Analysis of MSMEs' Cassava Production Efficiency Using a Comparison of Machine Learning Models in Jember Regency

Danang Kumara Hadi*[1,2] and Yuta Sato[2]
[1]Department of Agroindustrial Technology, Faculty of Agriculture
Universitas Muhammadiyah Jember, Jl. Karimata No. 49, Jember 68124, Indonesia.
[2]Graduate School of Engineering, Ibaraki University, Japan
Email: danangkumara@unmuhjember.ac.id*

Received: May-23-2025; Accepted: June-25-2025; Published: June-30-2025

Abstract

Cassava is one of Indonesia's agro-industrial commodities, but many Micro, Small, and Medium Enterprises (MSMEs) in the cassava processing industry face difficulties in achieving optimal production efficiency. This study aims to evaluate the efficiency of cassava processing production systems in MSMEs in Jember by comparing machine learning algorithms (Linear Regression, Random Forest, Support Vector Regression (SVR), and XGBoost) to predict output and key efficiency factors. The data used consists of 250 data points: 80% for model training and 20% for testing to build a machine learning-based prediction model, with input features production processing as the X-axis, and output in the form of production volume as the Y-axis. Data preprocessing, exploratory data analysis, and modeling were conducted using Python, with evaluation based on MAE, RMSE, and R² metrics. Among the tested models, Random Forest demonstrated the best performance with an R² value of 0.990. Sensitivity analysis revealed that production output increases significantly with the addition of labor and machines, with an optimal configuration of 15–20 workers and 2–3 machines per batch. The study concludes that focusing on overall production efficiency rather than merely increasing resources is the most effective strategy.

Keywords: *Cassava, Efficiency Analysis, Machine Learning Algorithm, Prediction Model.*

1. INTRODUCTION

Cassava (*Manihot esculenta*) is one of the strategic food commodities in Indonesia that has high economic value and great potential as an agro-industrial raw material, such as the tapioca flour industry (Aminanti Suraya Putri et al., 2020), modified starch (Hamidi and Banowati, 2019), bioethanol (Shanavas et al., 2011), and snacks or chips (Hadi et al., 2021). Jember regency has the advantage of cassava commodity which lies in its high productivity, tolerance to marginal land, as well as a relatively fast planting cycle and suitable soil for planting (BPS kabupaten jember, 2020). However, in practice, cassava micro, small and medium enterprises (MSMEs) in Jember Regency often face challenges in achieving optimal production efficiency.

The efficiency of the production system is strongly influenced by many factors, such as the number of workers (Sosa-Perez et al., 2020), machine capacity (Niekurzak et al., 2023), process time (Rosova et al., 2022) at each stage of production (washing, peeling, drying), to the work system per shift. Inefficiency in the management of the production process can lead to decreased productivity, waste of resources, and high operational costs (Kumara et al., 2023). Therefore, an analytical approach is needed to understand the dynamics of the production system and identify crucial points that affect efficiency (Hadi et al., 2023).

Along with the development of data technology in the era of artificial intelligence (AI), machine learning (ML) based approaches have become one of the effective methods in predicting, classifying, and optimizing industrial systems (Kreuzberger et al., 2023; Yan, 2022; Zhang et al., 2022). ML is able to model complex relationships between various production variables and provide accurate predictions of output or work efficiency. Another advantage is ML's ability to learn patterns

ISSN 2252-6137 (print), ISSN 2302-3848 (online), doi: 10.22146/aij.v12i1.106018

from historical data and recommend data-driven decision making strategies for system improvement (Kumar et al., 2020; Moosavi et al., 2020; Zaki et al., 2020). However, the application of machine learning in small and medium scale agro-industries (Benos et al., 2021; Elbasi et al., 2023), especially in the context of cassava production, is still very limited. Many industry players do not have a systematic approach in measuring and improving their production efficiency. In fact, with the right approach, this technology can provide strategic recommendations based on actual and historical data simulations (D'Amour et al., 2022).

This research was conducted to examine how the performance of the cassava production system is viewed from actual process data, identify the factors that have the most influence on production efficiency, and determine the most accurate machine learning algorithm in predicting output or work efficiency at XYZ MSME. The main objective of this research is to build a Python-based predictive model using several algorithms such as Linear Regression, Random Forest, Support Vector Regression (SVR), and XGBoost, and evaluate the performance of each model based on standard metrics (Subasi, 2020). The results of this research are expected to make a real contribution to the development of production efficiency strategies in the cassava agro-industry sector, especially for small and medium scale industry players who want to apply data-based analytical approaches in a practical and measurable manner. This comparative analysis also shows that Random Forest excels in handling non-linear data and has good feature interpretability, but requires more computational resources than Linear Regression. Thus, this research not only provides a predictive model, but also provides practical considerations regarding algorithm selection based on data complexity and user requirements. Until now, there has been no research that directly compares the performance of several machine learning algorithms in the context of output prediction and production efficiency in the cassava processing industry in Jember. Therefore, this research makes a new contribution with an ML model comparison approach that is applicable and specific to local agroindustry conditions.

2. MATERIAL AND METHODS

2.1 Data Collection

Data is collected from the actual production process and the results of simulations that have been carried out previously at XYZ MSME. The dataset used amounted to 250 observation data. List the input features including the number of workers, number of machines, number of shifts, washing time, peeling, drying, total production time, and process efficiency as the X-axis and output in the form of production quantities as the Y-axis. The resulting dataset contains a number of important variables that represent operational conditions and process performance, such as the number of workers, the number of machines used, the number of work shifts, the time required at each stage of the process, the total production time, the amount of output produced, and the overall level of process efficiency. The research scheme is presented in Figure 1.

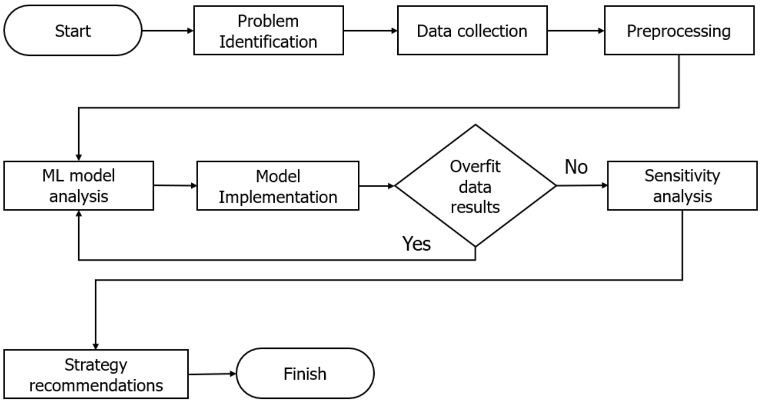


Figure 1. Research scheme of efficiency analysis in cassava processed product production system using Machine Learning Algorithms

The dataset is then randomly divided into two parts: 80% as training set and 20% as testing set. This division aims to allow the model to learn patterns from historical data and be tested using previous data. The training data is used to train the machine learning model to understand the pattern of the production process, while the testing data is used to test the performance of the model against new data that has never been seen before. This separation is important so that the model can produce accurate predictions and not experience overfitting.

2.2 Preprocessing & Exploratory Data Analysis (EDA)

It is important to prepare and understand the data before modeling. Using Pandas, NumPy, Matplotlib, and Seaborn, the data is analyzed through feature scaling, value distribution, correlation between variables, and outlier identification. The goal is to ensure the data is ready and relevant for the model training process (Subasi, 2020). The selection of machine learning algorithms in this study is based on data characteristics and prediction needs. Linear Regression was chosen as the basic model that is easy to interpret. Random Forest was chosen for its ability to handle non-linear data and robust to outliers. SVR was chosen for its ability to work well on high-dimensional data and limited sample size. XGBoost was chosen for its efficiency in boosting and its ability to produce high accuracy on complex data. This diverse approach allows a thorough evaluation of model performance based on the context of cassava production data (Kreuzberger et al., 2023; Kumar et al., 2020; Subasi, 2020).

2.2.1 Linear Regression

Linear Regression is used to model the linear relationship between input (feature) and output (target) variables (Manurung et al., 2024). General formula:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \quad (1)$$

Where:

$\hat{y}$	: predicted output
xi	: i-th feature
βi	: regression coefficient
ε	: error/residuals

2.2.2 Random Forest Regressor

Random Forest is a bagging-based ensemble algorithm that uses many decision trees to generate average predictions (Erkamim et al., 2023).

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (2)$$

Where:

T	: number of trees
$h_t(x)$	: prediction of the t tree
$\hat{y}$	: average prediction of all trees

2.2.3 Support Vector Regression (SVR)

SVR seeks a regression function that has a minimum margin of error ε and is tolerant of small deviations (Sepri and Fauzi, 2020).

$$min\frac{1}{2}\|w\|^2 \text{ subject to: } \left|y_i - (w.x_i + b)\right| \leq \varepsilon \quad (3)$$

Where:

w	: weight vector
b	: bias/intercept
ε	: error tolerance

2.2.4 XGBoost Regressor

XGBoost is a boosting algorithm that builds the model incrementally, by adding new trees that focus on correcting the errors of the previous trees (Syafei and Efrilianda, 2023).

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{4}$$

Where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{5}$$

Where:

l : loss function (e.g. RMSE)
$\Omega(f)$: regularization to avoid overfitting
T : number of leaves in the decision tree
λ, γ : regularization parameters

2.2.5 Cross-Validation dan Hyperparameter Tuning k-Fold Cross-Validation

Cross-validation is used to ensure the model does not overfit or underfit. The data is divided into k parts. Each part becomes the test set once and the training set k-1 times (Kusunartutik and Dwidayati, 2022).

$$Score = \frac{1}{k}\sum_{i=1}^{k} Metric \tag{6}$$

3. RESULTS AND DISCUSSION

3.1 Statistical Analysis of Data

The analysis is carried out in stages, starting from data exploration, predictive model building, model performance evaluation, to interpretation of prediction results and sensitivity analysis of important variables (Bhargav et al., 2024). Each stage is discussed in detail to illustrate the accuracy of the model as well as the relevance of the results to the actual production system being analyzed.

Descriptive statistics are presented in Figure 2 to provide an overview of the distribution of data obtained from the production system in cassava processing MSMEs. This visualization includes key variables such as washing, peeling, drying, total process time, total output, and production efficiency. Each boxplot displays the minimum, median, and maximum values, as well as the possibility of outliers, thus helping to understand the characteristics and stability of the production process in the field. From the image data, data transformation using scaling and data normalization is divided into 80% for training and 20% for testing. Figure 3 shows the support data of human resources and machinery on production execution.

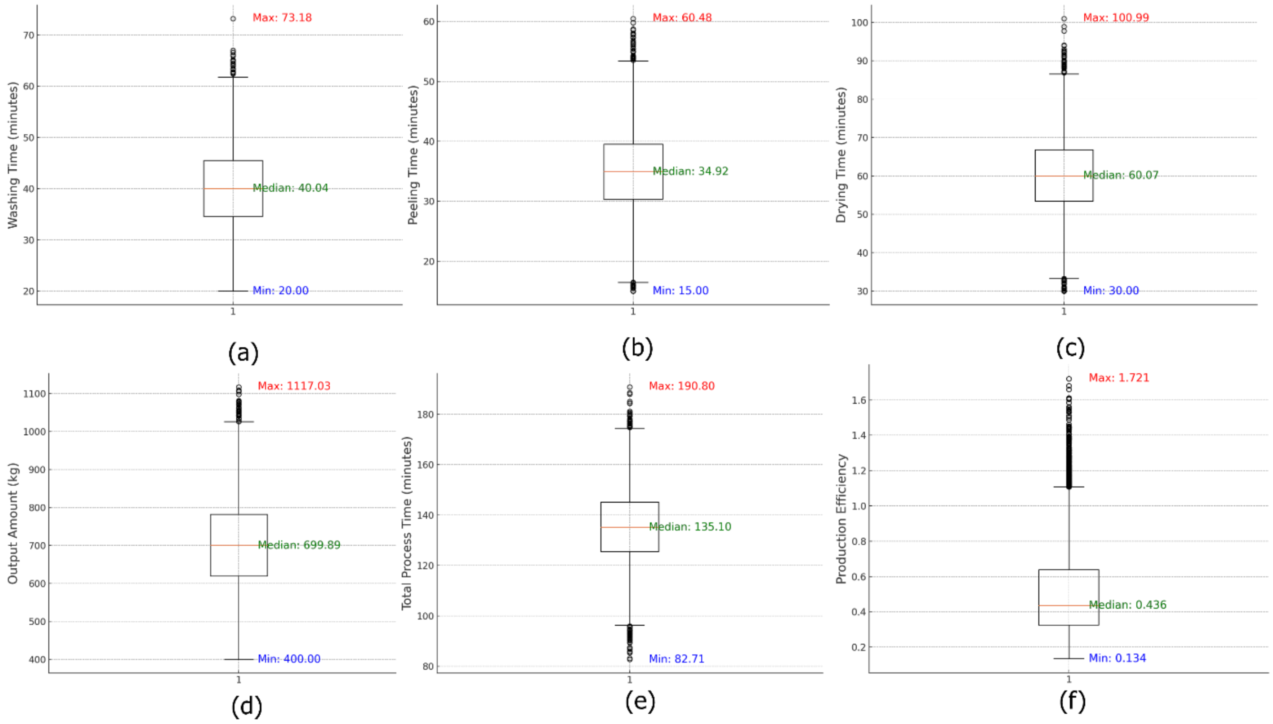


Figure 2. Box-and-whisker plot overview of the distribution of data obtained from the production system in cassava processing MSMEs (a) Washing Time (b) Peeling Time (c) Drying Time (d) Output Amount (e) Total Process Time (f) Production Efficiency

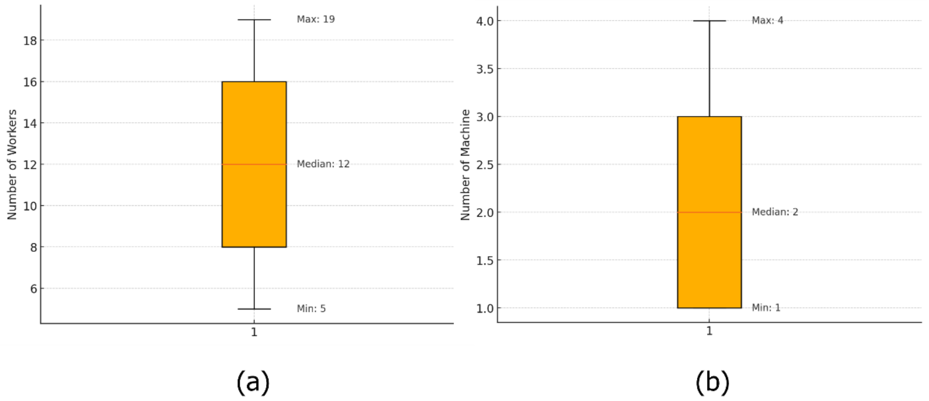


Figure 3. Supporting data for production implementation each batch (a) Number of workers data (b) Number of machine data

3.2 Modeling Implementation Results

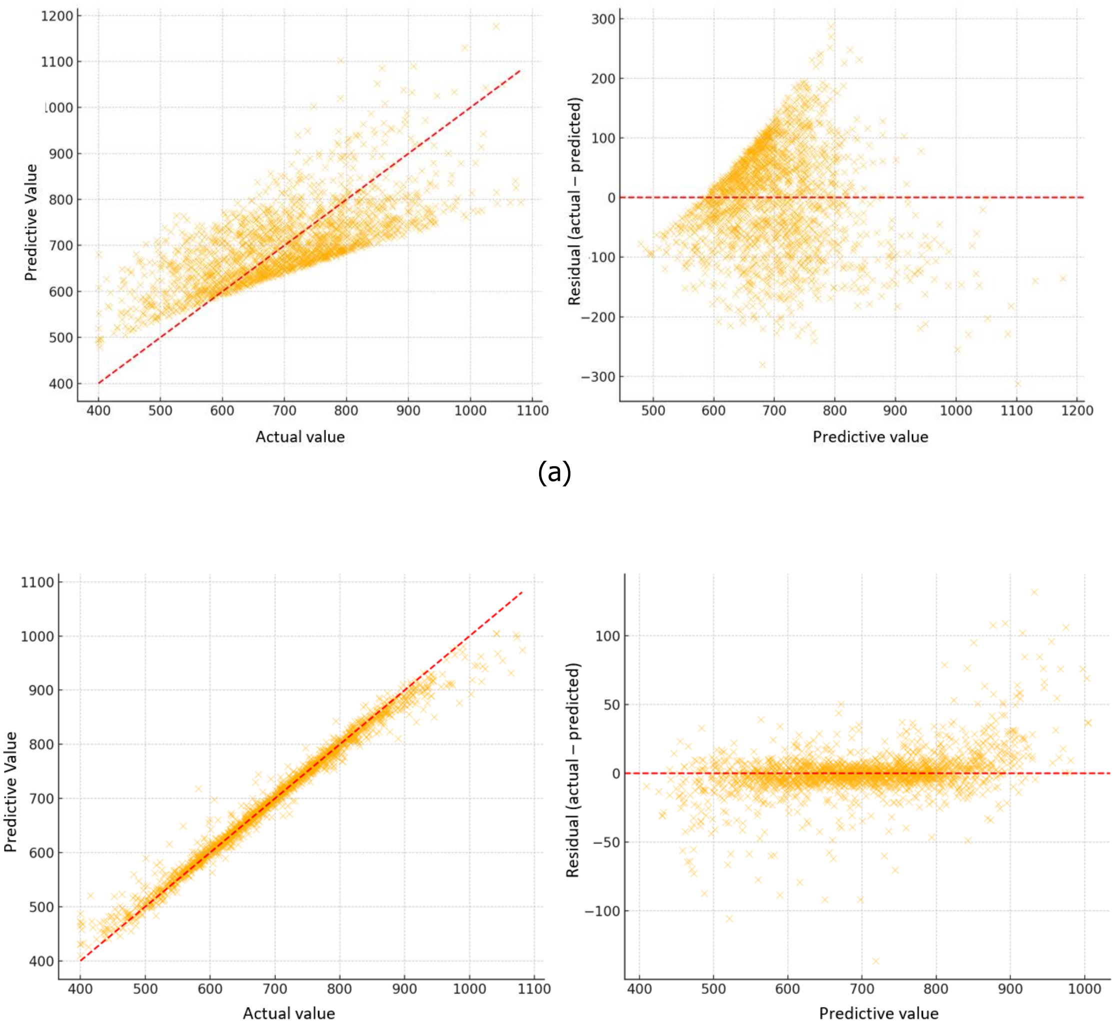
Model evaluation and interpretation were conducted to measure the performance of the machine learning algorithms used in predicting cassava production output. Evaluation is carried out using three main metrics, namely Mean Absolute Error (MAE) to determine the average absolute error, Root Mean Square Error (RMSE) to measure the deviation of predictions from actual values, and R² Score to see the extent to which the model is able to explain variations in output data. The comparison metric research results of the data obtained are presented in Table 1.

Table 1. Metric of model comparison of data

Model	MAE	RMSE	R² Score
Linear Regression	69.52	85.97	0.48
Random Forest Regressor	9.54	16.84	0.99
Support Vector Regression (SVR)	5.30	11.74	0.98
XGBoost	24.14	33.38	0.92

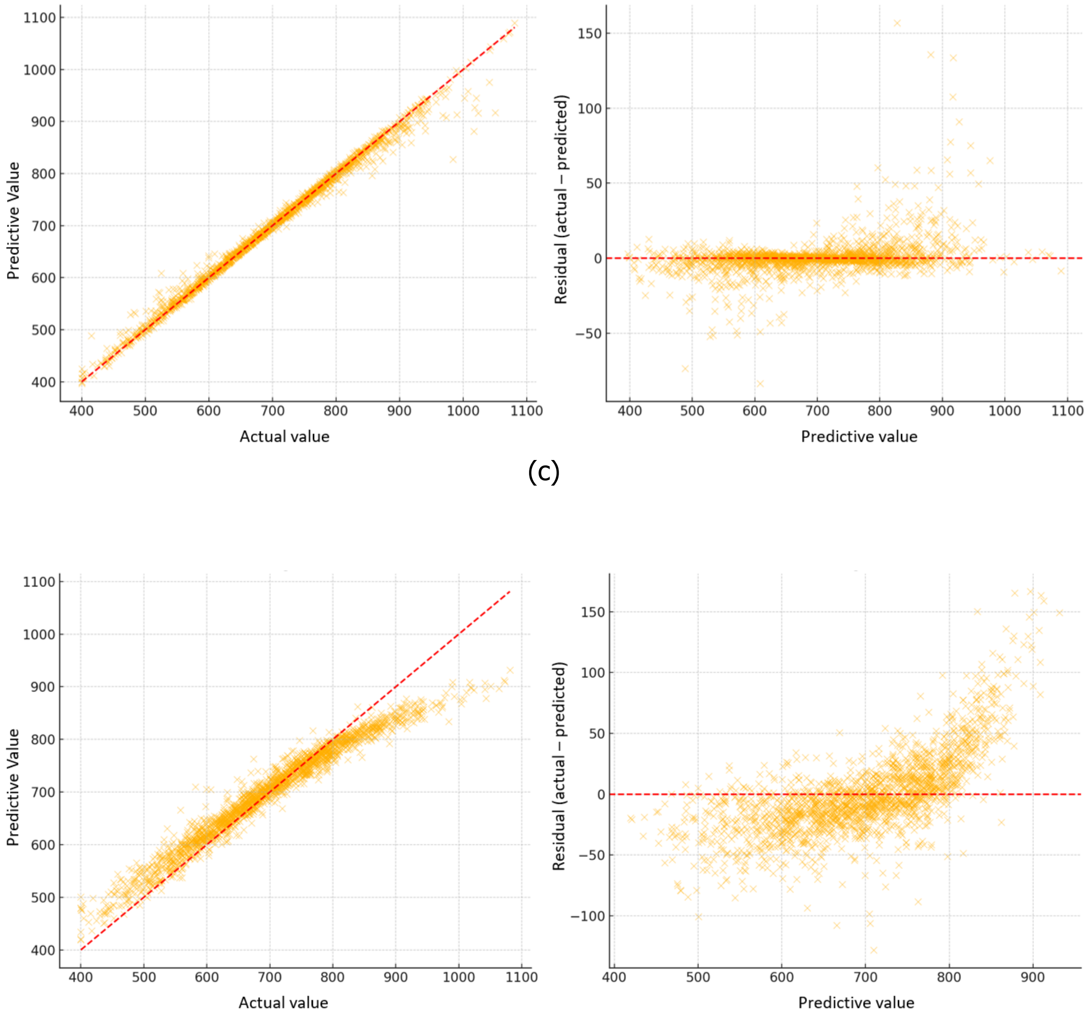
Based on the very high R² value of 0.990 (close to 1), the Random Forest model is proven to be able to explain almost all variations in cassava production output data. This value makes Random Forest one of the best models in this study, especially when considering the balance between predictive accuracy, stability, and interpretation. In addition to the very high R² value (0.99) which shows that the Random Forest model is able to explain almost all variations in the production output data, the MAE and RMSE values also support the accuracy of this model. The MAE of 9.55 indicates that the average prediction error of the model is only about 9.55 output units, while the RMSE of 16.84 indicates a relatively low level of deviation of predictions from actual values. When compared to other models, Random Forest provides the best balance between prediction accuracy and model stability, outperforming SVR (MAE = 5.31; RMSE = 11.75 but with higher computational cost) and Linear Regression (MAE and RMSE are much larger, indicating the model's mismatch with the data pattern).

In addition, the model predictions were compared with the actual data through scatter plot and residual plot visualizations to identify the accuracy and error patterns of the model presented in Figure 4. Model interpretation is complemented by feature importance analysis to identify the input factors that have the most influence on the amount of output, so that they can be used as a basis for decision-making and production optimization strategies.



(a)

(b)



(c)

(d)

Figure 4. Scatter plot and residual plot visualizations to identify the accuracy and error patterns of the model, the left picture shows the regression between prediction and actual, the right picture shows the residual plot (a) Linear regression (b) Random Forest (c) SVR (d) XGBoost.

Feature importance is based on the results of model interpretation based on the Random Forest Regressor algorithm, which is excellent for measuring the relative influence of each feature as well as the correlation to production output. The implementation of the feature importance model and correlation analysis with production output/efficiency is presented in the heatmap in Figure 5. From the data in Figure 5, to increase cassava production output, MSMEs should focus on improving overall production efficiency rather than just increasing the number of workers or speeding up process time.

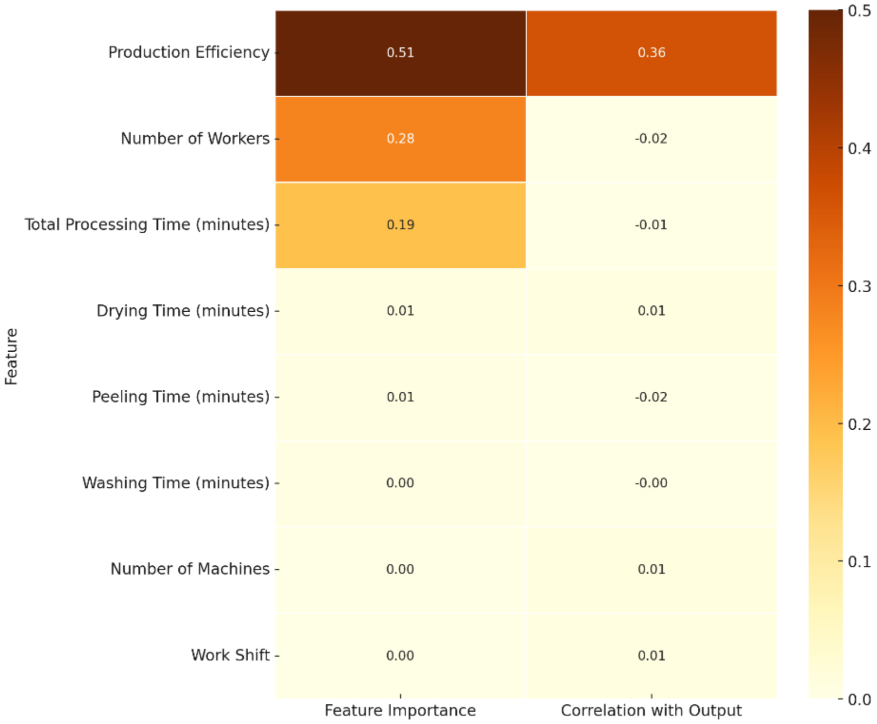


Figure 5. Heatmap of feature importance model and correlation analysis with production output/efficiency

3.3 Sensitivity Analysis

Sensitivity analysis was conducted based on the results of scenario simulation tests based on the Random Forest Regression model. The analysis was conducted by modifying the labor variables and the number of machines. Other scenarios (shift, processing time, etc.) were assumed to be constant at the average value. The results of the sensitivity analysis are presented in Figure 6. From the results, an increase in the number of workers significantly increases production output, an increase in the number of machines also has an effect, although not as great as labor. The combination of optimal labor and sufficient number of machines yields the best output prediction.

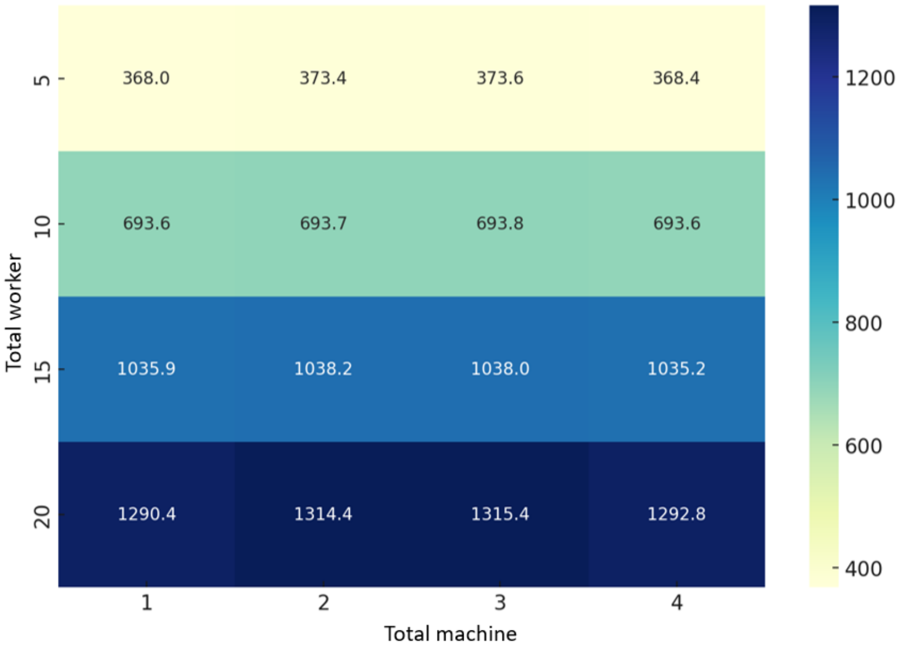


Figure 6. The heatmap results of the sensitivity analysis

Based on the results of the study, the recommended optimal strategy to increase output and efficiency of cassava production is to focus efforts on improving overall production efficiency through more measurable management of labor, process flow, and working time. The optimal number of workers is in the range of 15-20 people per batch, supported by 2-3 units of machinery for maximum results without wasting resources. The application of machine learning-based prediction models, especially Random Forest Regression, can be an accurate tool in operational decision-making. In addition, companies are advised to start implementing simple monitoring systems and data-driven approaches to create a more adaptive, efficient, and sustainable production process.

As a complement to the sensitivity analysis, a simple ANOVA test was conducted on the simulated output results based on variations in the number of workers and machines. The results show that the variation in the number of workers has a statistically significant effect on production output ($p < 0.05$), while the effect of the number of machines is significant only to a certain degree. This suggests that strategic decisions in labor allocation have more impact on increasing output than adding machines, in the context of the existing system.

3.4 Critical Discussion

The implication of this study is the need for a more widespread data-driven approach among MSMEs to support operational decision-making and efficiency. Based on the results of feature importance analysis in the Random Forest model, the features that have the most influence on production output are the number of workers, number of machines, and total processing time. This is consistent with the theory of production in cassava MSMEs, where labor and machine capacity are the main drivers of productivity.

Model evaluation showed that the Random Forest model had an R^2 of 0.990, while the Linear Regression model only reached 0.48. This very high R^2 value can lead to indications of overfitting, but the results of the cross-validation test (5-Fold CV) show that the difference in performance between training data and test data is relatively small, which indicates that the model does not experience significant overfitting. Linear models also tend to be underfitting because they are unable to capture the complexity of relationships between variables.

When compared to the research of Nur et al., 2023, the Random Forest and XGBoost approaches also show high performance in predicting agricultural yields, but this study emphasizes more on simulating combinations of operational variables for sensitivity analysis, which has not been done specifically in the cassava processing MSME sector.

Although the model has high accuracy, there are several limitations, including simulative and historical data from one MSME, so generalization to other industries or regions is still limited. Some important variables such as raw material quality and weather are not included, which could have an impact on prediction accuracy. The model does not test cost efficiency, focusing only on physical output.

4. CONCLUSIONS

Applying several machine learning algorithms, this research successfully built a prediction model for cassava production output with a very high level of accuracy. The Random Forest Regression model showed the best performance with an R^2 value of 0.990, outperforming SVR, Linear Regression, and XGBoost.

Through sensitivity tests and scenario simulations, it was found that the optimal combination of labor (15-20 people) and number of machines (2-3 units) gave the highest predicted output. The addition of work shift variables does not have a significant impact without good efficiency management. Therefore, the recommended approach is to prioritize overall efficiency, not just the addition of resources. By utilizing machine learning as a prediction and decision-making tool, cassava agro-industry companies can optimize resources, increase productivity, and move towards a smarter, data-driven production system.

ACKNOWLEDGEMENT

The authors would like to thank the Research Team and Crew UMKM XYZ for the cooperation and support provided during this research.

REFERENCES

Aminanti Suraya Putri, B., Dian Wisika Prajanti, S., Pujiati, A., 2020. The Effect of Capital, Labor and Raw Materials Toward Production Value (Study on Tapioca Flour Industry in Margoyoso District, Pati Regency). J. Econ. Educ. 9, 143–149.

Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D., 2021. Machine learning in agriculture: A comprehensive updated review. Sensors.

Bhargav, C.S., Nethi, A., Maruti, G., Valluri, S., Sameksha, A., 2024. INTERNATIONAL JOURNAL OF Comparing Data Structures for Efficient Search Algorithms 7. https://doi.org/10.15680/IJMRSET.2024.0712058

BPS kabupaten jember, 2020. Berita Resmi Statistik. Bps.Go.Id.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D. and Hormozdiari, F., 2022. Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(226), pp.1-61.

Elbasi, E., Zaki, C., Topcu, A.E., Abdelbaki, W., Zreikat, A.I., Cina, E., Shdefat, A. and Saker, L., 2023. Crop prediction model using machine learning algorithms. Applied Sciences, 13(16), p.9288.

Erkamim, M., Suswadi, S., Subarkah, M.Z., Widarti, E., 2023. Komparasi Algoritme Random Forest dan XGBoosting dalam Klasifikasi Performa UMKM. J. Sist. Inf. Bisnis 13, 127–134. https://doi.org/10.21456/vol13iss2pp127-134

Hadi, D.K., Putri, R.A., Farida, S.N., Santoso, I., 2021. Application of Cleaner Production in a Fruit Chips Industry. Ind. J. Teknol. dan Manaj. Agroindustri 10, 162–171. https://doi.org/10.21776/ub.industria.2021.010.02.7

Hadi, D.K., Setiawan, A.P., Indrian, O.V., Rosyid, E.F., 2023. Evaluation of Sustainability Supply Chain Performance in the Food Industry: A Case Study. J. Tek. Ind. 24, 95–108. https://doi.org/10.22219/jtiumm.vol24.no2.95-108

Hamidi, M.A., Banowati, E., 2019. Utilization of mocaf flour (modified cassava flour) for revitalization of the use tapioca flour in communities for empowering hulu-hilir human resources in wonogiri regency. IOP Conf. Ser. Earth Environ. Sci. 243. https://doi.org/10.1088/1755-1315/243/1/012081

Kreuzberger, D., Kühl, N., Hirschl, S., 2023. Machine learning operations (mlops): Overview, definition, and architecture. IEEE access.

Kumar, Y., Kaur, K. and Singh, G., 2020, January. Machine learning aspects and its applications towards different research areas. In 2020 International conference on computation, automation and knowledge management (ICCAKM) (pp. 150-156). IEEE.

Kumara, D., Assadam, A., Eka, Z., Renadi, B., 2023. Supplier Selection Based on Green Procurement of Agricultural Commodities of Cassava : Environmental Perspective From Jember Regency 1, 67–75.

Kusunartutik, F., Dwidayati, N.K., 2022. Pemilihan Titik Knot Optimal Menggunakan Metode GCV Dalam Regresi Nonparametrik Spline Truncated. Indones. J. Math. Nat. Sci. 45, 69–76. https://doi.org/10.15294/ijmns.v45i2.39727

Manurung, B.A., Gea, A., Silalahi, A.P., Samosir, N., 2024. Penerapan Algoritma Regresi Linear Untuk Memprediksi Jumlah Wisatawan. J. Ilm. Sist. Inf. 4, 1–9.

Moosavi, S.M., Jablonka, K.M., Smit, B., 2020. The role of machine learning in the understanding and design of materials. J. Am. …. https://doi.org/10.1021/jacs.0c09105

Niekurzak, M., Lewicki, W., Coban, H.H., Bera, M., 2023. A Model to Reduce Machine Changeover Time and Improve Production Efficiency in an Automotive Manufacturing Organisation. Sustain. 15. https://doi.org/10.3390/su151310558

Nur, N., Wajidi, F., Sulfayanti, S., Wildayani, W., 2023. Implementasi Algoritma Random Forest Regression untuk Memprediksi Hasil Panen Padi di Desa Minanga. J. Komput. Terap. 9, 58–64. https://doi.org/10.35143/jkt.v9i1.5917

Rosova, A., Behun, M., Khouri, S., Cehlar, M., Ferencz, V., Sofranko, M., 2022. Case study: the simulation modeling to improve the efficiency and performance of production process. Wirel. Networks 28, 863–872. https://doi.org/10.1007/s11276-020-02341-z

Sepri, D., Fauzi, A., 2020. Prediksi Harga Cabai Merah Menggunakan Support Vector Regression. Comput. Based Inf. Syst. J. 8, 1–5. https://doi.org/10.33884/cbis.v8i2.1921

Shanavas, S., Padmaja, G., Moorthy, S.N., Sajeev, M.S., Sheriff, J.T., 2011. Process optimization for bioethanol production from cassava starch using novel eco-friendly enzymes. Biomass and Bioenergy 35, 901–909. https://doi.org/10.1016/j.biombioe.2010.11.004

Sosa-Perez, V., Palomino-Moya, J., Leon-Chavarri, C., Raymundo-Ibañez, C., Dominguez, F., 2020. Lean Manufacturing Production Management Model focused on Worker Empowerment aimed at increasing Production Efficiency in the textile sector. IOP Conf. Ser. Mater. Sci. Eng. 796. https://doi.org/10.1088/1757-899X/796/1/012024

Subasi, A., 2020. Practical machine learning for data analysis using python. books.google.com.

Syafei, R.M., Efrilianda, D.A., 2023. Machine Learning Model Using Extreme Gradient Boosting (XGBoost) Feature Importance and Light Gradient Boosting Machine (LightGBM) to Improve Accurate Prediction of Bankruptcy. Recursive J. Informatics 1, 64–72. https://doi.org/10.15294/rji.v1i2.71229

Yan, Y. (2022). Machine learning fundamentals. Machine Learning in Chemical Safety and Health: Fundamentals with Applications, 19-46. https://doi.org/10.1002/9781119817512.ch2

Zaki, M.J., Jr, W.M., Meira, W., 2020. Data mining and machine learning: Fundamental concepts and algorithms. books.google.com.

Zhang, X., Tian, Y., Chen, L., Hu, X. and Zhou, Z., 2022. Machine learning: a new paradigm in computational electrocatalysis. The Journal of Physical Chemistry Letters, 13(34), pp.7920-7930. https://doi.org/10.1021/acs.jpclett.2c01710