive approach to agricultural governance.

METHODOLOGY

This study employs a quantitative research design, utilizing machine learning techniques—decision trees and random forests—to analyze and predict IA membership. The dataset used for this study was sourced from an IA profiling database, containing various features such as the number of farm beneficiaries, service area, years since organization, and other socio-economic indicators. A total of 234 IA records were used for analysis. The data preprocessing, model training, and evaluation were conducted using Python in Google Colab.

Data preprocessing involved handling missing values through the K-Nearest Neighbors (KNN) imputer, encoding categorical variables using label encoding, and feature engineering to enhance predictive accuracy. The KNN imputation method was chosen as it fills in missing values based on the average of neighboring data points, ensuring that the missing values are estimated based on similar entries, which helps maintain the integrity of the dataset. Label encoding was applied to categorical variables, converting them into numerical values to enable the machine learning models to process them effectively. Feature engineering was conducted by selecting variables relevant to irrigation and agricultural organizations, such as the number of farm beneficiaries, service area, and leadership structure. The target variable for prediction was the number of IA members, with features selected based on their significance in agricultural organizations.

The dataset was split into training and testing sets using an 80-20 ratio, ensuring a robust model evaluation. Decision trees were

optimized through hyperparameter tuning, using grid search to identify the best combination of tree depth, minimum samples per split, and minimum samples per leaf. The random forest model was trained with 100 estimators and a maximum depth of 5 to balance performance and interpretability.

The models' performance was evaluated using Mean Absolute Error (MAE), as it provides a transparent and interpretable measure of prediction accuracy. MAE was selected over metrics such as Mean Squared Error (MSE) because it directly reflects the average error magnitude and is less sensitive to outliers. This makes it more appropriate for this study, where a straightforward understanding of prediction accuracy is crucial. By focusing on MAE, the model's errors can be more easily understood and compared, which is essential when interpreting the practical implications for IA membership predictions.

Visualization techniques, such as decision tree plotting, enhance model interpretability. These visualizations helped understand the decision-making process of the tree models, providing insights into the most significant features influencing IA membership prediction. The results were analyzed to determine the key predictors of IA membership and to assess the potential of machine learning in supporting agricultural policy and planning.

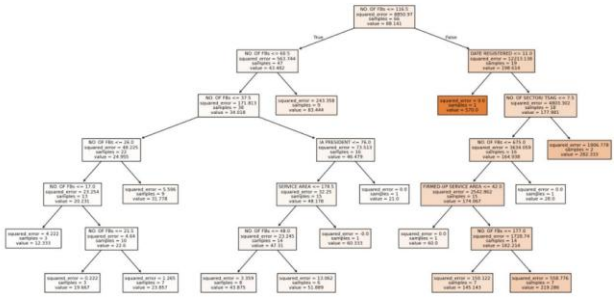

Figure 1
Optimized Decision Tree for IA Membership Prediction

RESULTS AND DISCUSSION

The predictive modeling for the Irrigators' Association (IA) membership using Decision Trees and Random Forests yielded valuable insights into the factors influencing membership numbers. Two models were employed: a Decision Tree Regressor and a Random Forest Regressor. The results indicate that the Random Forest model achieved a lower Mean Absolute Error (MAE) than the Decision Tree, suggesting superior predictive performance.

Table 1
Model Performance

Model	MAE	RMSE	R^2
Optimized Decision Tree	24.92	54.20	0.34
Random Forest	19.49	50.83	0.42

The Decision Tree model produced an MAE of 24.92, while the Random Forest model achieved a lower MAE of 19.49. The lower error in the Random Forest model suggests that it provides more accurate membership predictions than the Decision Tree. This aligns with prior studies highlighting Random Forests' ability to mitigate overfitting and capture complex patterns more effectively than a single Decision Tree (Sagi & Rokach, 2020; Schonlau & Zou, 2020). The improvement in prediction accuracy can be attributed to the ensemble nature of Random Forests that combine multiple decision trees to generate a more robust and generalized model.

Figure 1 presents the structure of the optimized Decision Tree model, which provides insights into the most significant variables affecting IA membership. The tree begins with the Number of Farm Beneficiaries (NO. OF FBs) as the primary splitting criterion. This suggests that the number of farm beneficiaries is the most influential factor in determining IA membership levels. Subsequent splits involve variables such as Date Registered, Number of Sectors/TSAG, Service Area, and IA President's Age, indicating their relevance in predicting membership counts.

Key Patterns in the Decision Tree Model:

- The first node (root node) splits on NO. OF FBs (Number of Farm Beneficiaries) $\leq$ 116.5, highlighting this variable's critical role in

membership prediction. Higher numbers of farm beneficiaries tend to correlate with larger membership sizes.

- The second major split at NO. OF FBs (Number of Farm Beneficiaries) $\leq$ 60.5 further refines the segmentation, showing that IA membership patterns change significantly at this threshold.

- Date Registration is another critical factor, suggesting that older associations tend to have more stable or significant memberships.

- The Number of Sectors (TSAG) and Service Area size also influence membership, demonstrating that larger service areas and more structured organizations tend to attract higher memberships.

- The IA President's Age also plays a role, which may indicate that leadership experience contributes to effective association management and growth.

While the Decision Tree model provides interpretability, it is prone to overfitting, which may not generalize well to unseen data. The Random Forest model overcomes this limitation by aggregating multiple trees, leading to a lower MAE 19.49. The ensemble approach reduces variance and improves overall prediction accuracy, making it a more reliable choice for IA membership forecasting.

These results reinforce previous findings that Random Forests outperform individual Decision Trees in prediction tasks due to their ensemble structure (Rokach, 2016). The reduction in MAE from 24.92 to 19.49 highlights the advantage of using ensemble methods for predictive modeling in the agricultural domain.

Figure 2 shows the feature importance for the Random Forest model, emphasizing the Number of Farm Beneficiaries, Number of Sectors, and Date Registered as the most influential features for predicting IA membership.

The findings suggest that farm beneficiary numbers, organization age, service area, and leadership characteristics significantly impact IA membership. Future work can explore additional machine learning techniques, such as gradient boosting or deep learning models, to refine predictions. Additionally, incorporating external socio-economic factors and irrigation infrastructure data could enhance model performance and applicability.

This study provides a data-driven approach to understanding IA membership dynamics, supporting policymakers and stakeholders in making informed decisions to strengthen agricultural organizations.

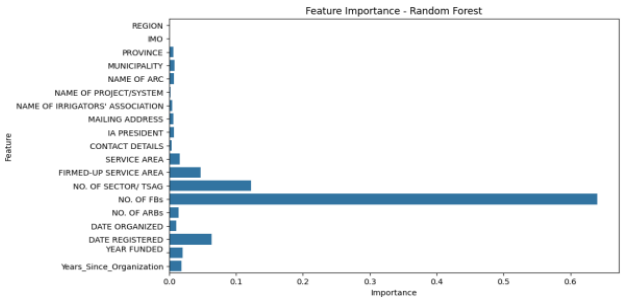

Figure 2
Feature Importance of Random Forest

CONCLUSION AND RECOMMENDATION

This study demonstrated the utility of machine learning, particularly Decision Tree and Random Forest algorithms, in predicting the membership size of Irrigators' Associations (IAs) using organizational profile data. Among the two models evaluated, the Random Forest Regressor outperformed the Decision Tree, achieving a lower MAE of 19.49, RMSE of 50.83, and an R^2 of 0.42, compared to the Decision Tree's MAE of 24.92, RMSE of 54.20, and R^2 of 0.34. These results indicate that the Random Forest model is better at generalizing patterns in the data and capturing complex relationships between features.

Feature importance analysis identified key predictors of IA membership: the number of farm beneficiaries, the number of sectors (TSAG), registration date, and the firmed-up service area.

These findings reflect the practical realities of IA operations—larger farm beneficiary bases and organized internal structures likely facilitate stronger membership retention and growth. The model's outputs offer actionable insights for planning interventions, prioritizing support for IAs with growth potential, and allocating resources based on structural and historical characteristics.

However, limitations remain. The current dataset is cross-sectional and limited in size and scope, focusing only on one geographic region. It does not include broader socio-economic, climatic, or policy factors critical in shaping IA dynamics. Furthermore, while Random Forests provide strong performance, their complexity can hinder interpretability, which is essential for policy-level applications.

Recommendations for future work include:

1. Expanding the dataset to include other provinces or regions for improved generalizability.

2. Integrating time-series data to uncover long-term trends and seasonal effects on IA membership.

3. Exploring advanced models such as Gradient Boosting Machines or Neural Networks to enhance predictive accuracy.

4. Socio-economic and environmental variables, such as subsidy access, rainfall patterns, or irrigation infrastructure, must be incorporated to improve model richness.

5. Enhancing model explainability through SHAP values or interpretable ML frameworks, ensuring stakeholders can trust and act on the findings.

By embracing data-driven modeling, agencies and local governments can better support IA development, refine outreach strategies, and promote sustainable irrigation practices tailored to the needs of rural farming communities.

Acknowledgments. The authors extend their thanks to the Research Office and the College of Technology Office, Bohol Island State University-Bilar Campus, and Cebu Institute of Technology-University for allowing and funding this study.

REFERENCES

Beaulac, C., & Rosenthal, J. (2018). Predicting university students' academic success and major using Random Forests. *Research in Higher Education, 60*(7), 1048–1064. https://doi.org/10.1007/s11162-019-09546-y

Gragasin, M., Datu, B. M., & Paras, C. (2005). Irrigators' Association and farm productivity: A comparative study of two Philippine irrigation systems. *Japanese Journal of Rural Economics, 7*, 1–17.

Klusowski, J., & Tian, P. (2021). Large-scale prediction with decision trees. *Journal of the American Statistical Association, 119*(525), 1–13. https://doi.org/10.1080/01621459.2022.2126782

Kotsiantis, S. (2011). Decision trees: A recent overview. *Artificial Intelligence Review, 39*, 261–283. https://doi.org/10.1007/s10462-011-9272-4

Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion, 27*, 111–125. https://doi.org/10.1016/j.inffus.2015.06.005

Salman, H., Kalakech, A., & Steiti, A. (2024). Random Forest algorithm overview. Babylonian *Journal of Machine Learning, 2*(1), 45–53. https://doi.org/10.58496/bjml/2024/007

Schonlau, M., & Zou, R. (2020). The random forest algorithm for statistical learning. *The Stata Journal, 20*(1), 29–53. https://doi.org/10.1177/1536867X20909688

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2017). Decision tree and random forest models for outcome prediction in antibody-incompatible kidney transplantation. *Biomedical Signal Processing and Control, 52*, 456–462. https://doi.org/10.1016/j.bspc.2017.01.012

Tsang, S., Kao, B., Yip, K., Ho, W., & Lee, S. (2011). Decision trees for uncertain data. *IEEE Transactions on Knowledge and Data Engineering, 23*(1), 64–78. https://doi.org/10.1109/TKDE.2009.175