# Evaluation of machine learning models based on household food insecurity data in Indonesia

*Herlin* Fransiska[1,2], *Agus Mohamad* Soleh[1*], *Khairil Anwar* Notodiputro[1], and Erfiani[1]

[1]School of Data Science, Mathematics, and Informatics, IPB University, Meranti Wing Street, Dramaga Campus IPB, Bogor, West Java, 16680, Indonesia
[2]Faculty of Mathematics and Natural Sciences, Bengkulu University, Kandang Limun Street, Bengkulu, 38371, Indonesia

**Abstract.** Household food insecurity is a critical issue, and accurate prediction models are essential for identifying at-risk households and guiding policy decisions to address this issue. This study compared the effectiveness and stability of two machine learning models: random forests (RF) and generalized random forests (GRF). Predicting household food insecurity using food insecurity experience scale data in West Java, Indonesia. The evaluation showed that the GRF model performed best and exhibited more consistent predictions. The important variables that influence household food insecurity in West Java are household size, type of house floor, bank savings account ownership, type of house wall, sanitation facility adequacy status, cash transfer program status, land ownership status, and food assistance recipient status.

## 1 Introduction

Sustainable Development Goal (SDG) 2.1 aims to end hunger by 2030. Household food insecurity is a serious problem related to the prosperity, health, and socioeconomic stability of a society. Food insecurity occurs when households lack adequate or stable access to sufficient and nutritious food and can have far-reaching consequences [1]. Therefore, effective research and intervention on this issue are of paramount importance. Predicting and understanding the factors influencing household food security using machine learning models can assist policymakers in designing more effective programs to address food insecurity and enhance food security.

Ensemble learning is a machine learning technique that aims to combine several weak models into a stronger and more accurate model. One of the most popular ensemble learning methods is Random Forest (RF) [2]. Random forests operate by constructing multiple decision trees, each of which contributes to the final prediction. Traditional random forests have become popular in many disciplines and have shown good results when used to solve traditional regression or classification problems [3]. Although RF is effective in many applications, it has limitations in terms of model flexibility and the ability to capture heterogeneity in data. Random Forest assumes that each tree operates independently and

* Corresponding author: agusms@apps.ipb.ac.id

identically, which can be challenging when complex relationships and interrelated variables are present in the data [4].

The Generalized Random Forest (GRF) was developed as an extension of the RF. The strength of the GRF in capturing complexity and variation in data makes it a highly powerful tool for data analysis, particularly in situations where complex data and high heterogeneity present challenges. The GRF can account for variations among different trees and accommodate more complex relationships within the data, enabling a more accurate model to capture existing heterogeneity [4]. The Generalized Random Forest (GRF), a popular nonparametric approach, is a flexible method that accommodates various objectives [4,5]. Additionally, the GRF produces accurate and consistent predictions, as well as variance estimates with valid confidence intervals and is robust. The GRF is also computationally efficient [4, 6].

The primary focus of this study was to predict household food insecurity in West Java, Indonesia, using the GRF method. This study aimed to identify the factors that determine food insecurity and provide accurate and reliable predictions through the implementation of machine learning. By utilizing the GRF model, this study seeks to deepen the understanding of the determinants of food insecurity and assist policymakers in designing more effective interventions to address food insecurity in West Java, Indonesia.

## 2 Materials and methods

This section explains some of the details of the methods and materials used. About random forest, generalized random forest, evaluation, and food insecurity.

### 2.1 Random forest

Random forest is an ensemble learning technique that uses multiple decision trees to make predictions (either classification or regression) [2]. Decision trees serve as core components in the training process. In each iteration, the random forest randomly selects a subset of features from the data to build each tree, thereby generating diverse trees [7]. Once all trees have completed the classification or regression, the random forest determines the result based on the majority vote for classification or the average for regression [2, 7-9]. This makes the random forest a robust technique that is commonly applied in various machine learning applications. The random forest algorithm has two stages: training and testing. In the training stage, the process involves bootstrap sampling and building decision trees, where the features used are randomly selected or extracted from the data. Once the trees are built, the next step is testing. In this stage, predictions were made by all trees. The final decision is made by applying a majority vote or averaging the predictions of all trees [2, 9].

### 2.2 Generalized random forest

The Generalized Random Forest (GRF) is a non-parametric statistical technique developed as an extension of the traditional random forest. The primary distinction of the GRF lies in its ability to estimate outcomes that depend not only on simple input variables but also on more complex structural relationships. This is made possible by the application of local moment equations. Local moment equations are used to identify the relationships between variables within a statistical model [4–6]. In the GRF, the data are partitioned into small sections, and the estimation is performed locally within each section. This approach allows the GRF to capture local variations in the data that may be overlooked by traditional random forests. The GRF provides more flexibility in handling more complicated statistical problems

owing to this capacity. With this capability, the GRF offers greater flexibility in addressing more complex statistical problems. Additionally, the GRF can handle data heterogeneity and produce more accurate estimates at the local level, making it a highly useful method for various statistical applications [4, 10]. The Generalized Random Forest (GRF) consists of two stages. The main algorithm of GRF describes the main structure of GRF, in which multiple trees are constructed using subsampled data, and the results are combined using weighted-averaging. In general, the algorithm uses the concepts of honesty and sub-sampling to improve accuracy and reduce bias. To improve estimation precision, the prediction results employ a neighborhood-based weighting technique rather than being based only on majority vote or average, as in conventional RF [4]. Each individual tree in the forest is grown using gradient-based methods to determine the best split during the tree-building process.

### 2.3 Food insecurity

Food insecurity is a situation in which individuals or households experience limitations in the quantity and quality of food required for a healthy life due to factors such as economic instability and limited access to food resources [11, 12]. Indonesia, like all other nations, views food insecurity as a serious issue that must be addressed. Finding the indicators of household food insecurity is crucial for the government to effectively prepare appropriate measures. It is possible to determine the characteristics of a household experiencing food insecurity by identifying the indicators of these households [13, 14]. In 2013, the Food and Agriculture Organization (FAO) identified a lack of effective data to measure the level of food insecurity. The FAO then developed the Food Insecurity Experience Scale (FIES). This instrument measures the experiences of individuals or households related to food insecurity. In addition to aiding data collection, the FIES supports various aspects of food security governance. For instance, the FIES facilitates better planning and decision-making, enhances transparency in institutions related to food distribution, promotes more equitable resource allocation, and assists in the development of more coordinated and coherent policies [7, 11, 4]. The Food Insecurity Experience Scale (FIES) consists of eight questions related to household access to sufficient food, with responses of "yes" or "no." In Indonesia, the FIES survey is conducted by the Central Statistics Agency (BPS) through the National Socioeconomic Survey [7].

## 3 Result and discussion

This study analyzed the status of household food insecurity in West Java, Indonesia, in 2021. A household is categorized as food insecure if there is a "yes" response to any of the eight questions in the Food Insecurity Experience Scale (FIES) survey. Data source: National Socio-Economic Survey (2021). Consists of 1 response variable, namely the classification of household food insecurity experience (Y) and 23 predictor variables

X1: Number of household members
X2: Gender of head of household
X3: Age of head of household
X4: Illiterate status
X5: The highest education of the head of the household
X6: Number of bank savings account ownership
X7: Health insurance contribution assistance recipient status
X8: Ownership of health insurance
X9: Smoker status of head of household
X10: Home ownership status

X11: House size
X12: Type of house wall
X13: Type of house floor
X14: Adequacy of home sanitation
X15: Feasibility of drinking water sources
X16: People's business credit status
X17: Bank/cooperative loan status
X18: Village-Owned Enterprises acceptance status
X19: House/land assets
X20: Prosperous Family Card recipient status
X21: Family Hope Program recipient status
X22: Non-Cash Food Assistance recipient status
X23: Status of recipients receiving other routine assistance.

The data consisted of 25,890 observations (households), but 25,873 were used in this study. This is because of the data cleaning/deletion process of household data that answered do not know or refused to answer the question.

In the data exploration stage, the distribution of the response variable was visualized in the form of a pie chart (Figure 1). with two categories: 0 and 1. Category 0 represents 77% of the total data, indicating households that do not experience food insecurity, and category 1 represents 23%, indicating households that experience food insecurity. This distribution provides an initial overview of the proportion of households based on their food insecurity status.

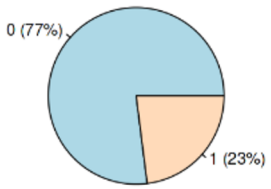


**Fig. 1.** Proportion of food insecurity household status.

The predictor variables used in this study consisted of two types: categorical and numerical. There are two variables whose diversity is very small, so they are not included in the analysis, namely X4 and X18.

In the modeling, the data were divided into 70% for training and 30% for testing. The modeling was conducted on the training data using two machine learning algorithms: random forest and generalized random forest. Then, the predictions of the testing data were made. The model evaluation was then calculated based on the testing data using metrics such as accuracy, sensitivity, specificity, and balanced accuracy with 100 iterations. The performance of the models generated by both algorithms was evaluated using boxplot visualization and difference testing. The boxplot results are shown in Figure 2.

.

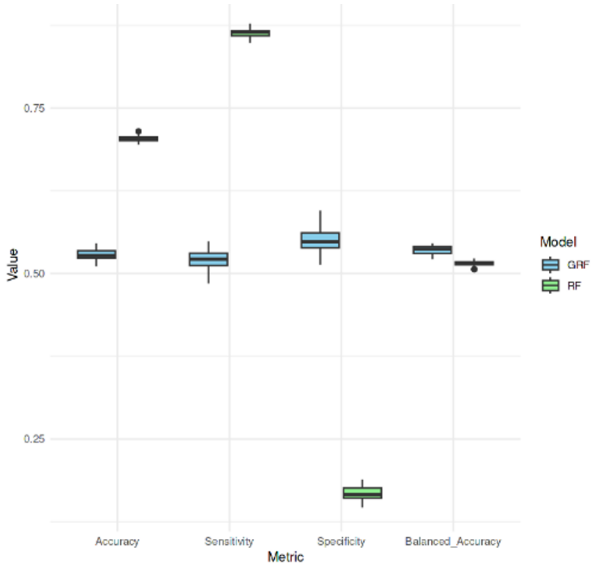


**Fig. 2.** Metric evaluation of RF dan GRF models.

**Table 2.** Testing of metric evaluation

| Evaluation | p-value | Description |
|---|---|---|
| Accuracy | <2.2e-16 | RF tends to have higher accuracy compared to GRF. |
| Sensitivity | < 2.2e-16 | RF tends to have higher sensitivity compared to GRF. |
| Specificity | < 2.2e-16 | GRF tends to have very higher specificity compared to RF. |
| Balanced accuracy | < 2.2e-16 | There is a significant difference in balanced accuracy between RF and GRF, with GRF showing better performance in balancing between sensitivity and specificity. |

The Generalized Random Forest (GRF) model was chosen because it demonstrated better performance in terms of specificity and balanced accuracy compared to Random Forest (RF). Given the importance of minimizing errors in detecting food insecurity cases, GRF's ability to balance sensitivity and specificity makes it the more suitable choice. Additionally, GRF is more effective in handling imbalanced data, which is often encountered in food insecurity datasets. With better performance in identifying both positive and negative cases, as well as greater model stability, GRF is the more optimal choice compared to RF for this study.

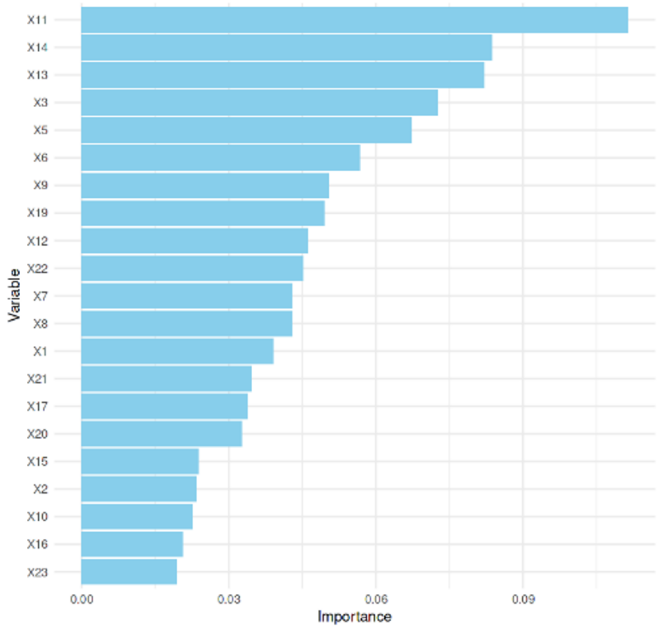


**Fig. 3.** Variable importance plot based on GRF model

The variable importance in the generalized random forest (GRF) model can be seen in Figure 3. Variables importance are house size (X11), adequacy of home sanitation (X14), type of house floor (X13), age of head of household (X3), the highest education of the head of household (X5), and number of bank savings account ownership (X6).

Some recommendations based on these results are that policies could focus on affordable housing initiatives, subsidies for home expansion, or improved access to housing loans. Expanding sanitation infrastructure and promoting education about hygiene and sanitation. Support low-cost, durable flooring materials for low-income households. Younger heads of households may need financial literacy training or job creation programs. Expanding access to vocational training for heads of households; ensuring higher levels of education for future generations through improved schooling systems. Promote awareness of banking services in underserved areas, support microfinance or savings groups for rural or low-income households, and encourage programs that educate households about the benefits of savings and investments.

## 4 Conclusions

This study demonstrated the effectiveness of the Generalized Random Forest (GRF) model in predicting household food insecurity in West Java, Indonesia. When compared to the traditional Random Forest (RF) model, GRF showed strong performance in terms of specificity and balanced accuracy, making it a more reliable tool for identifying households experiencing food insecurity. Importance predictors of food insecurity are house size (X11), adequacy of home sanitation (X14), type of house floor (X13), age of head of household (X3), the highest education of the head of household (X5), and number of bank savings account ownership (X6). The model's ability to highlight key predictors provides valuable insights into the structural and socioeconomic factors that contribute to food insecurity, thereby supporting the development of more precise and impactful policy measures.

## Acknowledgment

We would like to express our sincere gratitude to the Directorate General of Higher Education, Research, and Technology of The Ministry of Education, Culture, Research, and Technology for funding this research through the 2024 Doctoral Research Scheme in accordance with Research Contract Number: 027/E5/PG.02.00.PL/2024 dated June 11, 2024.

## References

1. C. A. Myers, Food insecurity and psychological distress: a review of the recent literature. Curr. Nutr. Rep. **9** (2), 107–118 (2020). https://doi.org/10.1007/s13668-020-00309-1
2. L. Breiman, Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
3. Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, X. Liang, An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Syst. Appl. **237**, 121549 (2024). https://doi.org/10.1016/j.eswa.2023.121549
4. S. Athey, J. Tibshirani, S. Wager, Generalized random forests. Ann. Stat. **47** (2), 1179–1203 (2019). https://doi.org/10.48550/arXiv.1610.01271
5. A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, J. Gama, Methods and tools for causal discovery and causal inference. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **12** (2), 1449 (2022). https://doi.org/10.1002/widm.1449
6. L. Lei and E. J. Candès, Conformal inference of counterfactuals and individual treatment effects. J. R. Stat. Soc. Ser. B Stat. Methodol. **83** (5), 911–938 (2021). https://doi.org/10.1111/rssb.12445
7. Rais, A. Mohamad Soleh, B. Susetyo, Rotation double random forest algorithm to predict the food insecurity status of households. J. RESTI. **8** (1), 33–41 (2024). https://doi.org/10.29207/resti.v8i1.5540
8. G. Shanmugasundar, M. Vanitha, R. Čep, V. Kumar, K. Kalita, and M. Ramachandran, A Comparative Study of Linear, Random Forest and AdaBoost Regressions for Modeling Non-Traditional Machining. Processes, **9** (11), 2015 (2021). https://doi.org/10.3390/pr9112015
9. S. Han, H. Kim, and Y. S. Lee, Double random forest. Mach. Learn. **109** (8), 1569–1586 (2020). https://doi.org/10.1007/s10994-020-05889-1
10. E. Zhou and D. Lee, Generative artificial intelligence, human creativity, and art. PNAS Nexus. **3** (3), 1–8 (2024). https://doi.org/10.1093/pnasnexus/pgae052
11. A. Tharwat, Classification assessment methods. Appl. Comput. Inform. **17** (1), 168–192 (2021). https://doi.org/10.1016/j.aci.2018.08.003
12. V. M. Hazzard, K. A. Loth, L. Hooper, and C. B. Becker, Food insecurity and eating disorders: a review of emerging evidence. Curr. Psychiatry Rep. **22** (12), 1–9 (2020). https://doi.org/10.1007/s11920-020-01200-0
13. M. Nidia, B. Sartono, Indahwati, A.F.Hadi, and E. Ramadhani, A study in determining indicators of food-insecure households using SHAP and Boruta SHAP, In AIP Conf. Proc. **2720**, 020011 (2023). https://doi.org/10.1063/5.0137150
14. R. Evi, B. Sartono, A. F. Hadi, W. D. Safitri, and S. Ufa, Study on identification of main indicators of food insecurity households in Aceh Province in 2019-2020 using classification tree, In AIP Conf. Proc. **2556**, 050002 (2023). https://doi.org/10.1063/5.0111055