

Explainable artificial intelligence for retrospective analysis of scheduled replanting intervals in Eucalyptus plantations

Nurhayati Sembiring^{1*}, Muhammad Rayhan Azbi¹

¹ Department of Industrial Engineering, Faculty of Engineering, Universitas Sumatera Utara, Jl. Almamater, Medan, Indonesia

* Corresponding author's e-mail: nurhayatipandia68@usu.ac.id

ABSTRACT

Replanting schedules play a crucial role in maintaining plantation productivity and supporting sustainable forest management. However, evidence on how well past replanting schedules can be explained and predicted using plantation management data and soil characteristics remains limited. This study aims to examine factors associated with scheduled replanting intervals in industrial “Eucalyptus” plantations and to assess the reliability of data-driven predictions. A retrospective analysis was conducted using historical plantation management records to investigate patterns in replanting scheduling and their relationship with stand characteristics and soil conditions. This study serves as a retrospective audit of past scheduling decisions, rather than an assessment of biologically verified readiness or the actual outcomes of the replanting efforts. Machine learning methods were applied to analyse these patterns, and predictive performance was evaluated across various harvest years and compared against simpler predictive approaches. The results indicate that plantation age and specific soil characteristics significantly influence the model's predictions. However, more complex models did not outperform simpler reference approaches; in the 2025 evaluation, they yielded an average prediction error of 27.81 days, compared to 16.15 days. These findings suggest that identifying influential variables does not automatically guarantee predictive reliability for operational planning purposes. This study transparently evaluates past replanting decisions and underscores the importance of assessing predictive reliability before implementing artificial intelligence in plantation management.

Keywords: operational scheduling; plantation management; temporal holdout; model benchmarking; feature attribution.

INTRODUCTION

Industrial scale Eucalyptus plantations require sustained wood productivity across successive planting rotations. A study on plantation forestry in Sumatra has documented the management of long-term productivity during the transition from “Acacia” to “Eucalyptus” (Hardiyanto et al., 2024). Research on successive “Eucalyptus” plantings has also examined changes in tree growth and soil quality (Dai et al., 2023). While these ecological studies highlight the importance of considering stand and soil variables, they do not validate companies' replanting intervals as benchmarks for biological readiness.

Nutrient availability and soil conditions remain critical considerations for sustainable “Eucalyptus” rotations. A recent chronosequence study of “Eucalyptus pellita” measured soil nutrient reserves and stoichiometric ratios (Powae and Rao, 2024), while research on successive plantings investigated shifts in soil quality (Dai et al., 2023). These findings provide an ecological context for recorded soil pH data and nutrient predictors; however, they do not establish a causal link to operational scheduling decisions.

Operational replanting schedules may reflect logistical and managerial constraints, as well as on-the-ground conditions. Consequently, retrospective models trained on historical scheduling data must be distinguished from site-suitability

models that have been biologically verified. Research on machine learning in environmental contexts also emphasises that spatial dependency requires careful attention during data splitting and validation (Jemeljanova et al., 2024). This analysis evaluates a “temporal holdout” method based on harvest years, accompanied by a sensitivity analysis grouped by compartment; this compartment-based grouping approach differs from geographic spatial blocking. Previous EEET studies have mapped research trends in “Eucalyptus” replanting (Sembiring and Napitupulu, 2025) and optimized scheduling by considering land use and labor aspects (Sembiring and Ishak, 2025).

Explainable artificial intelligence (XAI) offers methods for interpreting fitted nonlinear predictive functions. Recent reviews discuss SHAP as a model-explanation tool, while cautioning that such explanations should not replace robust validation (Salih et al., 2025; Pai et al., 2025). SHAP attributes characterise a fitted predictive model; they do not establish causal ecological mechanisms. Therefore, we interpret SHAP results by considering out-of-time predictive performance and data-quality limitations.

This study distinguishes operational scheduling from ecological readiness. Company records define post-harvest scheduling intervals as the difference between the scheduled replanting date and the harvest date. This analysis is retrospective: it evaluates how well a Random Forest model reconstructs historical scheduling intervals and how SHAP characterises the relationships within that model rather than predicting a validated biological readiness endpoint.

Random Forest serves as the foundation for an ensemble-based nonlinear predictive framework. We assess its interpretability and predictive generalisation using rolling-origin testing (by harvest year), median benchmarks from the training set, and compartment-based sensitivity analysis. SHAP explanations both global and local are calculated using the same model (with fixed parameters). This study provides a transparent, methodologically documented assessment of negative predictive findings for operational scheduling, rather than making claims about readiness for practical implementation. In related research on tropical land management, Arif et al. (2025) found that the performance of “Random Forest” differs significantly between land suitability classification and rice productivity prediction; this finding underscores the importance

of task-specific evaluation, rather than assuming that nonlinear models possess an equally robust generalization capability across various distinct prediction targets.

Although previous studies have examined research trends and the optimization of “Eucalyptus” replanting schedules (Sembiring and Napitupulu, 2025; Sembiring and Ishak, 2025), as well as the application of explainable AI in agricultural and environmental modeling (Pai et al., 2025; Arif et al., 2025; Marwiji et al., 2026), these studies have yet to determine whether interpretable models can reliably reconstruct the “Eucalyptus” replanting intervals established by a company for various future harvest-year cohorts. Specifically, the relationship between SHAP-based explanations and the reliability of out-of-time predictions for this operational target remains insufficiently explored. Another conceptual gap concerns the discrepancy between recorded managerial schedules and independently verified ecological readiness. This gap requires evaluation against simple baselines time-based validation and compartment-based grouping rather than relying solely on model interpretability or goodness-of-fit as evidence of operational suitability. Recent research on machine learning for spatial environments further underscores the need to evaluate models using validation methods that account for data dependencies, rather than relying exclusively on random data splitting (Jemeljanova et al., 2024).

This study’s contribution is methodological and empirical, not algorithmic: it integrates a harvest-year holdout method, a rolling-origin approach, compartment-based sensitivity analysis, simple model benchmarks, and “TreeSHAP” explanations (both global and local) into a single retrospective audit of scheduled replanting intervals. Its primary contribution is examining the gap between explainability and reliability: a fitted model may produce interpretable feature attributions while failing to improve predictive accuracy over a benchmark based on the training data median values. This approach does not claim that combining these methods is novel, that the model can identify the causal effects of soil, or that the historical schedules reflect biologically optimal replanting times.

The specific objectives are to: (1) reconstruct and audit the company-established scheduled replanting interval; (2) evaluating a conservatively designed “Random Forest” model by comparing it against a median-value benchmark, using

a time-based held-out cohort (segregated by harvest year); (3) quantifying global feature contributions via SHAP and examining local SHAP explanations for specific compartments excluded from the model training process (“held-out compartments”); and (4) discussing the ecological context, predictive limitations, and requirements for the development of future, field-verified decision-support systems. This study does not determine an optimal replanting timeframe based on the model’s predictions.

Previous research has explored explainable AI in agriculture (Pai et al., 2025). However, this study intentionally limits its scope: it audits the company’s scheduling targets, compares out-of-time predictions against a simple baseline, and explains the resulting model rather than claiming any independently validated ecological readiness. An environmental modelling study by Marwiji et al. (2026) also used a combination of “Random Forest” and SHAP to interpret flood hazards; unlike that classification-based application, this study evaluates regression models for retrospectively recorded plantation scheduling intervals and reports predictive failures relative to a simple baseline.

MATERIALS AND METHODS

This section describes the retrospective operational dataset, the construction of scheduled replanting intervals as the response variable, and the seven predictors used in the analysis. It also details data quality checks, pre-processing, model specification, temporal holdout evaluation, sensitivity analysis grouped by compartment identifier, and SHAP-based interpretation. The analysis dataset comprises 1,181 records associated with 923 distinct compartment identifiers. These procedures are designed to audit the company’s previously established schedules, not to determine biological readiness or the optimal timing for replanting.

Data origin, observation unit and target

This study utilizes original operational worksheets provided by a commercial “Eucalyptus” plantation company in North Sumatra, Indonesia. The raw dataset was kept intact without modification, and all analyses were performed using an in-memory copy. The dataset comprises 1,181

operational records associated with 923 distinct compartment identifiers. Each record is treated as an individual observation; however, recurring identifiers may represent non-independent observations rather than geographically separate plantation units. Records are distributed by harvest year as follows: 2020 ($n = 1$), 2021 ($n = 315$), 2022 ($n = 198$), 2023 ($n = 282$), 2024 ($n = 185$), and 2025 ($n = 200$).

The data provider has not granted permission to disclose information that could identify the company, the precise location of the plantations, geographic coordinates, compartment boundaries, the total area under management, the specific species or clones used, or proprietary harvesting and replanting protocols. Consequently, these details are omitted, limiting independent assessment of site representativeness, silvicultural comparability, and geographic transferability. Thus, the compartment identifiers are treated solely as operational labels rather than as independently verified geographic units. Disclosing this restricted information would require explicit authorisation from the data owner.

The response variable the scheduled replanting interval (in days) is calculated as the number of calendar days between the “Scheduled Replanting Date” and the “Harvest Date”, as recorded by the company for the 1,181 observations, with values ranging from 45 to 658 days. This variable reflects the interval documented in the company’s operational schedule, not the actual time elapsed until replanting or an independently verified ecological readiness threshold.

Predictors and data-quality screening

Seven numerical predictors were included in a fixed order: crop age, yield (tonnes), soil moisture, nitrogen (mg/kg), phosphorus (mg/kg), potassium (mg/kg), and pH. We excluded target-derived variables such as dates, scheduled dates, and compartment identifiers from the predictor set to prevent target leakage and the inappropriate use of operational identifiers.

Data quality checks revealed identical values between rainfall and air temperature, and between soil temperature and air humidity, across all 1,181 records. The “Area (ha)” variable contained ambiguous entries and was therefore excluded from the analysis. Seven yield values were either unavailable or non-numeric. No source records were

deleted; missing values in the retained predictors were imputed using the median value, calculated exclusively from the corresponding training partition to prevent information leakage. Units of measurement, laboratory procedures, and the provenance of field sensors could not be independently verified.

Data dictionary and audit trail

Seven model predictors are listed above, with the target defined as the number of days between two recorded dates. A non-sensitive analytical data dictionary has been compiled to document the model variables, their analytical roles, and any known limitations regarding data quality. This documentation excludes individual observations and sensitive operational information. However, a fully verified source-level dictionary covering all original columns, permitted values, measurement dates, missing-value conventions, and measurement provenance is not yet available. Consequently, while the existing documentation supports methodological transparency, it does not allow for a complete independent verification against the original operational records. Recent research on machine learning reproducibility emphasises that documenting data provenance, computational procedures, and experimental conditions is essential for assessing which components of an analysis can be independently verified (Semmelrock et al., 2025).

Model specification and validation

The Random Forest Regressor was configured with `n_estimators=250`, `min_samples_leaf=3`, `max_features=0.8`, and `random_state=42`, along with other default library settings. We fitted a `SimpleImputer(strategy="median")` only to each training partition. For each test harvest year (2023, 2024, and 2025), we used all observations from preceding harvest years as the training cohort. The benchmarks included the median training target value, an age-only Ridge model (`alpha=1`), a Ridge model with seven features (`alpha=1`), and a `DecisionTreeRegressor` (`max_depth=5`, `min_samples_leaf=10`, `random_state=42`). The Ridge predictors were not standardised; thus, these models served as fixed descriptive benchmarks rather than parameter-optimised competitors. MAE, RMSE, and R^2 values were calculated for each test cohort (held-out cohort).

A separate sensitivity evaluation was conducted using “`GroupKFold(n_splits=5)`” with compartment identifiers as the grouping variable, ensuring that no compartment identifier appears in both the training and testing sets for any given fold. Because the fold partitioning mixes different harvest years, this approach constitutes neither spatial block validation nor prospective temporal validation. For the year 2025, paired bootstrap resampling performed on 200 test observations (with replacement) over 5,000 iterations, using a NumPy random seed of 20261009 was used to generate 95% percentile intervals for the Mean Absolute Error (MAE) of the Random Forest (RF) model, as well as for the RF values minus the median absolute error. Resampling at the observation level does not account for dependencies within compartments that are observed repeatedly.

While “`GroupKFold`” prevents the same compartment identifier from appearing in both partitions of a fold, it does not guarantee geographic separation between adjacent compartments. Without spatial coordinates or polygons, we cannot calculate or verify metrics such as buffer distances, spatial autocorrelation, or spatial block-based fold assignments. Consequently, the results from this grouped-fold method serve only as a sensitivity check against “identifier leakage,” and this study makes no claims regarding spatially independent generalisation. Future spatial validation efforts will require georeferenced compartment boundaries or centroids, predefined spatial blocks and buffer distances, and reporting on fold assignments and errors at the block level. Research on spatial cross-validation further indicates that geographical separation, as well as the disparity between training locations and prediction locations, can significantly influence performance assessment (Wang et al., 2023).

SHAP and case-selection procedure

The 2025 Random Forest model was trained on 981 data points collected before 2025. We applied `TreeExplainer` to the trained, frozen forest model across all 200 test observations after imputing median values from the training data. We determined global importance by calculating the mean absolute SHAP attribution within the test set, expressed in units of predicted days. The additivity property of SHAP was numerically verified. We selected local cases based on the Random

Forest model's minimum, median, and maximum absolute prediction errors, using a stable ascending sort of the errors. Local attributions represent the model's internal computational components and do not reflect the effects of interventions.

Reproducibility and independent verification

This analysis is documented through the definition of response variables, the ordering of predictors, median imputation performed exclusively on the training data ("training-only median imputation"), fixed model settings, data partitioning based on harvest year ("harvest-year temporal holdouts"), sensitivity analysis grouped by compartment identifier, observation-level "bootstrap" settings, and the TreeSHAP procedure. The authors report having re-executed the analysis locally using a restricted operational workbook, and have reconciled the aggregated model/fold metrics, global SHAP summaries, "bootstrap" intervals, and additivity diagnostics with the archived results. This process constitutes computational verification by the authors, not an independent numerical reproduction. External parties may review the disclosed methods and the aggregated results; however, an "executable verification" of the workflow would require non-confidential official scripts and a separately developed "synthetic test fixture". Such a test fixture could validate software execution, input validation, data-partitioning logic, output schemas, and determinism. Still, it cannot reproduce empirical results or establish the validity of confidential source records. This document makes no claims regarding the availability or successful implementation of such a test fixture by independent parties. Independent numerical reproduction of the published empirical results requires specifically authorised access to the operational data and its execution environment. The data provider restricts the distribution of observation-level records, whether original or anonymised, along with individual predictions, data-partition membership records, and individual SHAP values. The distinction between methodological transparency, author re-execution, and external reproduction draws on recent literature on reproducibility (Semmelrock et al., 2025; Werner et al., 2024).

Computational workflow and software

Qualification of spatial dependence. While "GroupKFold" prevents identical compartment identifiers from appearing in both the training and testing sets of a given fold, it does not prevent geographically adjacent compartments from being assigned to different folds. Coordinates, adjacency matrices, and mapped management blocks are not provided in the supplied dataset. Consequently, spatial block-based cross-validation, spatial autocorrelation diagnostics, and claims regarding spatial transferability cannot be substantiated. We retain the compartment-grouped analysis solely as a sensitivity analysis at the identifier level. This distinction aligns with recent geospatial machine learning research, which underscores the need for explicit spatial validation designs when geographic transferability is a consideration (Wang et al., 2023).

The computational workflow includes source fields, date-based target construction, predictor sequencing, tailored imputation within training partitions, model parameters, temporal holdouts, sensitivity evaluations by compartment group, bootstrap configurations, and case selection for SHAP analysis. The initial exploratory notebooks cannot be verified as the implementation underlying the final reported results. The process of local re-execution and reconciling its aggregates as described above should not be presented as an independently reproducible analysis. If permitted, a release-ready code package must include a proper executable entry point, defined dependencies, random seeds, expected input schemas, and output validation checks; it must also exclude any sensitive paths, values, identifiers, or metadata.

Data availability and computing environment. The data provider prohibits releasing observation-level records, whether original or de-identified, even upon request, and forbids disclosing operational or estate-level metadata that could identify the associated entities. Consequently, the underlying empirical datasets are not publicly available, and independent reproduction of the reported numerical findings cannot be guaranteed. Releasing any code, analytical dictionaries, aggregated outputs, or synthetic test datasets depends on authorisation from the data provider for the specific files and their contents. The final local rerun process was documented using Python 3.13.5, NumPy 2.3.5, scikit-learn 1.8.0, SHAP 0.50.0, SciPy 1.17.0, and Matplotlib 3.10.8. The

initial exploration environment was not preserved; while a version of the `artifact\_tool\_v2` (version 2.8.66) was subsequently located, its authenticity as the exact version used previously cannot be verified. These versions reflect the final local rerun process rather than bitwise reproducibility of the original execution.

RESULTS AND DISCUSSION

Source audit and predictive performance

A source-level examination confirmed 1,181 observations, 923 unique compartment identifiers, and accurate arithmetic calculations for the target variable in each observation. Two pairs of duplicate environmental columns were identified throughout the records. Seven missing or non-numeric “Harvest Yield” values were addressed via specific imputation during the training phase. No record-level exclusions were applied.

A comparison of predictive performance among the three hold-out harvest-year cohorts is presented in Table 1. It summarises the out-of-time evaluation. Each model was trained on all available observations from harvest years before its respective test year. Across the three test years, the median benchmark yielded the lowest MAE. Although some trained models showed better R^2 or RMSE for 2023, they did not improve the primary criterion, MAE; therefore, these results do not justify using Random Forest as a superior

predictor for scheduling. The visualization of the corresponding results is shown in Figure 1.

A lower MAE value indicates better predictive performance. The median benchmark derived from the training set achieved a lower MAE than the Random Forest method across all three test years.

The “training-median” predictor yielded the lowest MAE value among all models for each temporal holdout period. In 2025, the Random Forest MAE was 27.81 days (95% observation-level bootstrap interval: 24.80–30.88), whereas the median benchmark MAE was 16.15 days. The difference between the Random Forest and median-benchmark MAEs was 11.66 days (95% paired bootstrap interval: 8.89–14.67). These results do not support using this model as a superior predictor for operational scheduling.

Compartment-grouped sensitivity

The results of sensitivity validation by compartment group are presented in Table 2. It presents the sensitivity results, grouped by compartment. We assigned all data points with the same compartment identifier to the same fold, ensuring no identifier overlaps between folds. However, we pooled data from different years, and geographical proximity information was unavailable; consequently, this method is neither a spatial-block holdout validation nor a prospective validation.

The Random Forest model yielded higher MAE values than the training-median baseline

Table 1. Temporal holdout comparison of reference and trained models

Year	Model	Train	Test	MAE (d)	RMSE (d)	R^2
2023	Median	514	282	56.88	109.06	-0.177
2023	Age Ridge	514	282	60.52	100.77	-0.005
2023	Seven Ridge	514	282	60.10	99.13	0.028
2023	Decision tree	514	282	61.06	100.54	-0.000
2023	Random Forest	514	282	60.49	99.36	0.023
2024	Median	796	185	29.28	40.34	-0.067
2024	Age Ridge	796	185	48.44	53.76	-0.895
2024	Seven Ridge	796	185	51.18	58.81	-1.268
2024	Decision tree	796	185	42.26	57.39	-1.160
2024	Random Forest	796	185	48.94	61.01	-1.441
2025	Median	981	200	16.15	20.86	-0.000
2025	Age Ridge	981	200	29.05	33.52	-1.583
2025	Seven Ridge	981	200	30.60	36.19	-2.011
2025	Decision tree	981	200	30.59	44.90	-3.635
2025	Random Forest	981	200	27.81	35.50	-1.896

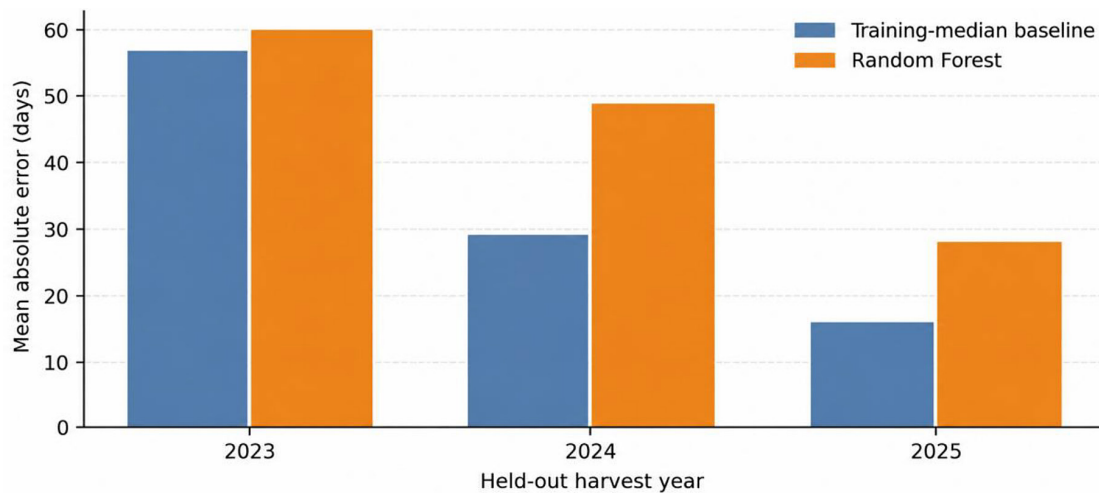


Figure 1. Temporal comparison of mean absolute error (MAE) across held-out harvest-year cohorts

Table 2. Five-fold compartment-grouped validation sensitivity

Fold	Train	Test	Test IDs	RF MAE	Median MAE	RF R ²
1	944	237	185	47.09	35.17	-0.219
2	945	236	184	43.91	37.47	0.062
3	945	236	184	51.53	45.90	0.042
4	945	236	185	48.72	41.88	-0.149
5	945	236	185	52.86	48.81	0.093

across all five compartment-based folds. This indicates that the simpler median-based approach consistently achieves superior prediction accuracy. These findings further underscore that the Random Forest model's increased complexity does not translate into improved predictive performance in a compartment-grouped validation scheme.

Global and local SHAP

The global SHAP attribution values for the 2025 held-out cohort are presented in Table 3.

Table 3 ranks seven predictors by mean absolute SHAP contribution, derived from 200 held-out observations (test data) from 2025. Plant age, phosphorus, and pH contribute the most to the variation in the model's adjusted predictions. These values measure the magnitude of contribution, not independent ecological effects. They should be interpreted with caution, given the low prediction accuracy for different time periods (out-of-time prediction accuracy). The corresponding results are shown in Figure 2.

The bars represent the average absolute contribution in predicted days. The magnitude

of the attribution reflects the fitted model and does not indicate predictive superiority or ecological causality. The corresponding results are shown in Figure 3.

The horizontal position indicates each feature's contribution to the predicted interval (in days), while the colour shows its relative magnitude. The relationships shown are non-causal (i.e., they do not imply cause-and-effect). The predicted SHAP expectation is approximately 95.91 days. The property of additivity has been numerically verified using a held-out cohort. The attribution ranking measures how the fitted model distributes

Table 3. Mean absolute SHAP attributions for the 2025 test cohort (n = 200)

Rank	Predictor	Mean SHAP (days)
1	Crop age	14.11
2	Phosphorus (mg/kg)	7.58
3	pH	6.86
4	Harvest yield (tonnes)	4.92
5	Potassium (mg/kg)	2.96
6	Soil moisture	2.53
7	Nitrogen (mg/kg)	1.14

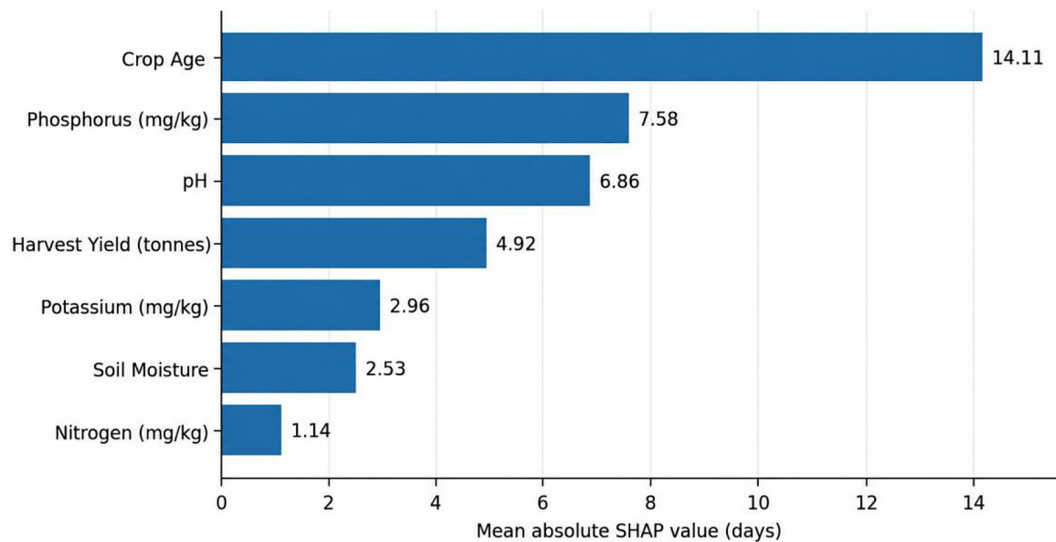


Figure 2. Global TreeSHAP feature attributions for the 2025 temporal holdout (n = 200)

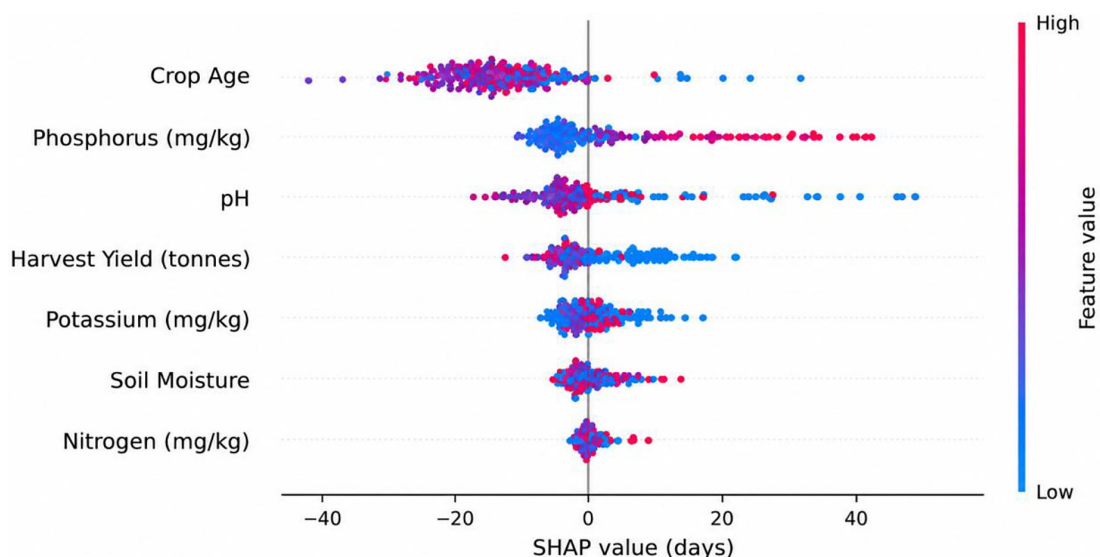


Figure 3. SHAP beeswarm plot for the set-aside 2025 cohort (n = 200)

its predictions across features, not how ecological readiness responds to interventions. Local TreeSHAP explanations have been internally reviewed for observations representing minimum, median, and maximum error levels; however, record-level examples, identifiers, and specific plots are excluded from this publication package due to data disclosure restrictions imposed by the data provider. Only aggregated global attribution summaries are published.

The data provider has not granted permission to publish details such as plantation identifiers, coordinates, land area, clone identities, or proprietary harvesting/replanting protocols. The 923 recorded compartment identifiers serve as unique labels within the dataset and should not

be interpreted as independent geographic spatial blocks. The information permitted for release does not allow for the assessment of spatial autocorrelation or distance-based data leakage.

The observed gap between explainability and reliability is supported by 2025 holdout data. In this dataset, the Random Forest model yielded a Mean Absolute Error (MAE) of 27.81 days—compared to an MAE of 16.15 days for the baseline (the median from the training data)—even though the model assigned the highest average absolute SHAP contributions to variables such as plant age, phosphorus levels, and pH. The MAE rankings remained consistent across all three temporal test sets and all five data folds (grouped by compartment). Consequently, the

primary conclusion highlights the limitations of using model-based explanations as evidence of predictive utility; these findings neither account for shifts in operational scheduling practices nor indicate whether soil characteristics causally influence replanting decisions.

The key finding is that while tailored nonlinear models consistently produce interpretable attributions, they do not outperform simple benchmarks in predicting future harvest cohorts. Therefore, this study's scientific contribution is methodological and diagnostic: it disentangles corporate scheduling records from ecological readiness, shows how testing across temporal dimensions and compartment-based groupings changes interpretations of seemingly explainable models, and uncovers specific limitations in source-data quality. These results do not constitute a field-validated guide for replanting. This caution is reinforced by evidence that data leakage and inadequate evaluation design can overestimate the superiority of complex machine learning methods over simpler benchmarks (Kapoor and Narayanan, 2023).

Scheduling decisions may depend on nursery capacity, labour availability, accessibility, weather-dependent time windows, and management policies; none of these factors is verified as model inputs. Target scheduling may also reflect organisational conventions rather than constraints imposed by soil conditions. Attribute values such as plant age, phosphorus content, and pH cannot establish causal relationships or optimal soil amendment strategies. The low baseline error rate in 2025, compared to the adjusted model, suggests a shift in operational practices or target distribution, though the exact cause remains unknown.

This study is limited by missing decision making timestamps, independently verified planting outcomes, measurement origins, geographic coordinates, and sensor calibration records. Observation level bootstrap intervals do not account for potential dependencies within compartments. The targets represent company-established scheduling intervals rather than biological readiness. While the reported data replays and aggregate reconciliations verify from the authors' perspective, and external inspection of the methods and aggregates is possible, independent numerical reproduction remains constrained by the lack of formal access to observation-level operational data. Any proposed execution test using synthetic data would only assess pipeline behaviour, not reproduce empirical estimates. Using compartment identifiers

for 'GroupKFold' does not enforce geographic separation and cannot substitute for spatial block validation. These limitations hinder verification and geographic transferability.

CONCLUSIONS

This study retrospectively examines the replanting intervals established by a company for its industrial Eucalyptus plantations, utilizing historical management records, stand characteristics, and soil properties. The results indicate that while a "Random Forest" model identifies interpretable relationships between plantation characteristics and replanting schedules, these relationships do not yield superior predictive performance. Across all three testing (hold-out) periods and five validation groups categorised by compartment, the median value from the training set consistently achieved a lower mean absolute error than the Random Forest model.

During the 2025 test period, which comprised 200 observations, the Random Forest model produced a mean absolute error of 27.81 days, compared to a benchmark median of 16.15 days. This 11.66-day difference, with a 95% confidence interval of 8.89–14.67 days derived via a paired bootstrap method at the observation level, highlights the limited reliability of the model's predictions. A global SHAP analysis identified stand age (14.11 days), phosphorus content (7.58 days), and soil pH (6.86 days) as the primary contributors to the model's predictions. However, these feature attributions reflect model-specific associations rather than evidence of ecological causality or predictive superiority.

These findings reveal a discrepancy between the ability to explain historical operational scheduling patterns and the reliability of predictions. The retrospective nature of the company's scheduling targets, the lack of independently verified planting-outcome data, uncertainty about the origin of measurements, and potential interdependencies among repeated plot records limit interpretation. Consequently, the observed scheduling intervals cannot be interpreted as biologically optimal replanting periods. Overall, this study demonstrates that explainable machine learning predictions do not consistently outperform simple statistical benchmarks in predicting historical Eucalyptus replanting schedules. Therefore, interpret the predicted results strictly as company

established retrospective scheduling intervals, not as biological readiness, actual planting timing, or recommendations for ecologically optimal replanting dates. The final pipeline was re-run locally against confidential workbooks, and the aggregated outputs were reconciled. Nevertheless, confidentiality constraints preclude unrestricted independent reproduction, and the absence of geographically referenced compartment information prevents spatial block validation. These limitations hinder independent verification and the geographic applicability of the findings.

Acknowledgements

The authors thank the company that provided plantation operational data for research purposes. The authors do not disclose the company's identity to maintain confidentiality.

REFERENCES

- Arif, S., Sakka, S., Nurmiaty, N., Rahmad, R., Syamsuddin, E., Wajedy, M. F., Laode, A. M. W., & Syamsuddin, S. (2025). Integrating random forest and irrigation management in geographic information systems-based land suitability and rice productivity modeling in tropical landscapes. *Ecological Engineering & Environmental Technology*, 26(7), 113–125. <https://doi.org/10.12912/27197050/205069>
- Dai, Q., Wang, T., Wei, P., & Fu, Y. (2023). Effects of successive planting of Eucalyptus plantations on tree growth and soil quality. *Sustainability*, 15(8), 6746. <https://doi.org/10.3390/su15086746>
- Hardiyanto, E. B., Inail, M. A., Nambiar, S., & Mendham, D. S. (2024). Sustaining plantation forest productivity in Sumatra over three decades: From acacias to eucalypts. *Forest Ecology and Management*, 553, 121613. <https://doi.org/10.1016/j.foreco.2023.121613>
- Jemeljanova, M., Kmoch, A., & Uuema, E. (2024). Adapting machine learning for environmental spatial data – A review. *Ecological Informatics*, 81, 102634. <https://doi.org/10.1016/j.ecoinf.2024.102634>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- Marwiji, M. Y. F., Soma, A. S., Arif, S., & Rahmat, S. (2026). Integration of Sentinel-1 synthetic aperture radar and random forest algorithm for high-precision flood hazard modeling in a data-scarce tropical watershed. *Ecological Engineering & Environmental Technology*, 27(2), 198–215. <https://doi.org/10.12912/27197050/217188>
- Pai, D. G., Balachandra, M., & Kamath, R. (2025). Explainable AI in agriculture: Review of applications, methodologies, and future directions. *Engineering Research Express*, 7(3), Article 032202. <https://doi.org/10.1088/2631-8695/ae086a>
- Powae, G. M., & Rao, B. K. R. (2024). Quantification of soil nutrient stocks and stoichiometric ratios in Eucalyptus pellita biomass plantation chronosequence in Papua New Guinea. *Catena*, 247, 108564. <https://doi.org/10.1016/j.catena.2024.108564>
- Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1), 2400304. <https://doi.org/10.1002/aisy.202400304>
- Sembiring, N., & Ishak, A. (2025). Optimization of eucalyptus replanting schedules: A dynamic model for efficient land use and workforce allocation. *Ecological Engineering & Environmental Technology*, 26(11), 17–33. <https://doi.org/10.12912/27197050/211645>
- Sembiring, N., & Napitupulu, H. L. (2025). Eucalyptus replanting a comprehensive bibliometric review of research trends. *Ecological Engineering & Environmental Technology*, 26(4), 130–149. <https://doi.org/10.12912/27197050/200822>
- Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., & Kowald, D. (2025). Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine*, 46(2), e70002. <https://doi.org/10.1002/aaai.70002>
- Wang, Y., Khodadadzadeh, M., & Zurita-Milla, R. (2023). Spatial+: A new cross-validation method to evaluate geospatial machine learning models. *International Journal of Applied Earth Observation and Geoinformation*, 121, 103364. <https://doi.org/10.1016/j.jag.2023.103364>
- Werner, L., Layaïda, N., Genevès, P., Euzenat, J., & Graux, D. (2024). Reproduce, replicate, reevaluate. The long but safe way to extend machine learning methods. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14), 15850–15858. <https://doi.org/10.1609/aaai.v38i14.29515>