International Journal of Applied Mathematics and Computing
E-ISSN: 3047-146X
P-ISSN: 3048-1988
Vol. 3 No. 2 April 2026

z (Research/Review) Article

Identification of Housing Eligibility Status Using Family Data in Samarinda City

Antonieta Aryuka Paskalia Nggotu[1], Hamdani Hamdani[2], Anindita Septiarini [3]

[1]Department of Informatics, Faculty of Engineering, Mulawarman University, Indonesia.
[2]Department of Informatics, Faculty of Engineering, Mulawarman University, Indonesia.
[3]Department of Informatics, Faculty of Engineering, Mulawarman University, Indonesia.
*Corresponding Author: antonietaaryuka@gmail.com

Abstract: The issue of uninhabitable houses still requires an accurate identification mechanism because the manual data collection process has the potential to be time-consuming, costly, and subject to subjectivity in determining aid priorities. This study aims to develop a classification model to identify habitable and uninhabitable houses based on family socioeconomic data using the Random Forest algorithm. The research method includes data preprocessing, data division using stratified split in three scenarios, baseline model development, and optimization through hyperparameter tuning using GridSearchCV with 3-fold cross-validation and balanced class_weight parameters. The data used includes variables such as education type, employment status, occupation type, number of family members, and family insurance type. The test results show that the 70:30 data division scenario after tuning provides the best performance with a recall value of 0.5797 for uninhabitable houses and an F1-score of 0.4746. Feature importance analysis shows that education type and employment status are the most influential variables in the classification. The results of this study show that the model built is capable of increasing sensitivity in detecting uninhabitable houses to support more objective field survey prioritization.

Keywords: Uninhabitable Housing; Random Forest; Classification; Hyperparameter Tuning; GridSearchCV.

1. Introduction

The provision of standardized and adequate housing serves as a pivotal foundation for public health and a critical social determinant of human well-being (Summer et al., 2026). Homes that do not meet building safety standards, adequate space, and environmental quality have the potential to impact the health and productivity of their occupants. Standardized housing represents more than just a physical building, acting as a crucial social environment that promotes community cohesion (Elsayed, 2025). Consequently, the quality of these housing conditions is heavily influenced by various socioeconomic factors, including the occupants' educational background, occupational status, and the number of dependents (Kayode et al., 2021). In reality, achieving livable housing is not easy for every family due to socioeconomic and poverty barriers (Robiah et al., 2024). With national figures showing that around 36.85% of households in Indonesia that still live in unfit housing (Perkim.id, 2024), this study focuses on the classification of housing suitability in the city of Samarinda by utilizing socioeconomic variables from the National Population and Family Planning Agency (BKKBN) dataset.

Advances in machine learning enable classification processes to be carried out more objectively and based on data in various sectors (Dritsas & Trigka, 2025). In the field of urban studies, both rule-based and data-driven approaches have been utilized extensively for building and housing classification. While traditional rule-based methods rely on rigid

Received: March, 6th 2026
Revised: March, 13th 2026
Accepted: March, 26th 2026
Published: April, 07th 2026
Curr. Ver.: April, 07th 2026

administrative criteria, data-driven approaches offer the flexibility to capture complex socioeconomic patterns (Kim et al., 2022). Machine learning itself uses various types of algorithm models that are capable of recognizing patterns in large data sets to produce accurate predictions. The use of machine learning algorithms has proven to be superior in handling socioeconomic data that has high dimensions and non-linear relationships between variables compared to traditional statistical approaches(Ca & Jobarteh, 2024). In the context of housing eligibility, various studies have been conducted to develop identification methods that are more efficient than manual data collection processes, which tend to be time-consuming and costly (Alsharkawi et al., 2021).

Various studies have demonstrated the superiority of ensemble methods in the social domain. A study in Greece found that Random Forest was the most successful classifier for predicting energy poverty with an accuracy of up to 94% (Kalfountzou et al., 2025). Similarly, research related to sustainable development goals (SDGs) confirmed that Random Forest provided the best results in classifying household economic status compared to SVM and ANN models (Sholihah & Hermawan, 2023). In the context of social vulnerability in Istanbul, the use of ANN with a subsampling strategy proved effective in distinguishing households vulnerable to disasters (Kalaycloğlu et al., 2023). In Indonesia, research in Yogyakarta highlights that physical features such as building area are strong predictors in machine learning-based models, but recommends adding socioeconomic variables for more relevant accuracy (Sudarawerti & Arif, 2025). Another study at the neighborhood level using Decision Tree shows that family size and occupation are dominant factors, but suggests using more valid classification methods to improve accuracy (Nengsi et al., 2025).

The main problem in this study is the gap in accuracy and efficiency in identifying uninhabitable houses if we only rely on physical variables that require intensive field surveys. A classification model is needed that is capable of accurately identifying habitability using only socioeconomic indicators. Most previous studies still focus on physical features of houses or model comparisons without performing optimization tuning. To overcome this problem, this study proposes a classification model using the Random Forest algorithm with hyperparameter tuning through GridSearchCV. Recent research has confirmed that Random Forest delivers superior predictive performance in housing studies, significantly outperforming other models such as Support Vector Machines (SVM) and traditional hedonic regressions (Zhang et al., 2026). This approach was chosen because Random Forest has proven to be superior in the social domain compared to SVM and ANN. The integration of GridSearchCV aims to improve the sensitivity of the model in detecting the minority class of uninhabitable houses, so that it can be an initial tool for the government in determining field survey priorities objectively. Similar data-driven approaches have been adopted in recent studies that use Random Forest models on administrative and survey data to support the targeting of vulnerable households in poverty and energy-poverty programs (Browne et al., 2021). The main contribution of this research is to provide a housing eligibility classification model that focuses on socioeconomic variables as an alternative to physical data collection and to implement systematic optimization through GridSearchCV on the Random Forest model for socioeconomic data cases in Samarinda.

2. Literature Review

2.1. Random Forest

Random Forest is one of the most widely used machine learning methods due to its ability to produce stable and accurate classification models (Salman et al., 2024). This algorithm works by randomly constructing a collection of decision trees to form a digital "forest." The more trees generated, the higher the prediction accuracy obtained. Unlike a single decision tree that relies on simple information gain, this ensemble method offers superior stability and accuracy in handling complex datasets (Ahmad et al., 2023). Although it operates as an ensemble, the fundamental mechanism of each constituent tree is rooted in information theory, specifically through the calculation of Entropy to determine the most informative splits within the forest's nodes (Kinasih et al., 2024). In the process of forming these individual decision trees, the data at the initial node is grouped based on the overall class label, then the information value is calculated using all available data. This calculation is used as the basis for determining the best attribute split, according to the formula shown in Eq. (1).

$$Info(D) = -\sum_{i=1}^{m} pi\ log^{2(pi)} \ldots\ldots (1)$$

Where the entropy value is calculated based on the probability of each class appearing in the dataset. Next, the partition value (Information Gain) is calculated using the formula shown in Eq. (2).

$$InfoA(D) \sum_{j}^{v} \frac{|Dj|}{|D|} x\ InfoA(Dj) \ldots\ldots (2)$$

The information value of each attribute is calculated to determine the quality of data separation, where the value |Dj|/|D| indicates the weight of each partition. The smaller the InfoA(D) value, the better the quality of the resulting partition. For continuous attributes, the separation is determined by finding the best split point based on the smallest information value. The information gain value will be calculated using Eq. (3), and the attribute with the highest gain value is selected as the decision tree branch. By aggregating thousands of these Entropy-based trees, Random Forest effectively reduces the variance typically found in a single decision tree.

$$Gain(A) = Info(D) - InfoA(D) \ldots.. (3)$$

The decision tree formation process is carried out repeatedly until it reaches the branching limit or all data has the same label, thus forming leaves with the majority value of the data (Oktafiani et al., 2024). The process in random forest for feature importance is as follows: The main advantage of the Random Forest algorithm lies in its ability to generate information about the contribution level of each feature to the model classification results. First, classification and regression trees are built to generate out-of-bag (OOB) sample data. Based on the OOB data, random forest can verify the importance of the input data and obtain an importance score for each feature, which is expressed through mean decrease accuracy (MDA) (Zhao et al., 2022).

2.2 Confusion Matrix

Confusion Matrix is a very important evaluation method in classification model performance analysis because it provides comprehensive visual results comparing model predictions and actual labels. Through the components of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), the confusion matrix enables the calculation of evaluation metrics such as accuracy, precision, recall, and F1-score, which provide an overview of the reliability of the model in performing classification (Sathyanarayanan & Tantri, 2024). The equations for each metric are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} x\ 100 \quad (4)$$

$$\text{Precison} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{F1-Score} = \frac{2 \text{ x Recall x Precision}}{\text{Recall + Precision}} \quad (7)$$

3. Materials and Method

This research was conducted through systematic stages, starting from data acquisition to model evaluation, as presented in the flowchart in Fig. 1. The initial stage involved Data Acquisition, where the dataset containing family socioeconomic variables was loaded as independent variables and housing eligibility status as dependent variables. Next, Data Preprocessing was carried out, which included handling missing values and analyzing class

distributions to identify potential data imbalances. The dataset was then divided into training and testing data using the Stratified Data Splitting technique to keep the proportions of each class consistent. The core process of this research involves Model Training using the Random Forest algorithm, which is further optimized through Hyperparameter Tuning with the GridSearchCV method to obtain the most effective model parameters. The final stage is Model Evaluation, which utilizes classification performance metrics to assess the accuracy and sensitivity of the model in identifying livable and uninhabitable houses with precision.

Figure 1. Research Flowchart

3.1. Data Acquisition

The initial stage of this research is the acquisition of a dataset derived from the 2024 Family Data Collection in Samarinda City, with a total of 180,848 family data. This dataset provides information on the socioeconomic conditions of families as predictor attributes and housing eligibility variables as target variables for training machine learning models. Data acquisition aims to obtain valid and relevant socioeconomic information on families as a basis for developing a housing eligibility classification model. Although housing eligibility status in the BKKBN dataset is determined based on physical building indicators, this study utilizes socioeconomic variables as predictors to examine their relationship with housing eligibility status. This dataset represents real conditions in the field, including subdistrict variables that are not used as modeling attributes but are used for the purpose of visualizing the distribution of uninhabitable houses per subdistrict. Some examples of raw data are shown in Table 1.

Table 1. Research Dataset

Number of People	Occupation Type	Employment Status	Insurance Type	Education Level	Housing Eligibility	Sub-district
3	9	4	1	8	1	Samarinda Ilir
4	11	5	1	8	1	Samarinda Ilir
5	11	5	4	8	0	Sungai Pinang
...	...	...	...	...	...	...
3	6	4	2	10	1	Sungai Kunjang
5	2	5	2	4	0	Sambutan

Furthermore, a detailed explanation of each variable used in the classification process, including details of attribute names, data types, and category descriptions for both input (independent) features and target (dependent) variables, is presented in Table 2. This data structure is the main reference in the preprocessing and modeling stages using the Random Forest algorithm. After the data was received, the dataset was prepared for the next stages, namely data preprocessing, modeling using the Random Forest algorithm, and evaluation of the classification model's performance.

Table 2. Dataset Structure and Feature Descriptions

Variable	Attribute Name	Data Type	Description
Number of People	jml_jiwa	Numeric	Total number of family members in a single household

Occupation Type	jns_pekerjaan	Numerik (Categorical)	1 = Unemployed/Not yet working 2 = Farmer 3 = Fisherman 4 = Trader 5 = State official/Regional head 6 = Civil Servant/Military/Police (PNS/TNI/POLRI) 7 = Private sector (Agriculture) 8 = Private sector (Industry) 9 = Private sector (Services) 10 = Retiree 11 = Freelancer 12 = Private employee (PK21) 13 = Entrepreneur (PK21)
Insurance Type	jns_asuransi	Numerik (Categorical)	1 = BPJS-PBI/Government-funded insurance 2 = BPJS Non-PBI/Paid insurance 3 = Private insurance 4 = None
Employment Status	status_pekerjaan	Numerik (Categorical)	1. Self-employed 2 Self-employed assisted by temporary/unpaid workers 3 = Self-employed assisted by permanent/paid workers 4 = Employee/Laborer 5 = Casual worker 6 = Family worker/Unpaid worker
Education Level	jns_pendidikan	Numerik (Categorical)	1 = No formal education 2 = Did not complete Primary School 3 = Currently in Primary School 4 = Completed Primary School 5 = Currently in Junior High School 6 = Completed Junior High School 7 = Currently in Senior High School 8 = Completed Senior High School 9 = Currently in University/Academy 10 = Completed University/Academy
Housing Eligibility	rumah_layak	Binary	0 = Ineligible 1 = Eligible

3.2. Data Preprocessing

This stage aims to convert raw data into data that is ready for analysis by correcting various problems such as missing values, noise, redundancy, and data imbalance(Shahidi et al., 2025). The preprocessing stage also ensures that the algorithm works on clean and structured data so that it can reduce bias and improve the consistency and reliability of the analysis results. With the increasing volume and variety of data, proper preprocessing has become an important factor in maintaining the effectiveness of analytical pipelines in various fields, including the government sector (Koukaras & Tjortjis, 2025). In this study, data preprocessing was carried out through three main stages. First, missing values were checked for each variable, then empty values were handled using the imputation method with the median for numerical data and the mode for categorical data to keep the data distribution stable. Second, a class distribution analysis was performed on the target variable rumah_layak to identify possible class imbalances. Third, the processed data was ensured to be clean and ready for use in the modeling stage.

3.3. Data Splitting

Data splitting is a crucial step in developing Machine Learning (ML)-based models, where choosing the right splitting strategy greatly determines the effectiveness of the model in accordance with specific research objectives (Aouichaoui et al., 2025). Data that has undergone preprocessing is then divided into training data and test data using the stratified

split technique. This data splitting is a technique for partitioning datasets, which is one of the factors that determines the performance of classification models in machine learning algorithms. The dataset is divided into three data splitting scenarios, namely 70:30, 80:20, and 90:10, to maintain the proportion of classes in each dataset. This technique aims to maintain the proportion of livable and unlivable houses in each dataset. At this stage, an analysis of the unbalanced data was also performed using bar chart visualization.

3.4. Model Training & Tuning

The training data was used to build a classification model using the Random Forest algorithm. Given the imbalance in the distribution of classes between livable and unlivable houses, the class_weight balanced parameter was applied during the training stage to give greater weight to the minority class (unlivable houses). This approach aims to make the model more sensitive in detecting unlivable houses and reducing classification errors in that class. This study uses the Random Forest algorithm to examine the relationship between family socioeconomic variables and the status of housing suitability. The process of forming decision trees in Random Forest is done by selecting the best attributes based on entropy and information gain calculations as explained in equations (1), (2), and (3). Random Forest was chosen because it is capable of handling nonlinear relationships between variables, works well with high-dimensional data, and has a relatively low level of overfitting. In addition, this algorithm provides feature importance information that can be used to analyze the most influential factors in determining housing eligibility. Furthermore, to overcome the imbalance in class distribution between eligible and ineligible houses, the class_weight balanced parameter was applied during the model training process so that the minority class received a greater weight. This approach aims to improve the sensitivity of the model in identifying uninhabitable houses. Next, hyperparameter tuning was performed using the GridSearchCV method with 3-fold cross-validation to obtain the best parameter combination. The selection of the best parameters was based on the F1-score value in the uninhabitable house class so that the model had a balance between precision and recall. The model with the best parameters was then used in the evaluation stage using test data.

3.5. Model Evaluation

The trained model was evaluated using test data with confusion matrix and classification report metrics. Feature importance analysis was then conducted. In addition, the model's classification results were visualized in the form of infographics showing the average probability of uninhabitable houses in each subdistrict in Samarinda City to indicate the level of housing suitability risk in each area based on the model's predictions.

4. Results and Discussion

This study successfully developed a housing suitability classification model using the random forest algorithm based on socioeconomic variables from the 2024 Samarinda City BKKBN Family Data Collection. The model was used to classify habitable and uninhabitable houses with a focus on the model's ability to detect uninhabitable houses as a minority class. The preprocessing results showed that most variables did not have missing values, except for the status_pekerjaan variable with a missing value of 36.737, which was then handled through an imputation process. The target class distribution showed that habitable houses were more dominant than uninhabitable houses. The class distribution visualization is shown in Fig 2. The target class distribution shows that livable houses account for 70.6% while uninhabitable houses account for 29.4%, so the dataset is classified as unbalanced, requiring optimization using hyperparameter tuning.

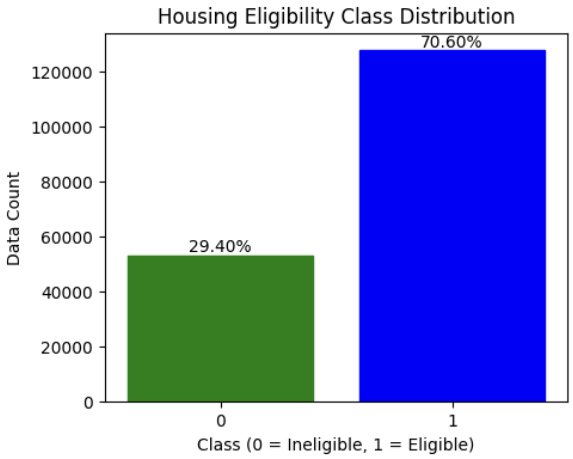

Figure 2. Class Distribution

4.1. Baseline Model Testing

Model performance was evaluated at the baseline stage and after the hyperparameter tuning process using three data split scenarios, namely 70:30, 80:20, and 90:10. Each scenario was evaluated using accuracy, precision, recall, and F1-score metrics with a focus on the uninhabitable house class (class 0). The scenario with the highest F1-score value in class 0 was selected as the best scenario for the advanced modeling stage. The initial random forest model was constructed using n_estimators parameters of 200 and class_weight = balanced and random_state with a value of 42. At the baseline stage, model evaluation can be seen in Table 3, where the 90:10 scenario produced the best performance.

Table 3. Evaluation of baseline scenario class 0 model

Skenario	Akurasi	Precision (0)	Recall (0)	F1-Score (0)
70 : 30	0,6261	0,4001	0,5437	0,4610
80 : 20	0,6228	0,3978	0,5512	0,4621
90 : 10	0,6221	0,3989	0,5627	0,4668

This study focuses on the model's ability to detect uninhabitable houses and minimize false negatives, so the 90:10 scenario was chosen as the best scenario before conducting optimization experiments such as tuning for each scenario. A visualization of the confusion matrix is also presented in Fig. 3.

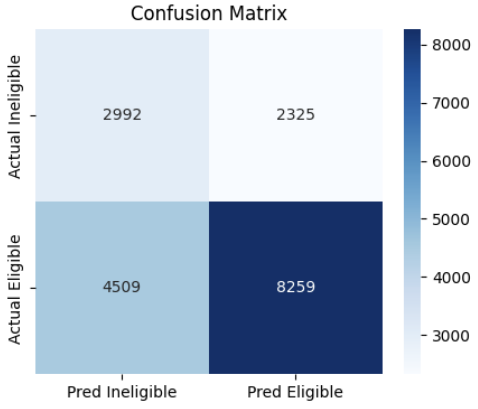

Figure 3. Confusion matrix baseline model

The baseline model evaluation was conducted using a 90:10 data split ratio, which resulted in a TP value of 2,992, FN of 2,325, FP of 4,509, and TN of 8,259. These results, based on the 10% testing subset, indicate that the model still produces a relatively higher FP value than TP. This disparity suggests the need for hyperparameter tuning to improve the balance between precision and recall, as well as testing across different split scenarios to ensure model stability. In addition to the confusion matrix, this study also presents the results of feature importance analysis, which shows the level of contribution of each variable in the process of

classifying houses as livable and unlivable. The importance value represents the magnitude of the influence of each feature in forming the decision tree in the model, as shown in Fig. 4.

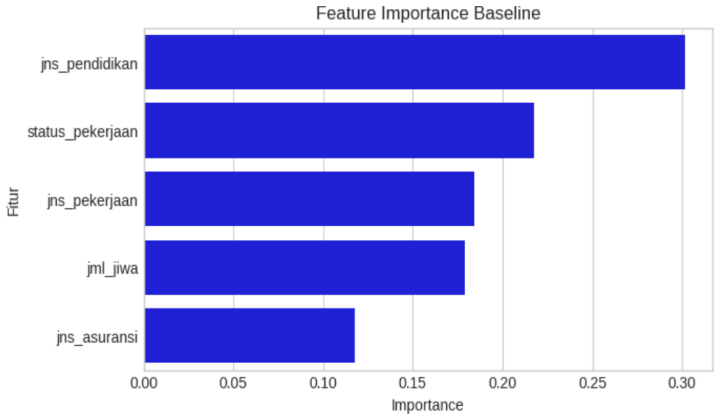

Figure 4. Feature importance baseline model

4.1. Hyperparameter Tuning Model Testing

After the best temporary scenario was obtained, the hyperparameter tuning process was carried out using the GridSearchCV method with 3-fold cross-validation. Previously, the best temporary scenario was obtained based on the baseline model, but all scenarios still went through the tuning process with the parameters tested including n_estimators (100, 200), max_depth (None, 10, 20), min_samples_split (2, 5), min_samples_leaf (1, 2), and class_weight (balanced). Model performance was evaluated based on the F1-score value for class 0. The model tuning scenario evaluation stage can be seen in Table 4, where the 70:30 scenario produced the best performance. Based on the results of testing three data distribution scenarios after the hyperparameter tuning process, the 70:30 scenario showed the best performance with an increase in class 0 recall to 0.5797 and an F1-score of 0.4746. All evaluation metrics improved compared to the baseline model, albeit to a limited extent, while the accuracy value decreased slightly due to the increased sensitivity of the model to minority classes, which resulted in an increase in the number of false positives.

Table 4. Evaluation of tuning scenario class 0 model

Skenario	Accuracy	Precision (0)	Recall (0)	F1-Score (0)
70 : 30	0,6227	0,4018	0,5797	0,4746
80 : 20	0,6201	0,3993	0,5794	0,4728
90 : 10	0,6201	0,3992	0,5787	0,4725

The evaluation results show that although the 90:10 scenario performed best at the baseline stage, after hyperparameter tuning, the 70:30 scenario produced more optimal performance. This study shows that parameter optimization affects the model's generalization ability to test data. The proportion of training data in the 70:30 scenario is considered sufficient to learn data patterns while providing representative test data, so that the model does not tend to overfit after the tuning process. The results of the model performance evaluation before and after tuning are presented in Table 5.

Table 5. Comparison of the best model performance

Model	Skenario	Accuracy	Precision (0)	Recall (0)	F1-Score (0)
Baseline Random Forest	90 : 10	0,6261	0,3989	0,5627	0,4668
Random Forest + Tuning	70 : 30	0,6227	0,4018	0,5797	0,4746

The best model evaluation is presented in the form of confusion matrix visualization, feature importance, and infographics to provide a clearer picture of model performance and the factors that influence classification results. Furthermore, the confusion matrix is used to show the number of correct and incorrect classifications in each class so that it can illustrate the model's ability to identify uninhabitable houses in detail. The evaluation results indicate that

the best model achieved an F1-score of 0.4746 and a recall of 0.5797 for the uninhabitable house class (minority class). This relatively limited performance is primarily attributed to the inherent complexity of socioeconomic data, where characteristics of families living in marginally uninhabitable houses often overlap with those in the habitable category. Furthermore, the decision to prioritize recall over precision was a deliberate strategy to minimize False Negatives. In the context of government assistance, it is more critical to avoid missing truly uninhabitable houses (False Negatives) than to occasionally misclassify habitable ones for further survey (False Positives). While alternative techniques such as SMOTE (Synthetic Minority Over-sampling Technique) could potentially increase these scores by generating synthetic data, this study strictly utilized the class_weight='balanced' parameter to maintain the integrity and original distribution of the 180,848 real-world records from BKKBN, ensuring that the model's insights remain grounded in actual field conditions (Zheng & McKenna, 2025).

The confusion matrix visualization for the best model with a 70:30 data split scenario after the tuning process is presented in Fig 5.

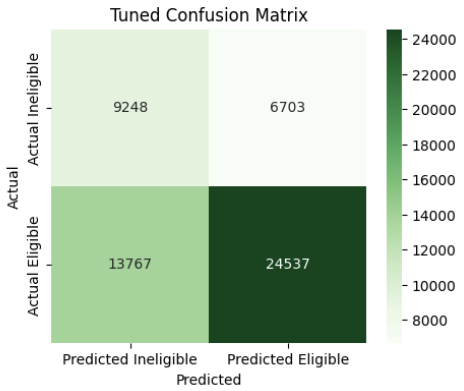

Figure 5. Confusion matrix final model

Based on the evaluation results, the Random Forest model with hyperparameter tuning in the 70:30 data split scenario was selected as the optimal model. Figure 6 presents the confusion matrix for this model, showing that it achieved a TN value of 24,537 (correctly identified livable houses) and a TP value of 9,248 (correctly identified unlivable houses). Meanwhile, the FP and FN values were 13,767 and 6,703, respectively. These results indicate that by prioritizing the minority class (unlivable houses) through hyperparameter optimization, the model demonstrates a balanced capability in distinguishing housing eligibility. Specifically, the model shows a higher sensitivity in capturing the target class of unlivable houses compared to the baseline model, which is essential for ensuring that government assistance is accurately targeted.

The next evaluation uses feature importance analysis to determine the contribution of each variable to the classification process. The feature importance results of the best model after tuning are shown in Fig 6.

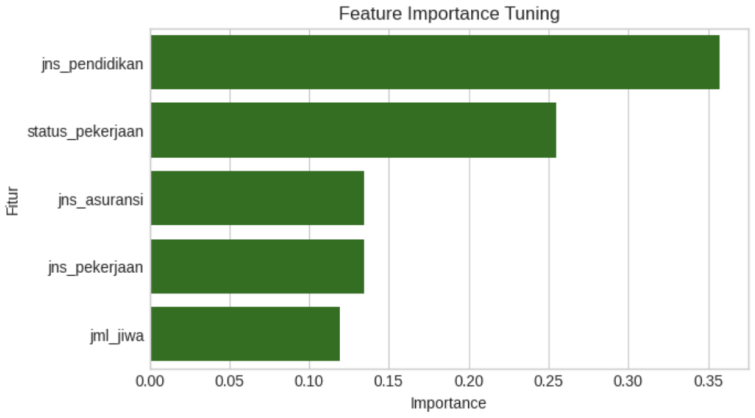

Figure 6. Feature importance final model

The feature importance results show that type of education is the most dominant variable with a value of 0.346650, followed by employment status at 0.251186. The variables of insurance type, occupation type, and number of family members have a smaller contribution but still play a role in the classification process. These results indicate that education and employment status are the main determinants in identifying uninhabitable houses.

Next, an infographic visualization of the average probability of uninhabitable houses in each subdistrict based on the model prediction results is shown in Fig 7.

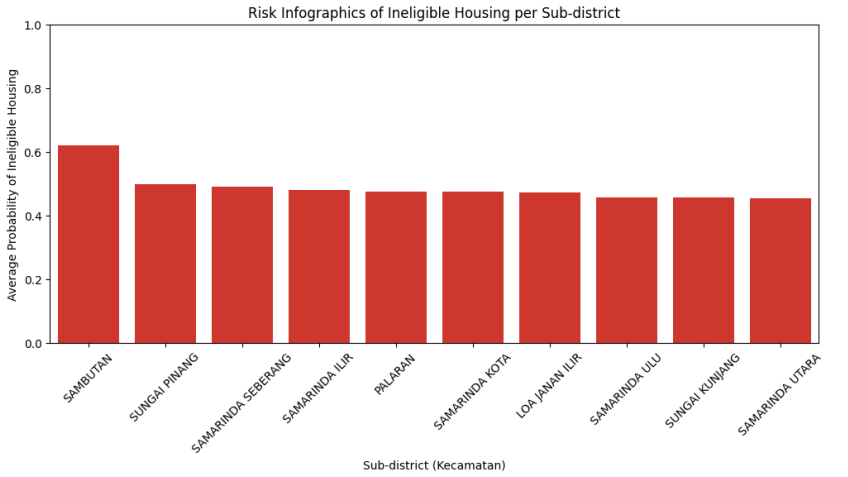

Figure 7. Model prediction

The Sambutan subdistrict has the highest probability value, indicating a relatively higher level of risk of uninhabitable houses, while Samarinda Ulu, Sungai Kunjang, and Samarinda Utara have lower probability values. The high probability of risk in the Sambutan subdistrict compared to other subdistricts such as Samarinda Ulu indicates disparities in socioeconomic conditions between regions in the city of Samarinda. This has policy implications for local governments to prioritize the allocation of survey resources to areas with higher probability of risk. Overall, the average probability of uninhabitable houses in the test data is 0.4819, which is higher than the actual proportion of 0.2940. This shows that the model tends to provide higher risk estimates in an effort to increase sensitivity in detecting uninhabitable houses.

The Random Forest algorithm demonstrated stable classification performance and effectively mitigated overfitting through its ensemble approach, consistent with previous socioeconomic studies. Unlike prior research, this study prioritizes recall and F1-score to enhance sensitivity in detecting uninhabitable houses, making it a viable tool for prioritizing housing assistance. However, limitations include a reliance on socioeconomic variables without physical building data and a relatively low precision for the minority class. Future research should incorporate diverse variables and advanced techniques for handling imbalanced datasets.

5. Conclusion

This study successfully proved that the Random Forest algorithm with hyperparameter tuning optimization can be effectively implemented to classify housing suitability based on family socioeconomic variables. Based on the results of testing three data division scenarios, it was found that the 70:30 scenario after the tuning process provided the most optimal performance, particularly in identifying the minority class of uninhabitable houses with an F1-score of 0.4746 and an increase in Recall to 0.5797. A synthesis of these findings shows that the use of the class_weight='balanced' parameter and GridSearchCV successfully achieved the research objective of overcoming data imbalance (class imbalance) in 180,848 family data sets from the BKKBN in Samarinda City. Feature importance analysis concluded that education type and employment status are the most dominant variables in determining housing eligibility status. Practically, this research contributes to local governments as an objective and efficient early screening tool to determine field survey priorities, especially in

high-risk areas such as Sambutan District. However, this research has limitations in that the model's prediction results still require factual validation by the authorities before final decisions are made. For further development, it is recommended to use more complex data imbalance handling techniques (such as SMOTE), compare with other machine learning methods (such as XGBoost or CatBoost), and add physical building predictor variables to improve the sensitivity and accuracy of the classification model.

References

Ahmad, M., Prabowo, H., Warnars, H. L. H. S., & Gaol, F. L. (2023). A Machine Learning Approach for Model Selection of Social Aid Beneficiaries. *Journal of System and Management Sciences*, *13*(6), 230–243. https://doi.org/10.33168/JSMS.2023.0614

Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., & Alyaman, M. (2021). Poverty classification using machine learning: The case of Jordan. *Sustainability (Switzerland)*, *13*(3), 1–16. https://doi.org/10.3390/su13031412

Aouichaoui, A. R. N., Liang, J., Abildskov, J., & Sin, G. (2025). Fairer benchmark of group contribution and machine learning models for property prediction: A new data splitting strategy. *Computers and Chemical Engineering*, *202*. https://doi.org/10.1016/j.compchemeng.2025.109271

Browne, C., Matteson, D. S., McBride, L., Hu, L., Liu, Y., Sun, Y., Wen, J., & Barrett, C. B. (2021). Multivariate random forest prediction of poverty and malnutrition prevalence. *PLoS ONE*, *16*(9 September). https://doi.org/10.1371/journal.pone.0255519

Ca, A. J., & Jobarteh, B. (2024). *Ensemble Learning: Methods, Techniques, Application*. https://doi.org/https://doi.org/10.13140/RG.2.2.28017.08802

Dritsas, E., & Trigka, M. (2025). Machine Learning and Data Science in Social Sciences: Methods, Applications, and Future Directions. In *IEEE Access* (Vol. 13, pp. 105334–105352). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2025.3578906

Elsayed, W. (2025). Social sustainability in housing as an entry point to achieving quality of urban life in Egypt. *Sustainable Futures*, *10*. https://doi.org/10.1016/j.sftr.2025.101045

Kalaycloğlu, O., Akhanll, S. E., Menteşe, E. Y., Kalaycloğlu, M., & Kalaycloğlu, S. (2023). Using machine learning algorithms to identify predictors of social vulnerability in the event of a hazard: Istanbul case study. *Natural Hazards and Earth System Sciences*, *23*(6), 2133–2156. https://doi.org/10.5194/nhess-23-2133-2023

Kalfountzou, E., Papada, L., Tourkolias, C., Mirasgedis, S., Kaliampakos, D., & Damigos, D. (2025). A Comparative Analysis of Machine Learning Algorithms in Energy Poverty Prediction. *Energies*, *18*(5). https://doi.org/10.3390/en18051133

Kayode, S. J., Muhammad, M. S., & Bello, M. U. (2021). Effect of Socio-Economic Characteristics of Households on Housing Condition in Bauchi Metropolis, Bauchi State, Nigeria. *Path of Science*, *7*(7), 2001–2013. https://doi.org/10.22178/pos.72-6

Kim, J., Hatzis, J. J., Klockow, K., & Campbell, P. A. (2022). *Building classification using random forest to develop a geodatabase for Probabilistic Hazard Information (PHI)*. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000561

Kinasih, A. N. S., Handayani, A. N., Ardiansah, J. T., & Damanhuri, N. S. (2024). Comparative analysis of decision tree and random forest classifiers for structured data classification in machine learning. *Science in Information Technology Letters*, *5*(2), 13–24. https://doi.org/10.31763/sitech.v5i2.1746

Koukaras, P., & Tjortjis, C. (2025). Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices. *AI*, *6*(10), 257. https://doi.org/10.3390/ai6100257

Nengsi, E. P. S., Komalla, D., Wulandari, A., Lorensya, C. N., & Aziz, M. F. (2025). Socio-Economic Status Classification of Neighborhood Residents Using the Decision Tree Algorithm. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, *4*(3), 2565–2569. https://doi.org/10.59934/jaiea.v4i3.1216

Oktafiani, R., Hermawan, A., & Avianto, D. (2024). Max Depth Impact on Heart Disease Classification: Decision Tree and Random Forest. *Jurnal RESTI*, *8*(1), 160–168. https://doi.org/10.29207/resti.v8i1.5574

Perkim.id. (2024). *36,85% Rumah Tangga Indonesia Masih Hidup dalam Kondisi Tidak Layak*. 155–173. https://doi.org/10.24815/jsu.v15i2.21452

Robiah, S., Amirullah, M. R., & Mulyadi, A. (2024). *Effectivness of the Program for Handling Uninhabitable Houses in Sukabumi City* (Vol. 22, Number 1). Jurnal Administrasi Publik. https://doi.org/https://doi.org/10.30996/dia.v22i01.8707

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/bjml/2024/007

Sathyanarayanan, S., & Tantri, R. B. (2024). Confusion Matrix-Based Performance Evaluation Metrics. *African Journal of Biomedical Research*, 4023–4031. https://doi.org/10.53555/ajbr.v27i4s.4345

Shahidi, S., Samadzai, A. W., & Shahbazi, H. (2025). Effective Data Preprocessing in Data Science: From Method Selection to Domain-Specific Optimization. *Journal of Advanced Computer Knowledge and Algorithms*, *2*(4), 84–90. https://doi.org/10.29103/jacka.v2i4.22886

Sholihah, N. N., & Hermawan, A. (2023). Comparison of Machine Learning Algorithms for Household's Economic Status Classification. *International Journal of Computer Applications*, *185*(50), 6–13. https://doi.org/10.5120/ijca2023923334

Sudarawerti, G., & Arif, F. (2025). *Improving Housing Price Prediction with Machine Learning: Evidence from Yogyakarta and Implications for Emerging Urban Markets*. https://doi.org/https://doi.org/10.31098/ijmesh..v9i1.3567

Summer, K., Anpalagan, K., Stacey, I., Stiles, S., Burgess, R., Wade, V., Bowen, A. C., Katzenellenbogen, J., & Wyber, R. (2026). Infectious disease outcomes associated with inadequate housing and access to healthy living practices in Australia: a systematic review. *BMJ Public Health*, *4*(1), e003531. https://doi.org/10.1136/bmjph-2025-003531

Zhang, F., Luo, Y., Dong, Y., Zhang, Q., & Han, A. (2026). Machine Learning Applications for Sustainable Housing Policy: Understanding Price Determinants to Inform Affordable Housing Strategies. *Algorithms*, *19*(2). https://doi.org/10.3390/a19020098

Zhao, Y., Zhu, W., Wei, P., Fang, P., Zhang, X., Yan, N., Liu, W., Zhao, H., & Wu, Q. (2022). Classification of Zambian grasslands using random forest feature importance selection during the optimal phenological period. *Ecological Indicators*, *135*. https://doi.org/10.1016/j.ecolind.2021.108529

Zheng, L., & McKenna, E. (2025). Machine Learning with Administrative Data for Energy Poverty Identification in the UK. *Energies*, *18*(12). https://doi.org/10.3390/en18123054