JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 287 – 292 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4455
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

PREDICTING TEA HARVEST PRODUCTION AT BAH BUTONG USING RANDOM FOREST AND HISTORICAL DATA

Hafizdh Huda Prayoga[1*], Yusuf Ramadhan Nasution[1]
[1]Ilmu Komputer, Universitas Islam Negeri Sumatera Utara
email: [*]hafizhudaprayoga@gmail.com

Abstract: Accurate forecasts of tea harvest production are important for workforce planning, factory operations, and marketing decisions, yet conventional estimation in plantations often relies on field experience and can be biased and less adaptive to changing conditions. This study aims to develop a Random Forest Regression model to predict tea harvest production at the Bah Butong tea plantation using historical operational and climate-related data. The dataset consists of 60 monthly records (2020–2024) with six predictor variables: rainfall (mm), number of rainy days, pest level, weed level, number of harvested trees and land area. Data were split into 80% training (48 samples) and 20% testing (12 samples). Model hyperparameters were optimized using RandomizedSearchCV with RepeatedKFold cross-validation (5 folds, 3 repeats). The tuned model achieved MSE of 668,980,524.45, RMSE of 25,864.66 kg, MAE of 19,838.69 kg, and MAPE of 7.59% on the test set. The results indicate that the model can provide practical production estimates, with errors averaging about 7–8% of the actual production. Feature importance analysis shows that the number of harvested tea bushes and cultivated area contribute most to predictions. Future work should extend the historical period and incorporate time-based features (seasonality/lag) for improved forecasting.

Keywords: hyperparameter tuning; production prediction; random forest; regression; tea harvest

Abstrak: Perkiraan akurat produksi panen teh sangat penting untuk perencanaan tenaga kerja, operasional pabrik, dan keputusan pemasaran, namun estimasi konvensional di perkebunan seringkali bergantung pada pengalaman lapangan dan dapat bias serta kurang adaptif terhadap perubahan kondisi. Studi ini bertujuan untuk mengembangkan model Regresi Random Forest untuk memprediksi produksi panen teh di perkebunan teh Bah Butong menggunakan data operasional dan data terkait iklim historis. Dataset terdiri dari 60 catatan bulanan (2020–2024) dengan enam variabel prediktor: curah hujan (mm), jumlah hari hujan, tingkat hama, tingkat gulma, jumlah pokok panen, dan luas lahan. Data dibagi menjadi 80% data pelatihan (48 sampel) dan 20% data pengujian (12 sampel). Parameter model dioptimalkan menggunakan RandomizedSearchCV dengan validasi silang RepeatedKFold (5 lipatan, 3 pengulangan). Model yang telah disempurnakan mencapai MSE sebesar 668.980.524,45, RMSE sebesar 25.864,66 kg, MAE sebesar 19.838,69 kg, dan MAPE sebesar 7,59% pada set data uji. Hasil tersebut menunjukkan bahwa model dapat memberikan estimasi produksi yang praktis, dengan kesalahan rata-rata sekitar 7–8% dari produksi aktual. Analisis kepentingan fitur menunjukkan bahwa jumlah semak teh yang dipanen dan luas lahan budidaya paling berkontribusi pada prediksi. Pekerjaan selanjutnya harus memperpanjang periode historis dan menggabungkan fitur berbasis waktu (musiman/lag) untuk peramalan yang lebih baik.

Kata kunci: panen teh; prediksi produksi; random forest; regresi; tuning parameter

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 287 – 292 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4455
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Accurate forecasting of tea harvest production is essential for workforce planning, factory operations, and marketing decisions in plantation-based industries [1], [2]. Tea yields are influenced by agroclimatic conditions and field management practices; for example, rainfall patterns and plucking standards can affect shoot population and yield dynamics [3]. In addition, rainfall extremes associated with climate change can increase production variability, making planning more challenging [4].

In practice, plantation planning is still often based on field experience and routine operational rules, which are subjective and less adaptive to changing conditions [5]. Leveraging historical production records together with machine-learning models offers a data-driven approach by learning patterns embedded in the data [6]. Previous studies have reported yield forecasting using multiple linear regression for several commodities in Indonesia; however, linear models are limited in capturing nonlinear relationships and interactions among factors [7].

Ensemble methods such as Random Forest are widely used because they reduce variance by aggregating many decision trees and have been shown to be robust in various prediction tasks [8]–[11]. However, studies focusing on tea plantations—particularly Bah Butong—remain limited. Few studies simultaneously combine climate indicators (rainfall, rainy days) and cultivation indicators (pest level, weed level, number of harvested bushes, and cultivated area) and provide interpretability analyses such as feature importance and residual diagnostics [2], [3], [9], [10]. Moreover, practical adoption requires reproducible workflows using accessible tools such as Python and cloud-based notebooks [11], [12], within a clear quantitative research framework [13].

Therefore, this study aims to develop a Random Forest Regression model to predict tea harvest production at the Bah Butong Plantation using historical data from 2020–2024. The model is optimized using Randomized SearchCV with RepeatedKFold and evaluated on a held-out test set using MSE, RMSE, MAE, and MAPE [8]. In addition to performance metrics, this study provides diagnostic and interpretability analyses, including actual-versus-predicted visualization, residual analysis, and feature importance, to support operational decision-making [9].

The contributions of this study are to develop a Random Forest Regression model for forecasting tea harvest production at the Bah Butong Plantation using historical data from 2020 to 2024, to optimize the model hyperparameters using RandomizedSearchCV with RepeatedK-Fold in order to obtain a more stable and reliable configuration, and to evaluate model performance using MSE, RMSE, MAE, and MAPE while improving interpretability through actual-versus-predicted plots, residual diagnostics, and feature importance analysis.

METHODS

The study uses 60 monthly records from 2020–2024. Predictor variables include rainfall (mm), number of rainy days, pest level, weed level, number of harvested bushes, and cultivated area. The target variable is production (kg). Numeric values were coerced to valid numbers, and missing values were handled using median imputation.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 287 – 292 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4455
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

The data were split into training and test sets at an 80:20 ratio (48 training samples and 12 test samples) using random_state = 42. The Random Forest model was trained on the training set and evaluated on the test set.

Hyperparameter optimization was performed using RandomizedSearchCV with RepeatedKFold cross-validation (5 folds, 3 repeats). The tuned parameters include n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, and bootstrap. The best configuration is shown in Table 1.

Data processing and modeling were conducted using Python in the Google Colab environment to facilitate experimentation and reproducibility [11], [12]. The research design is quantitative, using predictive modeling based on historical data [12]. The error metrics are defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum SE_i \quad (1)$$

$$RMSE = \sqrt{MSE} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n} |e_i| \quad (3)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left(\left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \right) \quad (4)$$

Model evaluation uses MSE, RMSE, MAE, and MAPE. Lower error-metric values indicate better predictive performance.

Table 1. Best hyperparameter settings for Random Forest Regression

Parameter	Value	Description
n_estimators	1000	Number of trees
max_depth	4	Maximum depth
min_samples_split	4	Minimum samples to split
min_samples_leaf	5	Minimum samples at leaf
max_features	log2	Features per split
bootstrap	True	Bootstrap sampling
random_state	42	Random seed
criterion	squared_error	Impurity criterion
imputer	median	Imputation strategy

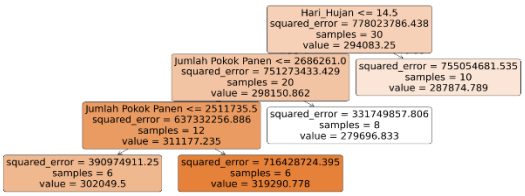


Figure 1. Example decision-tree structure in Random Forest Regression

RESULTS AND DISCUSSION

The tuned Random Forest Regression model was used to predict the test set (12 samples). The evaluation results are summarized in Table 2. The model achieved an MSE of 668,980,524.455 kg^2, RMSE of 25,864.658 kg^2, MAE of 19,838.689 kg, and MAPE of 7.59%. Because production is measured in kilograms, MSE has units of kg^2 (squared error), whereas RMSE and MAE are expressed in kg and MAPE is expressed in percent. A MAPE of 7.59% indicates that, on average, the prediction error is about 7–8% of the actual production.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 287 – 292 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4455
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Table 2. Model evaluation results on the test set

Metric	Value
MSE (kg^2)	668,980,524.455
RMSE (kg)	25,864.658
MAE (kg)	19,838.689
MAPE (%)	7.59%

Figure 2 compares actual and predicted production on the test set. Overall, the predictions follow the production pattern; however, for several samples with extreme production values, the model tends to produce predictions closer to the mean, resulting in larger errors.

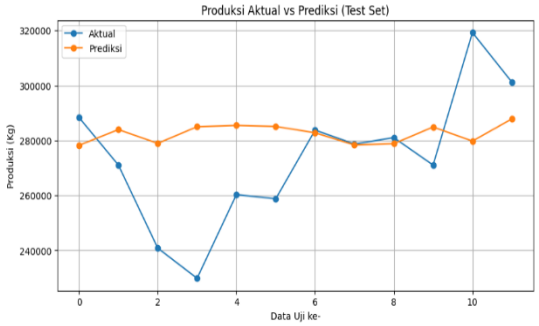


Figure 2. Actual vs. predicted production (test set)

Figure 3 shows that some points lie close to the diagonal line, indicating good predictions for part of the test data. However, several points are far from the diagonal, suggesting large prediction errors for some samples.

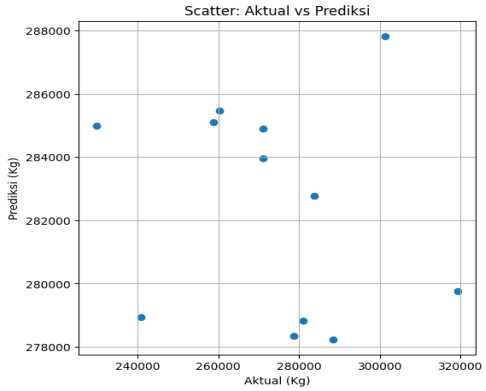
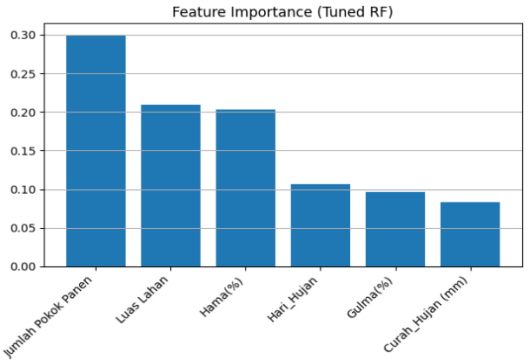


Figure 3. Scatter plot of actual vs. predicted values (test set)

The feature-importance analysis (Figure 4) shows that the number of harvested bushes and cultivated area contribute most to production predictions. Pest level (%) also contributes noticeably, whereas rainfall (mm) has a relatively smaller contribution in the built model.



Gambar 4. Feature importance model Random Forest (tuned)

Figure 5 shows the residual plot on the test set. Each point represents one test sample, and the residual axis indicates the magnitude of prediction error. Residuals close to zero indicate accurate predictions, while residuals far from zero indicate larger errors.

1. Positive residuals indicate underestimation (predicted values are lower than actual values).
2. Negative residuals indicate overestimation (predicted values are higher than actual values).

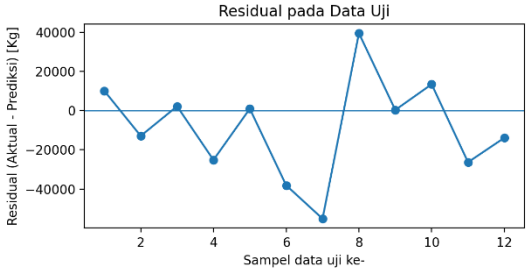


Figure 5. Residual plot on the test set

This pattern is consistent with the model evaluation results, i.e., MSE =

668,980,524.45 kg², RMSE = 25,864.66 kg, MAE = 19,838.69 kg, and MAPE = 7.59%. The relatively moderate MAPE indicates that the average percentage error is still acceptable; however, a few large residuals can increase MSE because the errors are squared.

Overall, the residual analysis suggests that the model is adequate for general estimation, but improvements are needed to reduce large errors, for example by adding more specific agronomic variables and applying time-aware validation if the data are sequential.

CONCLUSION

Based on Random Forest Regression modeling using rainfall (mm), number of rainy days, pest level, weed level, number of harvested bushes, and cultivated area as predictors, the model achieved MSE of 668,980,524.45 kg², RMSE of 25,864.66 kg, MAE of 19,838.69 kg, and MAPE of 7.59% on the test set. These results indicate that Random Forest can be used as a historical-data-based approach for tea production prediction to support operational planning.

This study has several limitations, including a relatively small dataset (60 monthly records), the presence of extreme production fluctuations in several months, and predictor variables that do not capture all agronomic factors (e.g., fertilization intensity, plant age, plucking standard, and soil conditions). In addition, the data split was random, so seasonality patterns were not explicitly modeled.

For future research, the historical observation period can be extended and additional agronomic variables can be included to improve data representativeness. It is also recommended to consider time-based features, such as month and lag variables of production or rainfall, as well as time-series evaluation schemes like TimeSeriesSplit to better capture seasonal patterns. In addition, alternative models, such as Gradient Boosting, may be evaluated as benchmark methods to obtain better predictive performance.

ACKNOWLEDGMENT

The authors thank the Bah Butong Tea Plantation for providing the data used in this study.

BIBLIOGRAPHY

[1] M. A. S. Dr. Ir. Anna A. Susanti, M.Si Rhendy Kencanaputra W. S.Si, *Outlook komoditas perkebunan*. 2024.

[2] I. N. Deva, H. Cipta, F. Rakhmawati, U. Islam, and N. Sumatera, "PTPN IV UNIT BAH BUTONG MENGGUNAKAN," vol. 4307, no. August, pp. 1103–1114, 2024.

[3] L. Long, Y. Shi, L. Ma, and J. Ruan, "Characterization of Young Shoot Population , Yield , and Nitrogen Demands of Tea (Camellia sinensis L .) Harvested under Different Standards," 2023.

[4] S. Suhadi, F. Mabruroh, A. Wiyanto, and I. Ikra, "Analisis Fenomena Perubahan Iklim Terhadap Curah Hujan Ekstrim," *Opt. J. Pendidik. Fis.*, vol. 7, no. 1, pp. 94–100, 2023, doi: 10.37478/optika.v7i1.2738.

[5] R. Affairs, "Christina Mey Rahayu1, Sofyan Zaman2*, Arya Widura Ritonga2," *Manaj*.

Pemanenan Kelapa Sawit (Elaeis guineensis Jacq.) di Kebun Tandun, Kabupaten Kampar, Riau, vol. 12, no. 2, pp. 266–275, 2024.

[6] N. Afrilia S, F. Frazna Az-Zahra, and P. Prajoko, "Prediksi Hasil Panen Wortel Menggunakan Algoritma Regresi Linear Berganda," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 5, pp. 10255–10262, 2024, doi: 10.36040/jati.v8i5.10954.

[7] R. Andia, K. Kaslani, S. Eka Permana, and T. Handayani, "Peramalan Hasil Panen Padi Kabupaten Cirebon Menggunakan Algoritma Regresi Linear Berganda," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 738–747, 2024, doi: 10.36040/jati.v8i1.8446.

[8] J. (2023). James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, "Statistical Learning 2.1," 2023, pp. 1–2.

[9] F. Al Farikhi and R. W. D. Pramono, "Perbandingan algoritma classification and regression tree (cart) dan random forest (rf) untuk klasifikasi penggunaan lahan pada google earth engine," *J. Spat. Wahana Komun. dan Inf. Geogr.*, vol. 23, no. 2, pp. 170–179, 2023, doi: 10.21009/spatial.232.09.

[10] M. Huda, "PENERAPAN METODE RANDOM FOREST PADA PREDIKSI PENILAIAN NILAI ASET KJPP SIG MALANG BERBASIS WEB," vol. 183, no. 2, pp. 153–164, 2023.

[11] M. R. S. Alfarizi, M. Z. Al-farish, M. Taufiqurrahman, G. Ardiansah, and M. Elgar, "Penggunaan Python Sebagai Bahasa Pemrograman untuk Machine Learning dan Deep Learning," *Karya Ilm. Mhs. Bertauhid (KARIMAH TAUHID)*, vol. 2, no. 1, pp. 1–6, 2023.

[12] R. Gelar Guntara, "Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 5, no. 1, pp. 55–60, 2023, doi: 10.47233/jteksis.v5i1.750.

[13] M. Waruwu, S. N. Pu`at, P. R. Utami, E. Yanti, and M. Rusydiana, "Metode Penelitian Kuantitatif: Konsep, Jenis, Tahapan dan Kelebihan," *J. Ilm. Profesi Pendidik.*, vol. 10, no. 1, pp. 917–932, 2025, doi: 10.29303/jipp.v10i1.3057.