# Optimizing Limited Class-Imbalanced Data for Classification of Regional Food Prices in Indonesia

I Made Sumertajaya, Embay Rohaeti, Anwar Fitrianto, Windhiarso Ponco Adi P.

***Abstract*—This study aims to classify 411 Indonesian regencies and cities with sparse historical food price data based on economic characteristics, addressing challenges related to class imbalance and limited data availability. Building on a previous classification of 103 regions, we utilize nine key economic indicators—Average School Duration, Expected School Duration, Expenditure, Food Security Index, Gross Regional Domestic Product, Life Expectancy, Poor Percentage, Rural Population, and Urban Population—to categorize these unclassified regions. We evaluate the performance of three machine learning models—Random Forest, Linear SVM, and RBF SVM—under eight preprocessing techniques for class imbalance handling, including naïve oversampling, SMOTE, and ADASYN. Hyperparameter tuning was conducted using bootstrap resampling with 100 repetitions, yielding a mean balanced accuracy of 0.998 (SD = 0.016) and mean ROC-AUC of 0.998 (SD = 0.017). The best-performing model, Random Forest with naïve oversampling (over-ratio = 0.75) and eight selected features, achieved perfect classification accuracy on the validation set. The Food Security Index and Gross Regional Domestic Product emerged as the most influential variables. The resulting classification framework provides a basis for identifying regions with similar food price dynamics, enabling policymakers to apply interventions based on comparable areas. By supporting data-driven decision-making for food price stability, this study contributes to economic resilience and sustainable development in Indonesia.**

***Index Terms*— bootstrap resampling, class imbalance, region classification, random forest**

## I. Introduction

MAINTAINING stability in food prices is a critical aspect of economic management, especially in large and diverse countries like Indonesia. Food price fluctuations can significantly disrupt the Food Security Index, which measures the availability and accessibility of food for the as the cost of essential goods rises, reducing the purchasing power of households which, in turn, decreases consumer spending, creating a ripple effect that hinders economic growth. This situation raises the idea of the importance of food price forecasting. Moreover, the ability to predict food prices accurately enables policy makers to implement timely interventions, but make sure that food remains affordable and accessible. By doing so, it is possible to maintain a stable Food Security Index, control inflation rates, and support sustained economic development. This strategy is essential for mitigating the adverse effects of price volatility, particularly in a country as large and varied as Indonesia, where regional disparities can further complicate economic management. Therefore, the importance of predicting food prices cannot be overstated, as it is a fundamental element in safeguarding economic stability and growth.

Various studies have been done to forecast food prices in Indonesia. For instance, [1] forecasted beef prices in eight provinces: Jakarta, Banten, West Java, East Java, East Nusa Tenggara, West Nusa Tenggara, Bali, and Lampung. The study also found that there was cointegration of beef prices. Another study focused on the price of red chili in Banyumas Regency [2] using ARIMA, and found that the red chili prices were fluctuating within the period of study. Some studies have also focused on rice prices, which are crucial due to rice's role as a staple food in Indonesia. [3] developed a model to forecast the average national rice price, [4] built a forecasting model for rice and corn prices in Central Sulawesi Province, while [5] created a forecast model for rice prices in six provinces: West Java, Central Java, East Java, Jakarta, Yogyakarta, and Banten. These studies have produced accurate models for forecasting specific commodities in their respective regions. However, food prices are often interrelated across different commodities and regions, suggesting that a more comprehensive study encompassing multiple commodities and regions would be ideal.

Previous studies conducted by Rohaeti e. al [6] formed four clusters of 103 regencies and cities (hereafter referred to as "survey regions") in Indonesia using historical prices of 13 food commodities through multivariate time series clustering (MTSClust). These clusters are important to identify generalized patterns across regions, which could be used to forecast prices in the survey regions. Although this method provided valuable insights, unfortunately it was constrained by data availability, and leave 411 regencies and cities (hereafter referred to as "non-survey regions") unlabeled. This situation creates problems in developing forecasting models that can serve as early warning systems to prevent

Manuscript received June 27, 2024; revised February 20, 2025.

This work was supported by the Ministry of Research, Technology and Higher Education of Indonesia.

I Made Sumertajaya is an Associate Professor at Department of Statistics, IPB University, Bogor Regency, West Java, 16144, Indonesia. (corresponding author; phone: +62 (251) 8624535, fax: +62 (251) 8624535, e-mail: imsjaya@apps.ipb.ac.id).

Embay Rohaeti is a senior lecturer at Mathematics Study Program, Pakuan University, Bogor City, West Java, 16129, Indonesia (e-mail: embay.rohaeti@unpak.ac.id).

Anwar Fitrianto is a senior lecturer at Department of Statistics, IPB University, Bogor Regency, West Java, 16144, Indonesia (e-mail: anwarstat@gmail.com).

Windhiarso Ponco Adi P. is the director of price statistics directorate of BPS-Statistics Indonesia, Central Jakarta, 10710, Indonesia (e-mail: windhiarso@bps.go.id).

food price inflation across all regions in Indonesia. Therefore, this study specifically aims to address the data limitation by accurately classifying the non-survey regions based on the classifications of the survey regions. By doing so, it seeks to extend the benefits of the early warning system to all 514 regencies and cities in Indonesia, enabling more comprehensive and proactive economic management.

This study aims to classify non-surveyed regions using existing classifications from surveyed regions, optimizing classification models to handle imbalanced class distributions effectively. By exploring various preprocessing techniques and classification approaches, the study seeks to improve predictive accuracy despite data limitations. Region-specific variables, such as population distribution, per capita expenditure, education indicators, life expectancy, poverty rate, Gross Regional Domestic Product, and the Food Security Index, serve as key predictors for model development.

One of the primary challenges is the inconsistency in data availability across regions, with some variables recorded up to 2022 and others extending to 2023. To ensure uniformity, the classification models are standardized using 2022 data. By leveraging optimized models for limited and imbalanced data, this research contributes to more effective economic planning and management of regional food markets.

## II. Random Forest

Random Forest (RF) is an ensemble learning model that combines multiple decision trees to improve prediction accuracy and robustness [7], [8]. By constructing a forest of decision trees, each trained on a random subset of features, the model aggregates the predictions from individual trees to enhance performance and reduce overfitting [9], [10], [11].

The RF model typically involves the following procedures:

1) Bootstrap sampling
   Multiple bootstrap samples are drawn from the training set, each used to construct a decision tree.
2) Random Forest Selection
   A random subset of variables is selected for each node in the tree. The best split, which maximizes a particular criterion, is chosen for each node. For classification, common criteria include Gini impurity, information gain, or classification error. In this study, Gini impurity was used, defined mathematically as follows [12], [13]:

$$I_G(p) = \sum_{i=1}^{J}\left(p_i \sum_{k\neq i} p_k\right) = \sum_{i=1}^{J} p_i(1-p_i) = 1 - \sum_{i=1}^{J} p_i^2 \quad (1)$$

   where $J$ is the number of classes and $p_i$ is the probability of selecting an item with label $i$, $i \in \{1,2,\dots,J\}$. Therefore, $\sum_{k\neq i} p_k = 1 - p_i$ is the probability of misclassifying that item.
3) Tree Construction
   Each tree is grown to its maximum extent, ensuring a diverse set of classifiers. This process allows individual trees to capture different patterns and reduce overfitting.
4) Aggregation
   For classification tasks, the final output is determined by majority voting among the individual trees, which can be mathematically expressed as follows:

$$\bar{y} = mode\{h_1(x), h_2(x), \dots, h_B(x)\} \quad (2)$$

   where $h_i$ is the prediction of the $i$-th tree, and $B$ is the total number of trees.

## III. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning model used for classification and regression. A linear SVM is a type of SVM model that aims to find a hyperplane that separates the data into two classes. The optimal hyperplane is chosen to maximize the margin, which is the distance between the closest points of each class. A hyperplane can be written as the set of points $x$ satisfying the following:

$$\boldsymbol{w}\cdot\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{0} \quad (3)$$

Where $\boldsymbol{w}$ is the normal vector to the hyperplane, $\boldsymbol{x}$ represents the variables, and $\boldsymbol{b}$ is the bias term.

For linearly separable binary data, the optimization problem can be formulated as follows:

$$\min_{\mathrm{w.b}} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{subject to } y_i(\boldsymbol{w}\cdot\boldsymbol{x_i} + \boldsymbol{b}) \geq 1 \quad (4)$$

Where $\boldsymbol{x_i}$ are the variable vectors and $\boldsymbol{y_i}$ are the classes.

This concept can be extended to a multi-class scenario. One common approach is One-vs-One (OvO), where the optimization problem is formulated as follows [14], [15]:

$$\min_{\mathrm{w_k.b_k}} \sum_{k=1}^{C}\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right) \quad \text{subject to } y_i(\boldsymbol{w_k}\cdot\boldsymbol{x_i} + \boldsymbol{b_k}) \geq 1 - \xi_i \mathrm{c} \quad (5)$$

where $\boldsymbol{w_k}$ and $\boldsymbol{b_k}$ are the parameters for class $\boldsymbol{k}$ and $\boldsymbol{\xi_i}$ are slack variables to handle non-separable cases.

The Radial Basis Function (RBF) SVM model is a type of SVM used for non-linear classification tasks, extending the SVM framework to handle complex decision boundaries using kernel functions. The RBF kernel can be mathematically expressed as follows [16], [17]:

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (6)$$

where $\|x_i - x_j\|^2$ is the Euclidean distance between points $x_i$ and $x_j$, and $\sigma$ is the variance.

## IV. Methodology

### A. Data

This study used region-specific data from 514 regencies and cities in Indonesia, among which 103 regencies and cities are labeled. This study employs machine learning models to classify the 411 unlabeled regencies and cities based on observed characteristics in the labeled dataset. To develop these models, we use 9 predictors as detailed in Table I.

TABLE I
PREDICTOR VARIABLES AND THEIR DESCRIPTIONS

| Variable | Description |
|---|---|
| Average School Duration (ASD) | The number of years of education completed by individuals aged 15 and above who have finished formal education (excluding repeated years). |
| Expected School Duration (ESD) | The number of years of education that children at a certain age are expected to complete in the future. |
| Expenditure | Consumption costs per resident, adjusted to purchasing power parity (thousand Rupiah per person per year). |
| Food Security Index (FSI) | An index defined by Indonesian's National Food Agency, based on nine indicators derived from three aspects of food security: availability, affordability and utilization of food. |
| Gross Regional Domestic Product (GRDP) | The total value (in Rupiah) of all final goods and services produced within a region, measured in Rupiah. This includes household consumption expenditure and non-profit private institutions. |
| Life Expectancy | The average estimated lifespan (years) of residents. |
| Poor Percentage | The percentage of the population living below the poverty line within a regency and city. |
| Rural Population | The total number of residents living in rural areas. |
| Urban Population | The total number of residents living in urban areas. |

### *B. Resampling*

To ensure a robust evaluation of the models, we randomly split the labeled dataset of 103 regencies and cities into training and testing sets, using an 80:20 ratio. This resulted in a training set of 82 regencies and cities and a testing set of 21. The split was stratified to maintain similar class distributions in both sets. The testing set was kept unseen by the models and used only to measure the final models' accuracy.

During the training process, a validation set is necessary, particularly for hyperparameter tuning. Given the small size of the training set, further splitting would hinder modeling. To address this, we employed bootstrapping as the resampling method. A bootstrap sample is created by sampling with replacement, producing a sample of the same size as the original dataset [18], [19]. Observations not selected in this process are used as the validation set. Bootstrapping allows the small training set to be replicated, enabling each iteration to be trained and validated on distinct data sets.

The bootstrapping process was iterated 100 times for each model and preprocessing method. This number of iterations was chosen to balance computational feasibility with the need to effectively test various model and preprocessing combinations. Fig. 1 shows the flowchart of this study.

## V. RESULT AND DISCUSSION

### *A. Class Imbalance*

Before the training process, it is important to understand the distribution of the target variable. As mentioned in the data sub-section, the labeled data consists of 103 observations from surveyed regions, divided into four classes. The class distributions are shown in Fig. 2.

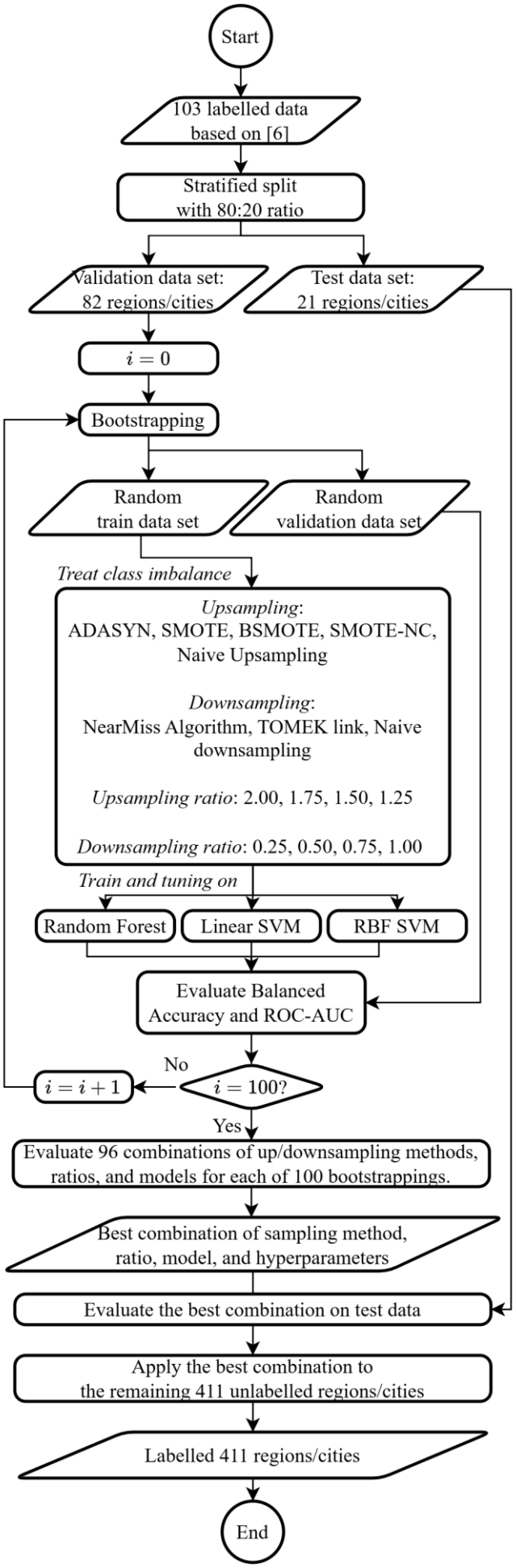


Fig. 1. Flowchart of the study

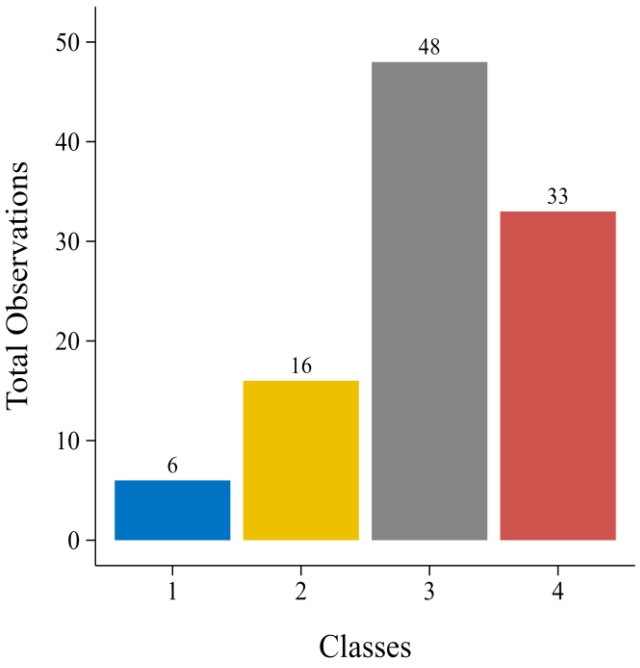


Fig. 2. Class distribution of the surveyed regions

As shown in Fig. 2, the target variable is class-imbalanced, with Classes 1 and 2 underrepresented compared to the other classes. Addressing class imbalance is crucial because models trained on imbalanced data tend to be biased toward the majority classes, which reduces the accuracy and robustness of predictions, especially for minority classes [20]. Failure to address this issue may result in high overall accuracy but poor balanced accuracy due to the misclassification of the minority classes.

To tackle this issue, eight upsampling and downsampling preprocessing methods were employed to balance the class distribution. The upsampling methods include:

1) Adaptive Synthetic Sampling (ADASYN) [21], [22]
2) Synthetic Minority Oversampling Technique (SMOTE) [22], [23], [24]
3) Borderline SMOTE (BSMOTE) [25]
4) SMOTE for Nominal and Continous (SMOTE-NC) [26]
5) Naïve upsampling [27]

The downsampling methods include:

1) NearMiss Algorithm [28]
2) TOMEK link [29], [30], [31]
3) Naïve downsampling [32]

Each method was tested with various ratios. The upsampling methods were tested with over-sampling ratios of 0.25, 0.50, 0.75, and 1.00, while the downsampling methods were tested with under-sampling ratios of 2.00, 1.75, 1.50, and 1.25. After resampling, these methods were compared, and the best method and ratio were chosen for final data training.

*B. Distribution of the Predictor Variables*

Understanding the distribution of predictor variables is equally important for building robust machine learning models. Assessing these distributions can reveal significant differences between classes in certain predictors, potentially identifying strong discriminative features for classification tasks. Fig. 3 presents the distributions of the predictor variables across the surveyed regions, grouped by class.

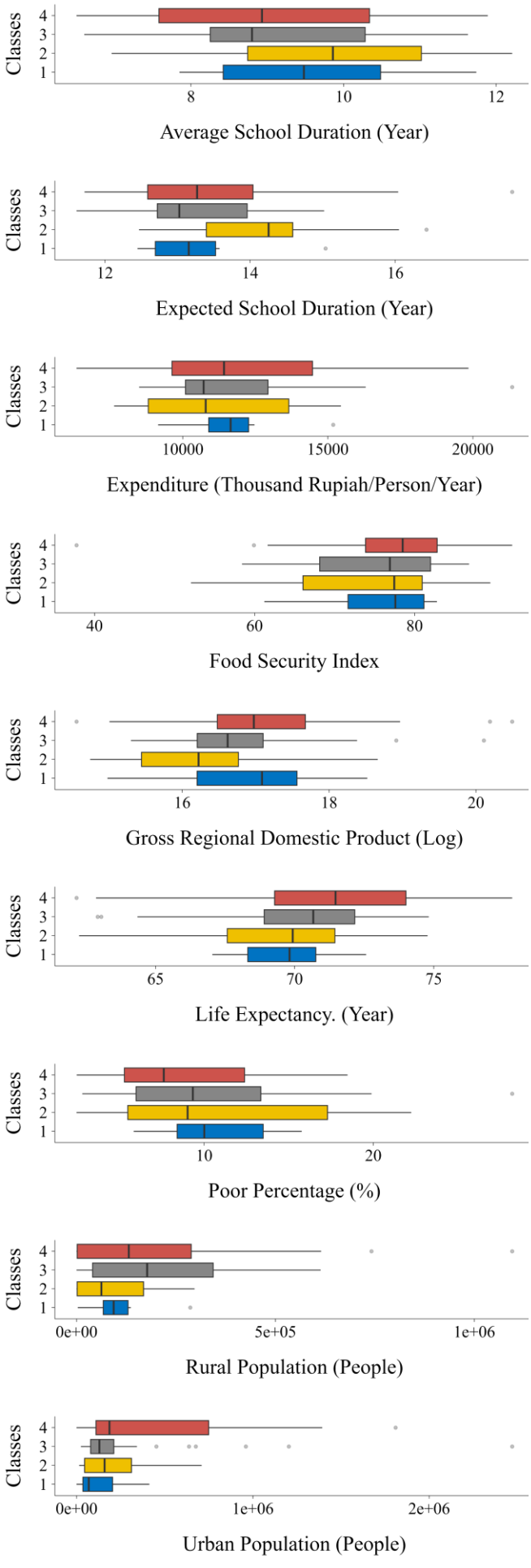


Fig. 3. Predictor distributions of the surveyed regions

Fig. 3 offers several key insights. Firstly, the scales of the predictor variables vary significantly, which is particularly relevant for SVM models. Since SVM relies on distance metrics, it is sensitive to the scale of input data. In contrast, scale differences have less impact on Random Forest (RF) models, which are tree-based rather than distance-based. Therefore, normalization will be applied before training the SVM models to ensure optimal performance.

Secondly, the boxplots in Fig. 3 indicate distinct distributions among classes for several predictor variables. These distinctions can enhance prediction accuracy, especially for tree-based models like Random Forest.

### *C. Training and Tuning Process*

To ensure comparability among models, the training process began by randomly bootstrapping the training data. These bootstrap samples were consistently used across all models and preprocessing methods, ensuring that each model worked with the same set of data. However, due to the random nature of bootstrapping, some samples may contain zero observations, particularly for the minority Class 1. In such cases, the sample was replaced with another random set of bootstrap samples, ensuring that the total number of samples remained 100. The observation counts for each class are shown in Fig. 4. As indicated, each class has at least one observation in every bootstrap sample, preventing modeling errors due to missing classes in the validation set.

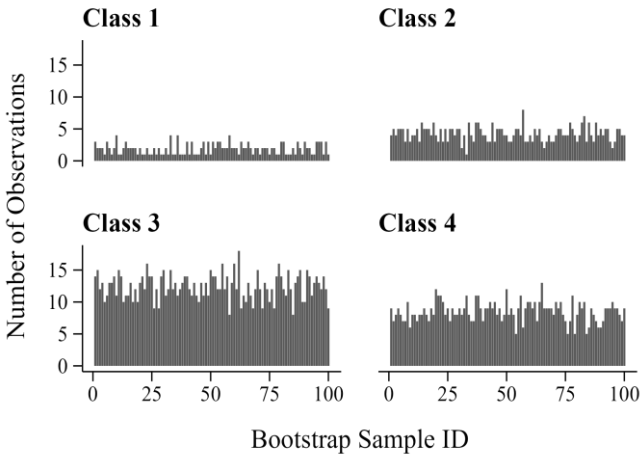


Fig. 4. Number of observations in validation set grouped by class.

### *Random Forest*

The tuning results for combinations of preprocessing methods, ratios, and hyperparameters were evaluated using Balanced Accuracy and ROC-AUC. The distributions for each preprocessing method are shown in Fig. 5. Each point in the boxplots represents the mean of 100 resamples, while the boxplot itself represents the distribution of mean values across hyperparameters, grouped by preprocessing method

As seen in Fig. 5, SMOTENC (over-ratio = 1), SMOTE (over-ratio = 1), and ADASYN (over-ratio = 1) were the top three preprocessing methods, with median Balanced Accuracy just below 1 and median ROC-AUC equal to 1. However, these methods occasionally resulted in errors, as indicated by the number of successful bootstrap samples being around 80-90. In contrast, naïve upsampling (over-ratio = 1) produced similar distributions but with 100 successful bootstrap samples, making it the most reliable method.

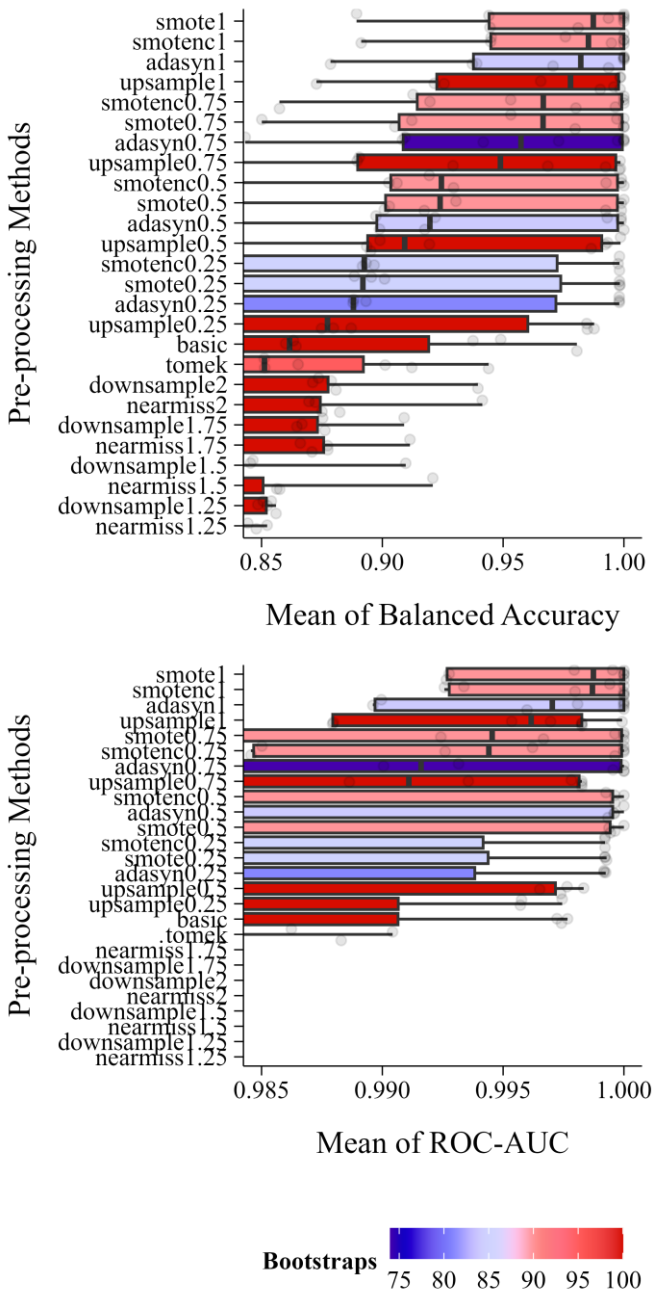


Fig. 5. Evaluation of Random Forest tuning on train data with 100 bootstraps

TABLE II
BEST TUNING RESULTS OF RANDOM FOREST MODEL

| Prep. Methods | Ratio | Var. | Min. Obs. | Boots. | BAL. ACC. | | ROC-AUC | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Mean | SD | Mean | SD |
| Up | 0.75 | 8 | 12 | 100 | 0.998 | 0.016 | 0.998 | 0.017 |
| Up | 1.00 | 8 | 12 | 100 | 0.998 | 0.016 | 0.998 | 0.017 |
| Up | 1.00 | 8 | 16 | 100 | 0.998 | 0.016 | 0.998 | 0.017 |
| Up | 0.75 | 8 | 16 | 100 | 0.998 | 0.016 | 0.998 | 0.017 |
| Up | 1.00 | 7 | 24 | 100 | 0.996 | 0.024 | 0.998 | 0.017 |
| Up | 0.50 | 8 | 12 | 100 | 0.993 | 0.030 | 0.998 | 0.018 |
| Up | 0.75 | 7 | 24 | 100 | 0.992 | 0.029 | 0.998 | 0.018 |
| Up | 0.50 | 8 | 16 | 100 | 0.992 | 0.034 | 0.997 | 0.019 |
| Up | 1.00 | 6 | 39 | 100 | 0.990 | 0.028 | 0.997 | 0.018 |
| Up | 0.50 | 7 | 24 | 100 | 0.986 | 0.040 | 0.997 | 0.019 |

Regarding the best hyperparameters, Table II shows the 10 best models across all approaches with 100 successful bootstrap samples. Table II confirms that among the methods with 100 successful samples, naïve upsampling is indeed the best preprocessing method for this case. The means of

Balanced Accuracy and ROC-AUC vary based on the over-ratio, but over-ratios of 1 and 0.75 perform slightly better.

Hyperparameters also play a crucial role. Table II suggests that the Random Forest model, with 8 random variables and a minimum of 12 observations, provides the best Balanced Accuracy and ROC-AUC.

In summary, the training and tuning process of the Random Forest model yielded an optimal model with 1,000 trees, 8 random variables, and a minimum of 12 observations. The preprocessing method is naïve upsampling with an over-ratio of 0.75. This optimal Random Forest model was tested on the validation data, achieving 100% correct predictions. The confusion matrix is shown in Table III, where all classes, including the minority class 1, are classified correctly.

TABLE III
CONFUSION MATRIX OF THE OPTIMAL RANDOM FOREST MODEL'S PREDICTION ON VALIDATION DATA

| | | Truth | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Predictions | 1 | 1 | 0 | 0 | 0 |
| | 2 | 0 | 4 | 0 | 0 |
| | 3 | 0 | 0 | 13 | 0 |
| | 4 | 0 | 0 | 0 | 9 |

Fig. 6 displays the variable importance of the optimal Random Forest model based on Gini impurity. As shown, the Food Security Index (FSI) and Gross Regional Domestic Product (GRDP) are the most important variables for predicting a region's class, followed by life expectancy and poverty percentage.

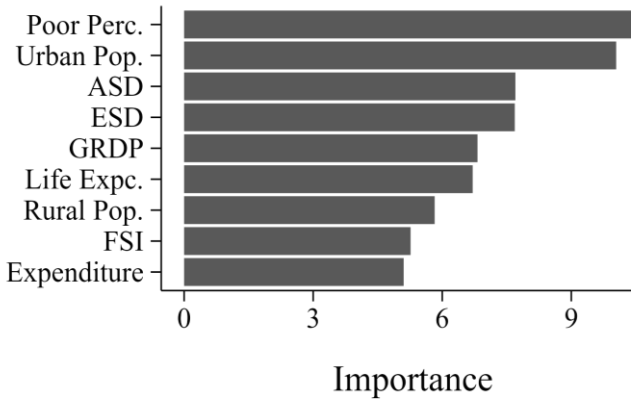


Fig. 6. Variable importance of the optimal Random Forest model

*Linear SVM*

Similar to the Random Forest model, Linear SVM was also evaluated using Balanced Accuracy and ROC-AUC. Fig. 7 shows the boxplots of the tuning results for the Linear SVM model, where each boxplot represents the same concept as those previously shown for Random Forest.

Fig. 7 indicates that the basic approach (no pre-processing) and TOMEK-link are two of the best approaches for the Linear SVM model. Although these methods did not achieve as many successful samples as the naïve upsampling method (over-ratio = 0.25), which had 100 successful samples, the basic and TOMEK-link approaches had 99 and 93 successful samples, respectively. Given the overall superior ROC-AUC values of the basic approach compared to the naïve upsampling (over-ratio = 0.25), the basic approach was selected as the best pre-processing method for Linear SVM.

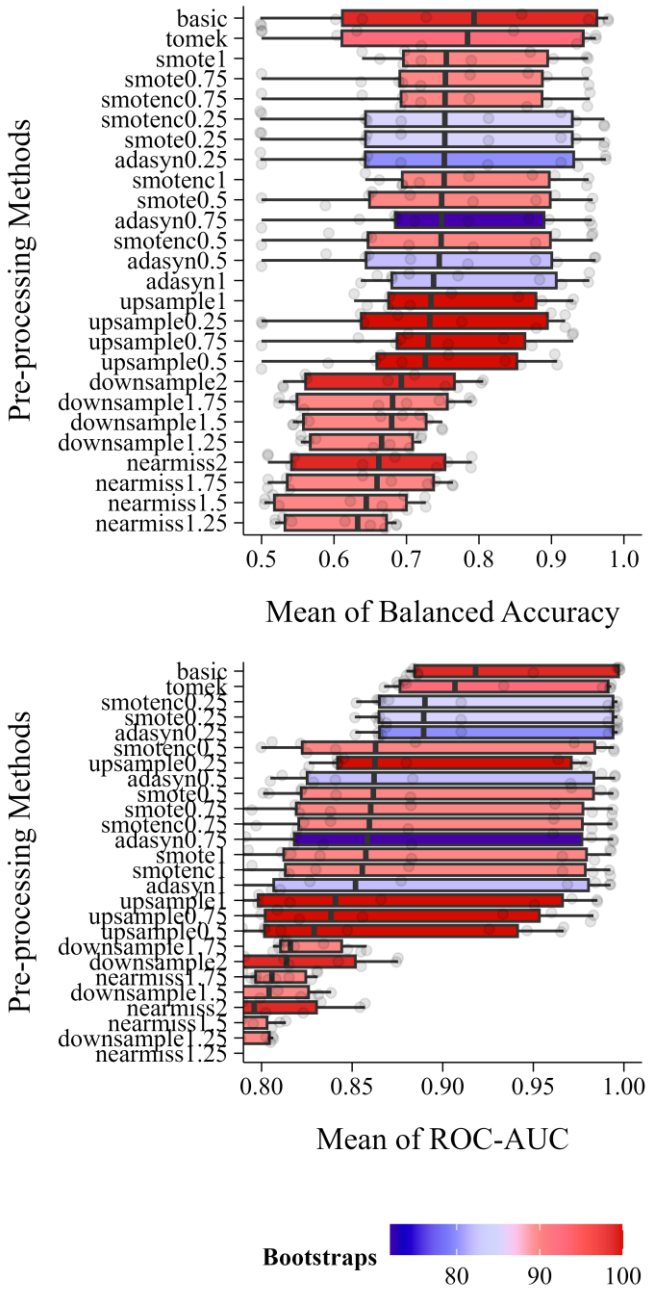


Fig. 7. Evaluation of Linear SVM tuning on train data with 100 bootstraps

TABLE IV
BEST TUNING RESULTS OF LINEAR SVM MODEL

| Prep. Methods | Ratio | Cost. | Boots. | Bal. Acc. | | ROC-AUC | |
|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Mean | SD |
| Basic | 0.00 | 15.880 | 99 | 0.978 | 0.046 | 0.998 | 0.008 |
| Basic | 0.00 | 6.611 | 99 | 0.978 | 0.046 | 0.997 | 0.009 |
| Basic | 0.00 | 1.574 | 99 | 0.966 | 0.050 | 0.998 | 0.007 |
| Basic | 0.00 | 0.957 | 99 | 0.951 | 0.053 | 0.996 | 0.008 |
| Up | 1.00 | 15.880 | 100 | 0.930 | 0.071 | 0.985 | 0.025 |
| Up | 0.75 | 15.880 | 100 | 0.930 | 0.072 | 0.983 | 0.034 |
| Up | 0.75 | 6.611 | 100 | 0.929 | 0.072 | 0.982 | 0.039 |
| Up | 1.00 | 6.611 | 100 | 0.927 | 0.071 | 0.985 | 0.024 |
| Up | 0.25 | 15.880 | 100 | 0.919 | 0.066 | 0.980 | 0.039 |
| Up | 0.25 | 6.611 | 100 | 0.919 | 0.066 | 0.979 | 0.039 |

The ten best preprocessing methods and hyperparameters for Linear SVM are shown in Table IV. Among the best tuning results, the naïve upsampling method is also among the top for Linear SVM, with 100 successful samples. Although the basic approach had only 99 successful samples,

its Balanced Accuracy and ROC-AUC were significantly better than those of the naïve upsampling method. Given its high success rate, with only one failed sample, the chosen method for Linear SVM is the basic approach, with a cost value of 15.880.

This optimal Linear SVM model was then applied to the validation set to estimate its performance on new data. Table V presents the confusion matrix of the predictions. As shown, all classes were correctly classified, including the minority class (Class 1). This result mirrors that of the Random Forest model.

TABLE V
CONFUSION MATRIX OF THE OPTIMAL LINEAR SVM MODEL'S PREDICTION ON VALIDATION DATA

| | | Truth | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Predictions | 1 | 1 | 0 | 0 | 0 |
| | 2 | 0 | 4 | 0 | 0 |
| | 3 | 0 | 0 | 13 | 0 |
| | 4 | 0 | 0 | 0 | 9 |

*RBF SVM*

Contrary to the good results from the Random Forest and Linear SVM models, the tuning results for RBF SVM showed generally low accuracy. Fig. 8 illustrates the distribution of mean values for Balanced Accuracy and ROC-AUC during the RBF SVM's tuning process.

Fig. 8 shows that while the ROC-AUC values are decent, ranging from 0.7 to 0.8, the Balanced Accuracy is poor, indicating severe misclassification of certain classes. This issue is also reflected in the 10 best tuning results shown in Table VI. Table VI indicates that the best hyperparameter tuning result for the RBF SVM model has a low Balanced Accuracy, with an average of only 0.632. The ROC-AUC is also low, at 0.447. This performance is further reflected when tested on the validation set, as shown in Table VII. The confusion matrix for the RBF SVM in Table VII shows that the majority classes are the most misclassified, which significantly contributes to the low Balanced Accuracy and

TABLE VI
BEST TUNING RESULTS OF RBF SVM MODEL

| Prep. Methods (Ratio) | Cost | Sigma | Boots. | Bal. Acc. | | ROC-AUC | |
|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Mean | SD |
| Up (1) | 0.008 | 0.002 | 100 | 0.632 | 0.063 | 0.447 | 0.261 |
| Up (1) | 0.001 | 0.098 | 100 | 0.629 | 0.071 | 0.429 | 0.260 |
| Up (1) | 0.120 | 0.000 | 100 | 0.629 | 0.063 | 0.448 | 0.259 |
| Up (1) | 18.698 | 0.000 | 100 | 0.628 | 0.062 | 0.448 | 0.261 |
| Up (1) | 0.568 | 0.000 | 100 | 0.628 | 0.063 | 0.450 | 0.261 |
| Up (1) | 1.985 | 0.000 | 100 | 0.628 | 0.064 | 0.439 | 0.252 |
| Up (1) | 4.060 | 0.000 | 100 | 0.627 | 0.062 | 0.438 | 0.250 |
| Up (1) | 0.022 | 0.000 | 100 | 0.619 | 0.062 | 0.346 | 0.132 |
| Up (1) | 0.038 | 0.000 | 100 | 0.585 | 0.061 | 0.346 | 0.132 |
| Down (2) | 0.001 | 0.098 | 99 | 0.538 | 0.052 | 0.697 | 0.152 |

TABLE VII
CONFUSION MATRIX OF THE OPTIMAL RBF SVM MODEL'S PREDICTION ON VALIDATION DATA

| | | Truth | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Predictions | 1 | 1 | 0 | 0 | 0 |
| | 2 | 0 | 2 | 4 | 5 |
| | 3 | 0 | 2 | 8 | 3 |
| | 4 | 0 | 0 | 1 | 1 |

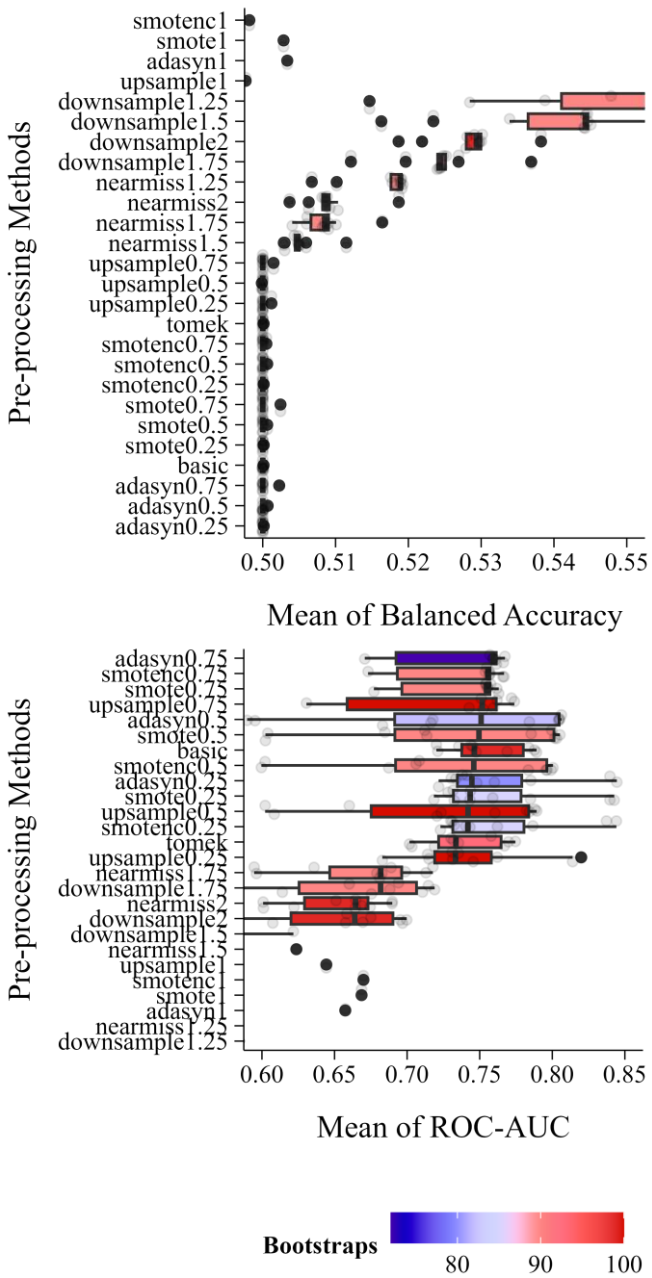


Fig. 8. Evaluation of RBF SVM tuning on train data with 100 bootstraps

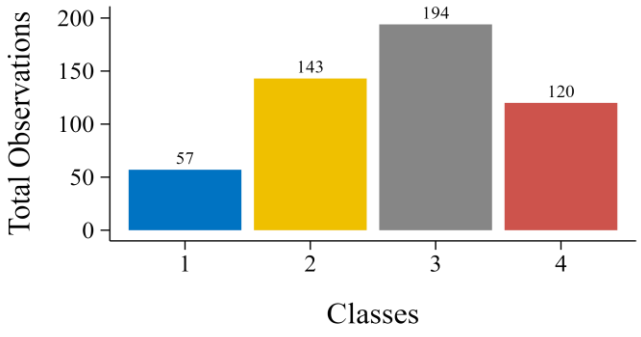


Fig. 9. Class distribution across all regions

ROC-AUC scores. The consistent misclassification of majority classes by the RBF SVM underscores its inability to handle imbalanced datasets, which is crucial for this task. Consequently, the Random Forest or Linear SVM models are likely to produce more accurate predictions for the 411 unlabeled/non-survey regions, ensuring better classification results for the entire dataset.

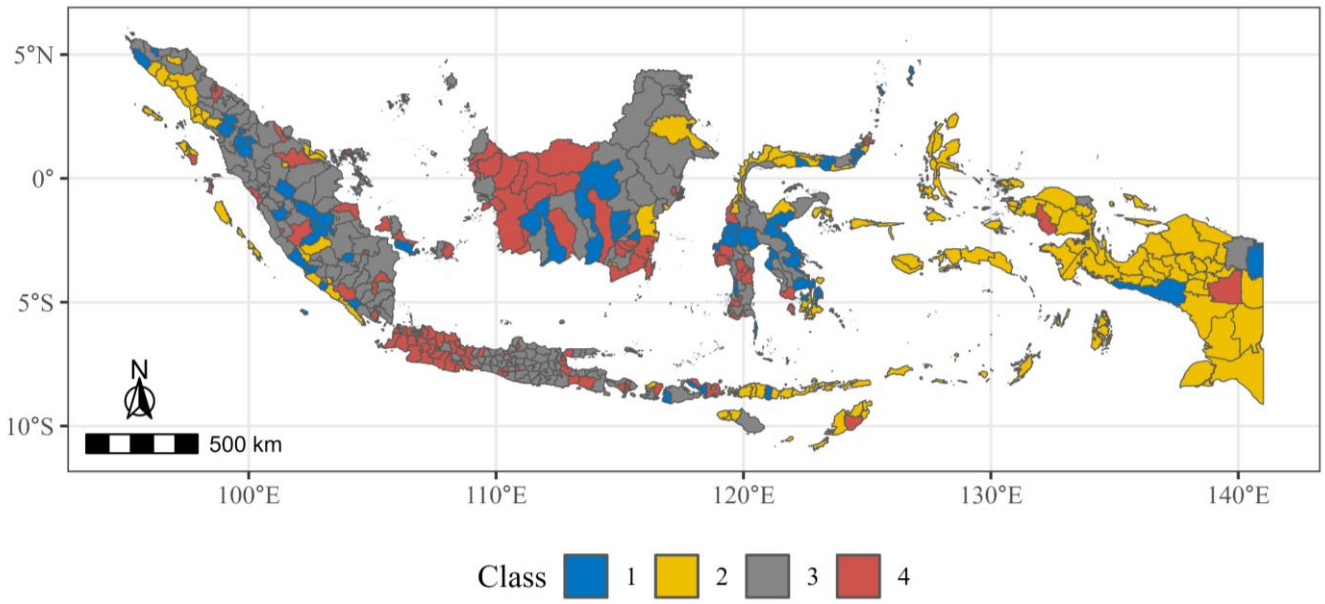


Fig. 10. Map of classification results across all regions.

*D.Final Predictions*

Based on the previous analysis, two models stand out as the best options:

1) Random Forest with naïve upsampling (over-ratio = 0.75), number of variables = 8, and minimum observations = 12.
2) Linear SVM with basic preprocessing and cost=15.880.

Both models achieved 100% accuracy on the validation set, suggesting highly accurate predictions on the test set. However, the optimal Linear SVM model only works with 99 samples, whereas the optimal Random Forest model successfully handles all samples. Additionally, the optimal Linear SVM model has slightly lower Balanced Accuracy and ROC-AUC scores, on average, compared to the optimal Random Forest. Therefore, the best model to predict the test data is the optimal Random Forest model.

The class distribution of the final predictions is shown in Fig. 9, which indicates a similar distribution to the 103 surveyed regions, where Class 1 is the minority class, and Class 3 is the majority class. A slight difference can be observed for Class 2. Although the surveyed data show more Class 4 regions than Class 2 regions, the classification model predicts that there are more Class 2 regions than Class 4 regions in the non-surveyed regions. Most of these Class 2 regions are in eastern Indonesia, as shown in Fig. 10.

## VI. Conclusion

By leveraging the classification of 103 surveyed regions, this study successfully extended the categorization to 411 Indonesian regencies and cities with sparse historical food price data. Evaluating multiple classification models and preprocessing techniques to address class imbalance, we identified the Random Forest model with naïve oversampling (over-ratio = 0.75) and eight selected economic indicators as the optimal approach. During hyperparameter tuning with bootstrapping, this model achieved a mean balanced accuracy of 0.998 (SD = 0.016) and a mean ROC-AUC of 0.998 (SD = 0.017) on the validation set, demonstrating strong generalization performance. When applied to the test set, it achieved perfect accuracy, ensuring reliable classification across all classes, including minority categories. Expanding this classification framework to all 514 regencies and cities enhances the effectiveness of early warning systems for food price stability, with the final class distribution closely aligning with that of the surveyed regions. These findings highlight the importance of robust modeling and preprocessing in addressing class imbalances and improving classification accuracy, ultimately supporting data-driven policy interventions for food security and economic resilience in Indonesia.

## References

[1] Firmansyah, P. Maruli, and A. Harahap, "Analysis of beef market integration between consumer and producer regions in Indonesia," *Open Agriculture*, vol. 8, no. 1, 2023.

[2] K. Prasetyo, D. D. Putri, I. Kartika Eka Wijayanti, and L. Zulkifli, "Forecasting of Red Chilli Prices in Banyumas Regency: The ARIMA Approach," in *E3S Web of Conferences*, vol. 444, 2023.

[3] C. Dewi, G. S. K. Prasatya, H. J. Christanto, S. O. B. Widiarto, and G. Dai, "Modified Random Forest Regression Model for Predicting Wholesale Rice Prices," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 23, pp. 7749-7759, 2023.

[4] Effendy, D. Evansyah, M. Antara, K. Noli, and M. F. Pratama, "Forecasting model of production and price of grains commodity in central Sulawesi," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 14, pp. 3555-3563, 2021.

[5] H. Primageza, R. A. Vinarti, R. Tyasnurita, E. Riksakomara, and A. Muklason, "Comparison of NNs-ARIMAX and NNs-GSTARIMAX on Rice Price Forecasting in Indonesia," in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp.1-8.

[6] E. Rohaeti, I. M. Sumertajaya, A. H. Wigena, and K. Sadik, "MTSClust with Handling Missing Data Using VAR-Moving Average Imputation," *Mathematics and Statistics*, vol. 11, no. 2, pp. 229-244, 2023.

[7] I. L. Simarmata and I. W. Supriana, "Music Genre Classification Using Random Forest Model," *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, vol. 12, no. 1, pp.83-88, 2023.

[8] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal*, vol. 20, no. 1, pp.3-29, 2020.

[9] K. Gross. (2020, April 6). Tree-Based Models: How They Work (In Plain English!). Data Iku. Available: https://blog.dataiku.com/tree-based-models-how-they-work-in-plain-english

[10] S. Han and H. Kim, "Optimal feature set size in random forest regression," *Appl. Sci.*, vol. 11, no. 8, 2021.

[11] R. B. Nair, G. N. Sundar, and D. Narmadha, "Educational Data Mining: A Systematic Review of Machine Learning Algorithms," in *International Conference on Computer Vision and Internet of Things 2023 (ICCVIoT'23)*, pp. 310-315.
[12] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini Importance?," *Bioinformatics*, vol. 34, no. 21, pp.3711-3718, 2018.
[13] M. Sandri and P. Zuccolotto, "A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees," *Journal of Computational and Graphical Statistics*, vol. 17, no. 3, pp.611-628, 2008.
[14] S. Chakraborty, S. Paul, and M. Rahat-Uz-Zaman, "Prediction of Apple Leaf Diseases Using Multiclass Support Vector Machine," in *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques*, pp.147-151.
[15] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the 20th International Conference on Machine Learning*, Washington, 2003.
[16] G. O. Anyanwu, C. I. Nwakanma, J. M. Lee, and D. S. Kim, "Optimization of RBF-SVM Kernel Using Grid Search Algorithm for DDoS Attack Detection in SDN-Based VANET," *IEEE Internet of Things Journal*, vol. 10, no. 10, pp. 8477-8490, May 2023.
[17] Xiaoxia Zhang, Ziqiao Yu, Yinyin Hu, and Jiao Yang, "Milling Force Prediction of Titanium Alloy Based on Support Vector Machine and Ant Colony Optimization," IAENG International Journal of Computer Science, vol. 48, no.2, pp223-235, 2021.
[18] H. Frick, F. Chow, M. Kuhn, M. Mahoney, J. Silgem, and H. Wickham. (2024, June). Bootstrap Sampling (Online). rsample.tidymodels.org. Available: https://rsample.tidymodels.org/reference/bootstraps.html
[19] Huafeng Xian, and Jinxing Che, "A Variable Weight Combined Model Based on Time Similarity and Particle Swarm Optimization for Short-term Power Load Forecasting," IAENG International Journal of Computer Science, vol. 48, no.4, pp915-924, 2021.
[20] E. Rohaeti and A. Andriyati, "Comparative Study of Predictive Classification Models on Data with Severely Imbalanced Predictors," *JUITA: Jurnal Informatika*, vol. 12, no. 1, pp. 121–129, 2024.
[21] M. Zakariah, S. A. AlQahtani, and M. S. Al-Rakhami, "Machine Learning-Based Adaptive Synthetic Sampling Technique for Intrusion Detection," *Appl. Sci.*, vol. 13, no. 11, 2023.
[22] Ke Zhou, Chunna Zhang, Yang Yu, Shengqiang Cong, and Xiaoping Yue, "Improving SMOTE Technology for Credit Card Fraud Detection Category Imbalance Issues," Engineering Letters, vol. 31, no. 4, pp1780-1785, 2023.
[23] J. Brandt and E. Lanzén, "A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification," Uppsala University, 2021.
[24] Juliana A. A., Samuel O., Pius A. O., Agnieta P., and Sunday O. O., "The Effect of Imbalanced Data and Parameter Selection via Genetic Algorithm Long Short-Term Memory (LSTM) for Financial Distress Prediction," IAENG International Journal of Applied Mathematics, vol. 53, no.3, pp796-809, 2023.
[25] S. S. Bagui, D. Mink, S. C. Bagui, and S. Subramaniam, "Determining Resampling Ratios Using BSMOTE and SVM-SMOTE for Identifying Rare Attacks in Imbalanced Cybersecurity Data," *Computers*, vol. 12, no. 10, 2023.
[26] I. D. Ratih, S. M. Retnaningsih, I. Islahulhaq, and V. M. Dewi, "Synthetic Minority Over-Sampling Technique Nominal Continous Logistic Regression for Imbalanced Data," in *AIP Conf. Proc.*, 2022.
[27] J. R. Barr, M. Sobel, and T. Thatcher, "Upsampling, a comparative study with new ideas," in *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pp. 318-321.
[28] J. Zhang and I. Mani, "kNN approach to unbalanced data distributions: A Case Study Involving Information Extraction," in *Proceedings of Workshop on Learning from Imbalanced Datasets II*, Washington, 2003.
[29] A. Bansal and A. Jain, "Analysis of focused under-sampling techniques with machine learning classifiers," in *2021 IEEE/ACIS 19th International Conference on Software Engineering Research, Management and Applications (SERA)*, pp. 91-96.
[30] Said Marso and Mohamed El Merouani, "Bankruptcy Prediction using Hybrid Neural Networks with Artificial Bee Colony," Engineering Letters, vol. 28, no. 4, pp1191-1200, 2020.
[31] Mar Mar Nwe, and Khin Thidar Lynn, "Effective Resampling Approach for Skewed Distribution on Imbalanced Data Set," IAENG International Journal of Computer Science, vol. 47, no.2, pp234-249, 2020
[32] T. M. Barros, P. A. S. Neto, I. Silva, and L. A. Guedes, "Predictive Models for Imbalanced Data: A School Dropout Perspective," *Educ. Sci.*, vol. 9, no. 4, 2019.