JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 8 (ISSUE 2) DECEMBER 2024: 229-244
https://doi.org/10.21009/JSA.08209

DETERMINATION OF IMPORTANT VARIABLES IN DIVORCE TYPE CLASSIFICATION USING THE RANDOM FOREST METHOD WITH SMOTE

Dania Siregar[1], Bintang Mahesa Wardana[1], Ahmad Syauqi Baihaqy[1], Liswatun Naimah[1], Almira Nindya Putri[1], Qorry Meidianingsih[2], Dini Safitri[3]

[1] *Statistics Study Program, Universitas Negeri Jakarta*
Jln. Rawamangun Muka, East Jakarta, DKI Jakarta, 13220, Indonesia

[2] *Mathematics Education Study Program, Universitas Negeri Jakarta*
Jln. Rawamangun Muka, East Jakarta, DKI Jakarta, 13220, Indonesia

[3] *Communication Science Study Program, Universitas Negeri Jakarta*
Jln. Rawamangun Muka, East Jakarta, DKI Jakarta, 13220, Indonesia

Corresponding author's e-mail: * ***dania-siregar@unj.ac.id***

ABSTRACT

Central Jakarta, a highly strategic area at the heart of Indonesia's capital, serves as the central hub for government activities, historical landmarks, tourism, and high-end shopping. It also provides convenient access to surrounding buffer zones. However, the availability of these facilities does not necessarily translate into the stability of domestic life within the community. This is evidenced by a rising divorce rate in the region since 2017, with a higher proportion of cases initiated by wives compared to husbands. Factors contributing to divorce filings include ongoing disputes, economic challenges, and domestic violence. These issues are closely tied to the demographic and socio-economic profiles of married couples, such as age, occupation, education level, and duration of marriage. This study aims to assess the level of importance of various variables in classifying wife-initiated and husband-initiated divorce cases in Central Jakarta using the Random Forest method. Random Forest, an enhancement of the CART (Classification and Regression Tree) method, incorporates bootstrap aggregating and random feature selection to improve classification accuracy. Due to the imbalance in the dataset, where wife-initiated divorce cases significantly outnumber husband-initiated cases, the SMOTE technique was applied to address this issue. The findings reveal that the plaintiff's age is the most important variable in classifying divorce cases, followed by the defendant's occupation, the defendant's age, and the plaintiff's occupation.

Article History:

Received: 10, December 2024
Revised: 24, December 2024
Accepted: 27, December 2024
Published: 31, December 2024
Available online.

Keywords:

Wife-initiated divorce, Husband-initiated divorce, *Random Forest, SMOTE.*

How to cite this article:
D. Siregar, B.M. Wardana, A.S. Baihaqy, L. Naimah, A.N. Putri, Q. Meidianingsih, D. Safitri, "DETERMINATION OF IMPORTANT VARIABELS IN DIVORCE TYPE CLASSIFICATION USING THE RANDOM FOREST METHOD WITH SMOTE", Journal Statistika dan Aplikasinya, vol. 8, iss. 2, pp. 229 – 244, December 2024

Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

Currently, the development of data analysis for classification modeling is increasingly diverse, one of the classification methods capable of classifying with high accuracy is the Random Forest method. Random Forest is one of the developments of the CART (Classification and Regression Tree) method by applying the bootstrap aggregating (bagging) method and random feature selection [1]. This modification increases the reduction of bagging variance by forming a multitude of mutually independent decision trees and can improve accuracy. Furthermore, Random Forest is also able to provide a measure of the variable importance of a classifier variable, this is certainly very useful for a variety of cases that require information about what variables make a major contribution to a classification result. Random Forest has been widely applied in various fields, including in the medical field which generally involves many variables and requires results with high significance. Barakat et.al [2] in their research using Random Forest successfully determined risk factors for predicting liver fibrosis. Bhagat & Patil [3] in their research found that the use of SMOTE (Synthetic Minority Over-sampling Technique) in Random Forest to address the imbalance in the data set successfully improved classification performance. Polat's research [4] concluded that the combination of SMOTE and Random Forest on Parkinson's disease classification data resulted in better classification predictions reaching 94% than 87% without SMOTE.

Analysis of divorce cases can also utilize the Random Forest, this method can be utilized to predict the classification of the type of divorce lawsuit, namely wife-initiated divorce (divorce lawsuit filed by the wife) or husband-initiated divorce (divorce lawsuit filed by the husband). Furthermore, this method can also be used to identify the variables that play an important role in classifying the type of divorce lawsuit. According to the Minister of Religious Affairs, the divorce rate in Indonesia continues to increase, previously in 2015 there were 398,245 divorce claims, consisting of 114,000 divorce claims filed by husbands and more than 281,000 divorce claims by wives. Meanwhile, in 2017 it increased to a total of 415,898 divorce lawsuits [5] The rise of divorce from a wife's lawsuit or called wife-initiated divorce, warrants serious attention and further study. The studies that have been conducted on divorce have focused on variables that are thought to affect divorce in general, such as those conducted by Handayani [6], which estimates the variables that affect divorce in Central Sulawesi Province using Probit Regression. Sari et.al [7] also focused on creating a system using the Naïve Bayes algorithm to estimate the chances of someone divorcing in Central Aceh Regency using variables such as age of marriage, number of children, and reasons for filing for divorce. Bolhari et.al [8] have also conducted research to determine the factors that influence divorce in Tehran using structured interviews and questionnaires with a qualitative approach. Akter & Begum [9] also explored the factors responsible for divorce among women undergoing divorce proceedings using purposive sampling technique with exploratory research design and qualitative methods. It is thus seen that research on the classification of divorce, namely wife-initiated divorce, and husband-initiated divorce, has not been conducted, while the rise of divorce originating from the wife's lawsuit or wife-initiated divorce continues to increase, especially in Indonesia.

Meanwhile, Jakarta as the economic center of Indonesia is also not immune from this problem, especially Central Jakarta which has the smallest city area, but the divorce rate is also quite high, reaching 1431 divorces with 327 cases of husband-initiated divorce and 1104 cases of wife-initiated divorce in 2016 and continuing to increase even until 2021 [10]. It is evident that this condition requires the serious attention of the government and the public. This study aims to determine important variables that classify the types of divorce that occur in the Central Jakarta area using the Random Forest method with SMOTE. SMOTE is used to address the issue of data imbalance that occurs in many types of divorce cases, namely wife-initiated divorce which is more than husband-initiated divorce. Divorce of a married couple is not inherently negative, it may even prove beneficial for both parties. However, that does not mean divorce should always be the solution to household problems. The discussion of these important variables is a matter of considerable interest, with the hope that the insights gained can be considered by couples facing challenges in their marriages and become material for making policies for the government in protecting the community.

2. METHODS

Material and Data

The data used is secondary data obtained from the Central Jakarta Religious Court regarding the types of divorce cases (wife-initiated divorce and husband-initiated divorce), spouse profiles, and reasons for divorce that occurred in the Central Jakarta area from 2017 to 2019. In this study, the spouse profiles and the reasons for divorce are placed as explanatory variables (X) while the type of divorce case is placed as the response variable (Y). Further explanation of the variables is provided in Table 1. Based on the categorization of the data, it was found that there were 2,752 (75.6%) wife-initiated divorce cases and 886 (24.4%) husband-initiated divorce cases. In these cases, plaintiff means the husband/wife who filed for divorce against their spouse while defendant means the husband/wife who was sued for divorce by their spouse.

Table 1. Research Variables

Symbol	Variables	Data Types
Y	Type of divorce case (wife-initiated divorce or husband-initiated divorce)	Categoric
X_1	Plaintiff's age	Numeric
X_2	Defendant's age	Numeric
X_3	Plaintiff's occupation	Categoric
X_4	Defendant's occupation	Categoric
X_5	Plaintiff's last education	Categoric
X_6	Defendant's last education	Categoric
X_7	Duration of marriage	Numeric
X_8	Ground of divorce	Categoric

Research Method

Random Forest

Random Forest is a widely used ensemble learning method, particularly effective for classification tasks due to its high accuracy, robustness to overfitting, and capability to handle large datasets. Introduced by Breiman in 2001, Random Forest operates by constructing a multitude of decision trees during training and outputs the mode of classes for classification (or mean for regression) as the final prediction. Each tree in the Random Forest is built using a bootstrapped sample of the training data, meaning that each tree is trained on a different subset of the data, selected with replacement [1]. During the construction of each tree, only a random subset of features is considered for splitting at each node, which introduces diversity among the trees and helps to reduce overfitting [11]. The final classification output is determined by aggregating the predictions of all the individual trees, typically through majority voting, where the class that receives the most votes becomes the final prediction [12]. This approach not only enhances predictive performance but also provides an estimate of feature importance, allowing users to identify which variables are most influential in making predictions [13]. The Random Forest mechanism effectively balances bias and variance, making it a powerful tool for handling complex datasets with high-dimensional spaces and imbalanced classes [14]. For instance, recent research in medical applications has shown Random Forest to be highly effective in disease prediction due to its ability to manage high-dimensional data [15]. Moreover, its feature importance scores offer interpretability, allowing researchers to identify which variables contribute most to the classification decision. Comparative studies have consistently shown Random Forest to outperform many traditional classifiers, including logistic regression and single decision trees, particularly on complex and imbalanced datasets. However, despite its strengths, Random Forest can be computationally intensive on large datasets due to the ensemble's size, and its interpretability, while improved relative to other ensemble methods, still lags behind simpler models.

Synthetic Minority Over-sampling Technique

SMOTE, or Synthetic Minority Over-sampling Technique, is a popular approach for addressing class imbalance in machine learning, introduced by Chawla et al. [16] to synthetically generate samples for the minority class, thereby balancing the dataset and improving model performance. In SMOTE, new instances are created by interpolating between existing samples within the minority class, which avoids simply duplicating minority instances and helps prevent overfitting. Since its introduction, SMOTE has inspired numerous variations to improve its effectiveness in different scenarios. For instance, Borderline-SMOTE [17] focuses on samples near the decision boundary, which are more likely to be misclassified, thereby enhancing classifier performance. Further adaptations, such as SMOTEENN and SMOTETomek [18], integrate under-sampling techniques like Edited Nearest Neighbors and Tomek Links to eliminate noise, leading to cleaner datasets that improve classifier robustness. SMOTE has found applications in critical fields with high-stakes imbalanced datasets, such as medical diagnosis [19] and credit scoring [20], showing how it can enhance predictive accuracy in identifying minority class events like diseases or loan defaults. Recent comparative studies, such as those by Fernández et al. [21] and Buda, Maki, & Mazurowski [22], explore SMOTE's performance against other techniques and within deep learning contexts, demonstrating its consistent utility but also revealing the challenges of class imbalance in neural networks. Douzas and Bacao [23] further extend SMOTE's utility by combining it with k-means clustering, making it adaptable to the high-dimensional data often encountered in deep learning. Lastly, SMOTE's integration with ensemble methods, as reviewed by Galar et al. [24], shows how techniques like bagging and boosting can leverage SMOTE's synthetic samples to further improve classifier performance, particularly in highly imbalanced scenarios. Overall, SMOTE remains a cornerstone technique in handling imbalanced datasets, with extensive adaptations and applications across various fields, underscoring its versatility and enduring relevance.

A synthetic sample unit can be written as follows:

$$\boldsymbol{x_{SMOTE\_i}} = \boldsymbol{x_i} + \left(\boldsymbol{x_i} - \boldsymbol{x_{NN\_i}}\right) \cdot b$$

where $\boldsymbol{x_{SMOTE\_i}}$ is a vector of size *k* that represents the predictor variables of the synthetic sample unit, $\boldsymbol{x_i}$ is a vector of size k representing the predictor variables of sample unit *i* drawn from the minority class population, $\boldsymbol{x_{NN\_i}}$ is a vector of size *k* representing the predictor variables of the nearest neighboring sample unit to sample *i*, *b* represents a uniform random variable with values in the interval [0,1], and *k* is the total number of predictor variables in the dataset.

Evaluation Metrics

The confusion matrix is a crucial tool in evaluating the performance of classification models, offering a breakdown of actual versus predicted outcomes across categories. Introduced in the mid-20th century in the field of information retrieval and adapted for machine learning, the confusion matrix allows for more nuanced analysis of classification performance [25]. It is typically structured as a 2x2 matrix for binary classification, comprising four key elements: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These elements provide the foundation for calculating essential metrics like Precision, Recall, and F1-score [26]. Precision, or Positive Predictive Value, is the ratio of correctly predicted positive observations to all predicted positives, offering insight into a model's false-positive rate, which is critical in applications where misclassification of positives carries significant cost [27]. Recall, also known as Sensitivity or True Positive Rate, is the ratio of correctly predicted positives to all actual positives, indicating the model's ability to identify true positives, which is essential in fields like disease screening [26]. The F1-score, the harmonic mean of Precision and Recall, balances the two by equally weighting them, making it ideal for imbalanced datasets by mitigating extremes in either metric [28]. Together, these measures derived from the confusion matrix provide a comprehensive framework for evaluating model performance, particularly in situations where class imbalance is present [29]. Such metrics thus enable practitioners to fine-tune models, adjust thresholds, and compare model efficacy across tasks with varying priorities on Precision and Recall [25]. The macro average F1-score calculates the F1-score for each class independently and averages them equally, making it ideal for balanced datasets or when all classes are equally important. The

weighted average F1-score, however, weights each class's F1-score by its prevalence, providing a better representation of performance on imbalanced datasets. In Python, tools like Scikit-learn allow easy calculation of both metrics, helping evaluate model performance based on specific dataset characteristics. Here are the formulas used to calculate precision, recall, and F1-score [26, 27, 28].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \times 100\%$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \times 100\%$$

$$F1 - Score = 2 \times \frac{Precision\ \times\ Recall}{Precision +\ Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \times 100\%$$

$$Macro\ F1 - score = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

$$Weighted\ F1 - score = \frac{\sum_{i=1}^{N} w_i \cdot F1_i}{\sum_{i=1}^{N} w_i}$$

Mean Decrease Impurity

In Random Forest classification, variable importance is a key feature that helps to identify which predictors contribute most to the model's predictions. Mean Decrease Impurity, also known as Gini Importance, is a widely used measure of variable importance in Random Forest classification. This metric quantifies the contribution of each feature to the overall predictive power of the model by calculating the total decrease in node impurity that occurs when a feature is used to split the data, averaged over all trees in the forest. Specifically, for a given feature *j*, its importance I_j can be calculated using the formula:

$$I_j = \sum_{t=1}^{T} \sum_{m=1}^{M_t} \frac{N_{m,j}}{N_m}\ \Delta G_m(t)$$

where *T* is the number of trees in the forest, M_t is the number of splits in tree *t*, N_m is the number of samples reaching node *m*, $N_{m,j}$ is the number of samples with feature *j* reaching node *m*, $\Delta G_m(t)$ is the decrease in impurity (such as Gini impurity or entropy) at split *m* in tree *t*. The higher the Mean Decrease Impurity score for a feature, the more important it is considered in making predictions. This measure effectively captures the influence of features on the decision-making process within the Random Forest model, aiding in feature selection and interpretation. Breiman [1] first introduced this concept in his seminal work on Random Forests, establishing it as a key technique in machine learning for assessing feature importance.

Data Analysis Stages

The method applied in this research is the Random Forest method which is a form of development of the Classification and Regression Trees (CART) method or more commonly called Decision Tree. Random Forest applies bootstrap aggregating (bagging) and random feature selection methods [1] and is included in a supervised learning algorithm that works by building several uncorrelated Decision Trees, this method can be used for classification or regression [30]. Random forest is more flexible in classifying a new observation than Decision Tree, this method also combines the flexibility and simplicity of Decision Tree to produce a significant increase in accuracy [31].

The data used in this study exhibits an imbalanced class in the response variable (Y), where the number of sample units in a particular class is considerably higher (majority class) compared to other classes (minority class). In this case, the number of cases (sample units) in the 'wife-initiated divorce'

class amounted to 2,752 cases or 75.6% of the total cases, while the number of cases in the 'husband-initiated divorce' class only amounted to 886 cases or only 24.4% of the total cases. Thus the 'wife-initiated divorce' class is defined as the majority class while the 'husband-initiated divorce' class is defined as the minority class. If there is a class imbalance in the response variable (Y) in a Decision Tree (and by extension, Random Forest), then the initial decision node in the Decision Tree will be more prone to prioritize the logic rule formulation that divides the majority class into pure groups and sacrifice the logic rule formulation that divides the minority class [32]. To address this issue, the SMOTE (Synthetic Minority Oversampling Technique) algorithm is applied to increase the number of sample units (in this case) in minority classes so that the number of sample units in minority classes is equivalent to the number of sample units in majority classes through the generation of new data (synthetic data) based on the k-nearest neighbor algorithm [33].

The analysis in this study was carried out entirely through the Python programming language with *random_state* = 2021, and followed the following stages of analysis:

1. The data was prepared and pre-processed with the assistance of the "Pandas" module [34].
    a. Determine the profile and background allegedly related to the type of divorce case according to the data provided by the Central Jakarta Religious Court.
    b. Select the data by discarding sample units with incomplete responses.
    c. Recode the responses so that the data format aligns with the provisions of the data format required in the subsequent stages of analysis.
2. The pre-processed data was then stratified into two mutually exclusive data sets with the assistance of the "Scikit-learn" module [35]. The two data sets are the original training data, comprising 70% of the pre-processed data, and the original testing data, comprising 30% of the pre-processed data.
3. The minority classes in the original training data were oversampled with the SMOTE algorithm to obtain a balanced set of synthesized data, called SMOTE training data, with the help of the "Imbalance-learn" module [36].
4. Perform Random Forest modelling on original training data and SMOTE training data with the help of the "Scikit-learn" module [35].
    a. Searching for the best Random Forest model parameter combination based on an iterative search conducted 10 times with Out-of-bag evaluation through Cross Validation Grid Search on original training data and SMOTE training data. The best Random Forest model parameter combination was searched based on the following three parameters:
        i. max_features: {None, "sqrt"};
        ii. criterion: {"gini", "entropy", "log_loss"};
        iii. class_weight: {None, "balanced"}.
    b. Perform two Random Forest modelling and estimation of explanatory variable importance scores on the original training data and SMOTE training data based on the best Random Forest model parameter combination for each training data found through iterative search.
5. Evaluate the modelling results of the Random Forest model in stage 4(a) and stage 4(b) through analysis of the precision, recall, F1-score, and confusion matrix values based on the class prediction results of the two models on the original test data with the help of the "Seaborn" [37] and "Matplotlib" [38] modules.
6. Assess which explanatory variables are believed to have an important role in the type of divorce case based on the importance score using the mean decrease in impurity.

This section contains data sources, research variables, sampling techniques, data collection methods and data analysis methods. The data analysis method used must be described in detail.

3. RESULTS

Random Forest Trained on an Imbalanced Dataset

Based on the iterative search, it was found that the best Random Forest model parameter combination for unbalanced original training data is as follows:

1. max_features = None;
2. criterion = gini;
3. class_weight = None.

Where the model built based on a combination of parameters on the original training data has a model performance on the original training data as can be seen in Figure 1 and Table 2 as follows:

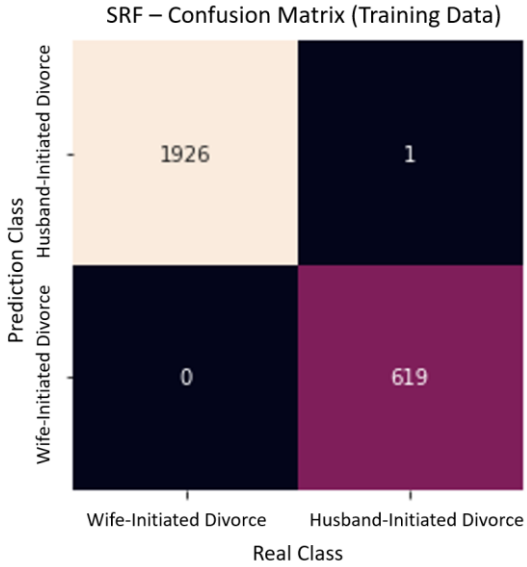

Figure 1. Confusion Matrix for Training Data

Table 2. Performance Metrics of the SRF Model for Training Data

Class	Precision	Recall	F1-score	Unit Sample
Wife-Initiated Divorce	99.948%	100.000%	99.974%	1.926
Husband-Initiated Divorce	100.000%	99.839%	99.919%	620
Macro Average	99.974%	99.919%	99.947%	
Weight Average	99.961%	99.961%	99.961%	

The weighted average F1-score of the resulting training data is 99.961%. Meanwhile, the results of the model's performance in making predictions on the original testing data can be seen in Figure 2 and Table 3 with the weighted average F1-score dropping to 86.287%.

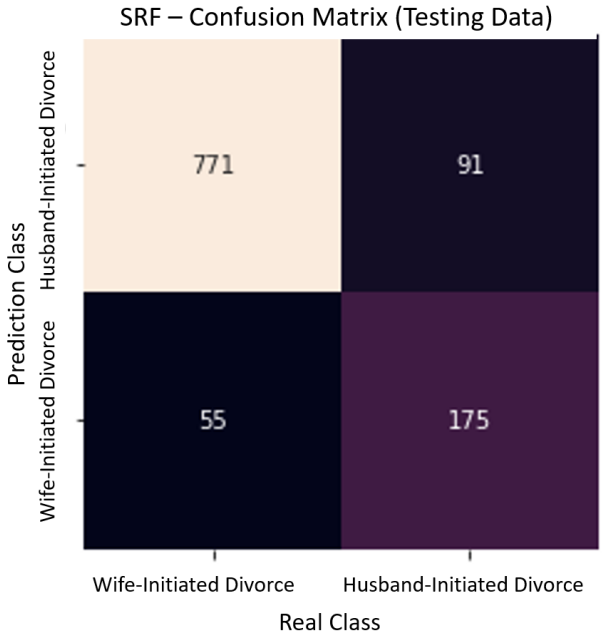

Figure 2. Confusion Matrix for Testing Data

Table 3. Performance Metrics of the SRF Model for Testing Data

Class	Precision	Recall	F1-score	Unit Sample
Wife-Initiated Divorce	89.443%	93.341%	91.351%	826
Husband-Initiated Divorce	76.087%	65.789%	70.565%	266
Macro Average	82.765%	79.565%	80.958%	
Weight Average	86.190%	86.630%	86.287%	

In this model, it is found that of the eight explanatory variables in the data, the explanatory variable that is thought to be the most important in classifying the type of divorce case based on the results obtained in Figure 3 is the Plaintiff's Age (X_1) followed by the Defendant's Occupation (X_4) and the Defendant's Age (X_2). This indicates that, based on results of the Standard Random Forest (SRF) modelling, the type of divorce filed is thought to have the strongest relationship with the defendant's occupation and the age of both parties. Further investigation into this relationship may be possible, but this research is limited to evaluating model performance and finding the most important variables.

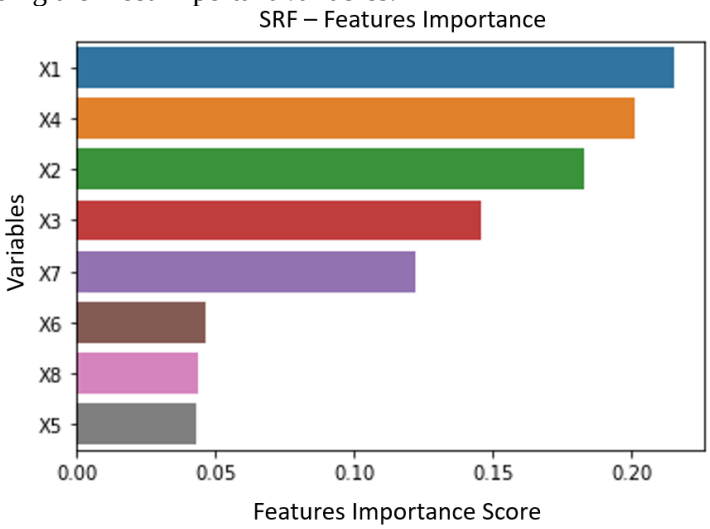

Figure 3. Features Importance of the SRF Model

Random Forest Trained on a Balanced Dataset Obtained Through SMOTE Algorithm

Based on the iterative search, it was found that the best Random Forest model parameter combination for the synthetically balanced SMOTE training data is as follows:

1. max_features = sqrt;
2. criterion = gini;
3. class_weight = balanced.

Where the model built based on the combination of parameters on the SMOTE training data has a model performance on the SMOTE training data as shown in Figure 4 and Table 4 below.

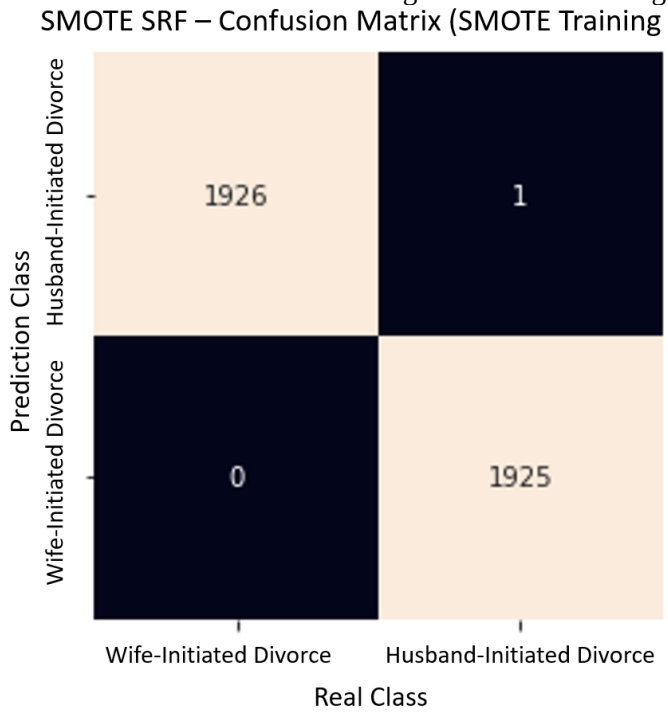

Figure 4. Confusion Matrix for SMOTE Training Data

Table 4. Performance Metrics of the SMOTE SRF Model for Training Data

Class	Precision	Recall	F1-score	Unit Sample
Wife-Initiated Divorce	99.948%	100.000%	99.974%	1.926
Husband-Initiated Divorce	100.000%	99.948%	99.974%	1.926
Macro Average	99.974%	99.974%	99.947%	
Weight Average	99.974%	99.974%	99.974%	

The weighted average F1-score of the training data using SMOTE is 99.974%. Meanwhile, the results of the model's performance in making predictions on the original testing data can be seen in Figure 5 and Table 5 with the weighted average F1-score dropping to 86.349%.

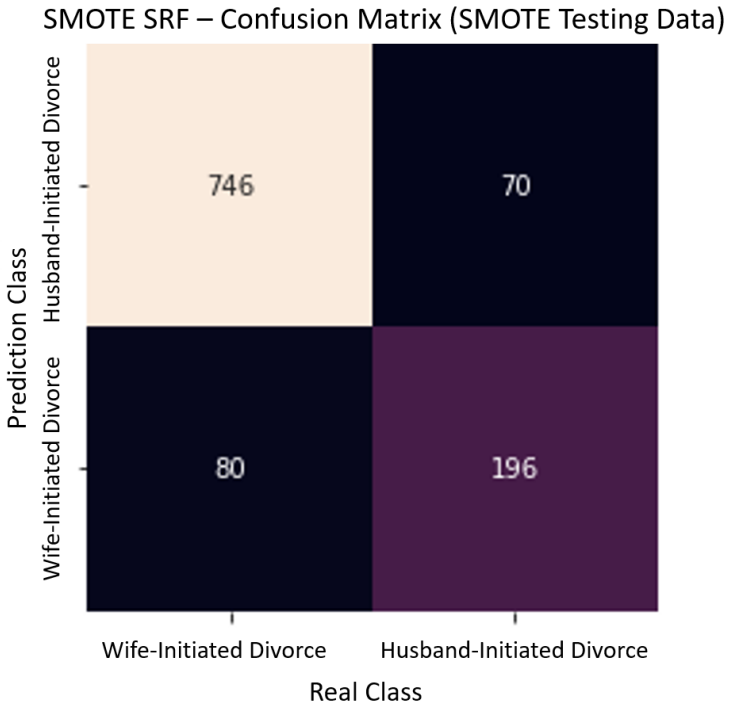

Figure 5. Confusion Matrix

Table 5. Performance Metrics Model of the SMOTE SRF Model for Testing Data

Class	Precision	Recall	F1-score	Unit Sample
Wife-Initiated Divorce	91.422%	90.315%	90.865%	826
Husband-Initiated Divorce	71.014%	73.684%	72.325%	266
Macro Average	81.218%	81.999%	81.595%	
Weight Average	86.451%	86.264%	86.349%	

In this model, it is found that of the eight explanatory variables in the data, the explanatory variable that is thought to be the most important in classifying the type of divorce case based on the results obtained in Figure 6 is the Plaintiff's Occupation (X_3) followed by the Plaintiff's Age (X_1) and the Defendant's Age (X_2). This indicates that, based on results of the SMOTE Standard Random Forest (SMOTE SRF) modelling, the type of divorce filed is thought to have the strongest relationship with the plaintiff's occupation and the age of both parties.

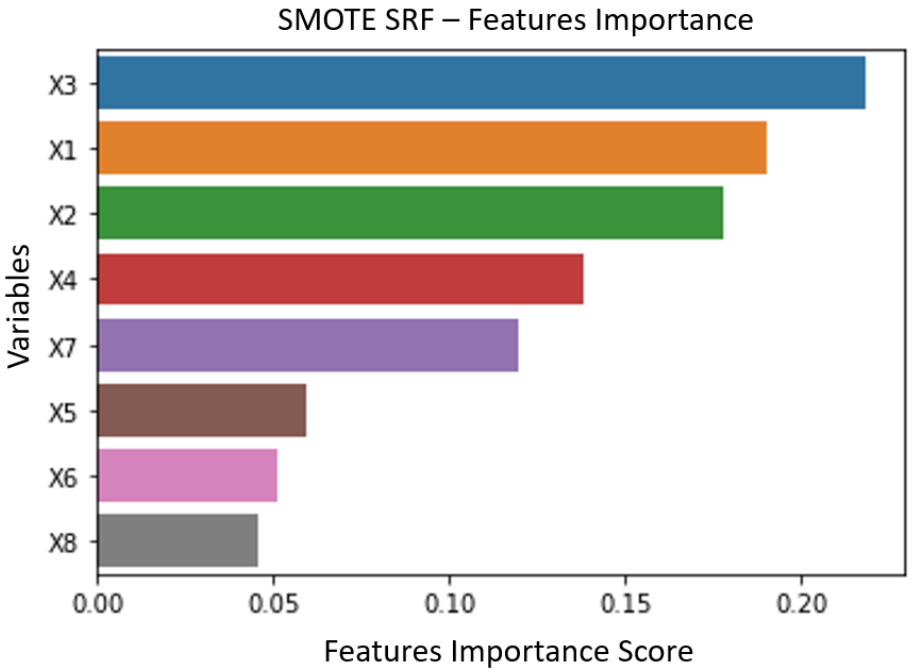

Figure 6. Features Importance of the SMOTE SRF Model

4. DISCUSSION

Differences Between the Two Random Forest Models

The data utilized for the analysis of divorce is inherently imbalanced, necessitating the use of a Weighted Average as the optimal measurement for evaluating model performance. The Weighted Average results for both models indicate that there is generally no improvement between the two models (SRF and SMOTE SRF). This indicates that the imbalance in the data does not significantly affect the initial Random Forest node. However, it can be seen that the modeling results on the SMOTE training data exhibit a more balanced proportion of precision and recall than the modeling results on the original training data. The modeling results on both test data has their own advantages. In consideration of the study's objectives pertaining to the influence of data imbalance, the modeling results on the SMOTE test data (SMOTE training data) are more suitable for utilization due to their stability in the proportion between precision and recall.

In the analysis of the most important explanatory variables for the classification of divorce types, it can be seen that between the two models there are five explanatory variables that are considered important, where the difference between the two models lies only in the placement of the order of the most important variables. Thus, based on the results of both modeling, which are respectively built on unbalanced data and balanced data, the most important explanatory variables for the classification of the proposed divorce type are the defendant's age and the plaintiff's age, the occupation of both parties followed by the duration of marriage.

Data Exploration of the Model's Most Important Variables

Based on the results of both Random Forest (SRF and SMOTE SRF) modeling, it is known that there are four explanatory variables that are thought to be important in classifying the Type of Divorce Case (Y). The four variables are Plaintiff's Age (X_1), Defendant'Age (X_2), Plaintiff's Occupation (X_3), and Defendant's Occupation (X_4). The following will explore the data on these four explanatory variables to further investigate these four variables.

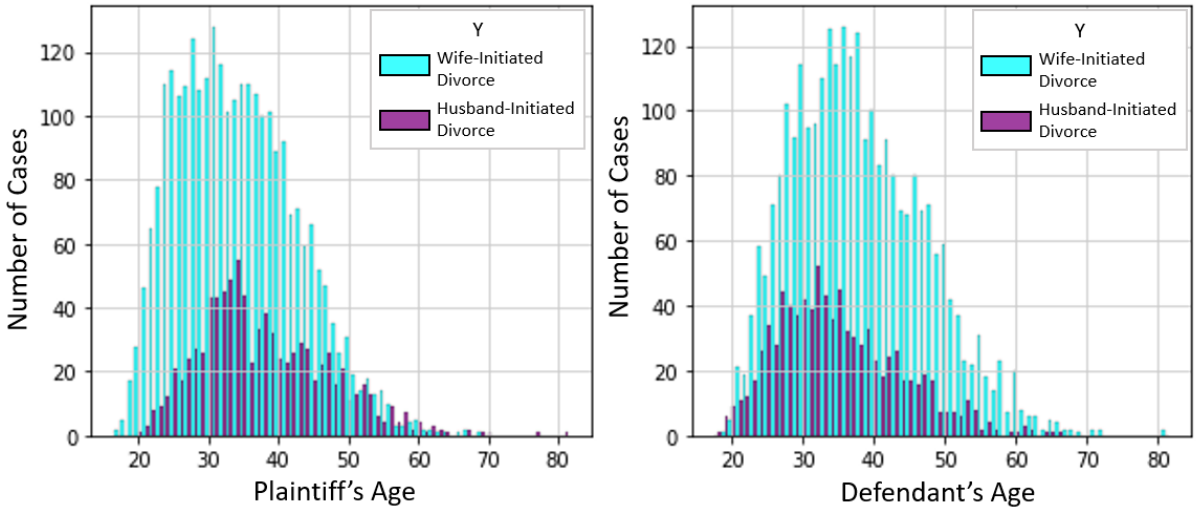

Figure 7. Distribution of Plaintiff's Age (X_1) and Defendant's Age (X_2)

Table 6. Distribution of Plaintiff's Age (X_1) and Defendant's Age (X_2)

Variable	Class	Mean	SD	Min	Q1	Q2	Q3	Max
	Wife-Initiated Divorce	34.375	8.632	17	28	34	40	69
X_1	Husband-Initiated Divorce	37.898	9.211	20	31	36	44	81
	Macro Average	38.106	9.398	19	31	37	44	81
X_2	Weight Average	34.913	8.715	18	28	34	40	66

Figure 7 and Table 6 show the distribution of divorce cases based on the plaintiff's age (X1) and the defendant's age (X2) in two categories of divorce. The figures present two types of divorce cases, namely Wife-Initiated Divorce (in light blue) and Husband-Initiated Divorce (in dark purple). Wife-Initiated Divorce is a divorce that is generally initiated by the wife, while Husband-Initiated Divorce is initiated by the husband. Analysis of these graphs reveals a consistent pattern regarding the age of the parties involved in the divorce. In the first graph, which depicts the plaintiff's age, most plaintiffs fall within the age range of 25-40 years old, with a peak at 30-35 years old. This is true for both Wife-Initiated divorce and Husband-Initiated divorce, although the number of Wife-Initiated divorce cases is much more dominant than Husband-Initiated Divorce. Divorce cases initiated by women (Wife-Initiated Divorce) appear to be more common and dominate in almost all age groups. The number of Wife-Initiated Divorce cases reached more than 100 cases at the peak of their age, while Husband-Initiated Divorce cases were far fewer with a similar distribution pattern. The second graph, depicting the defendant's age, shows a similar pattern to the plaintiff graph. Most defendants in divorce cases are also in the 30-40 year age range. As with the plaintiffs, Wife-Initiated Divorce cases continue to dominate the number of defendant cases across all age groups. On the other hand, cases of Husband-Initiated Divorce continue to show a smaller number of cases compared to Wife-Initiated Divorce, with the peak of cases also in the same age range, which is around 30-40 years old.

What is interesting about the two graphs above is that divorce is more common in the productive years, with the number of cases dropping significantly after the age of 50. This can be interpreted that couples in their young to productive years are more vulnerable to divorce than older couples. Overall, the distribution of divorce cases by age, for both plaintiffs and defendants, shows the dominance of Wife-Initiated Divorce cases over Husband-Initiated divorce cases, as well as the trend that divorces are most prevalent in the 30-40 year age bracket. This suggests that marital problems tend to peak at productive ages, while at older ages, divorce becomes less common.

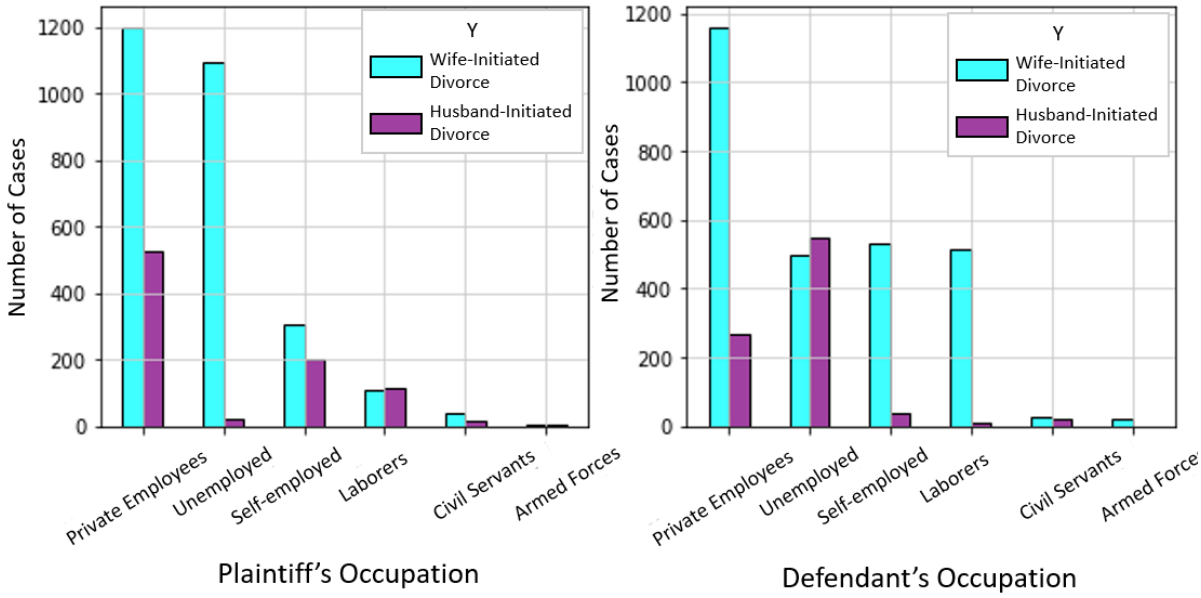

Figure 8. Distribution of Plaintiff's Occupation (X_3) and Defendant's Occupation (X_4)

Figure 8 shows a comparison of the number of divorce cases by occupation of the plaintiff and respondent in two types of divorce, namely Wife-Initiated Divorce (in light blue) and Husband-Initiated Divorce (in dark purple). The left graph shows the plaintiff's occupation, while the right graph shows the defendant's occupation. On the graph of the plaintiff's occupation, it can be seen that the Private Employee category dominates the number of cases, especially for the Wife-Initiated Divorce, with 1200 cases. This shows that women who work as private employees have a very large proportion in filing for divorce. Meanwhile, the non-working category also showed a significant number of cases, with more than 1000 cases, meaning that non-working wives also accounted for a very large proportion of divorce cases. As for Husband-Initiated Divorce cases, although the number of cases was smaller, the distribution pattern was slightly different, with plaintiffs or husbands from the categories of Private Employees, Self-Employed and Laborers filing more cases than other occupational categories. Other occupational categories such as Unemployed, Civil Servants and Armed Forces had a much smaller number of cases. In the graph of the respondent's occupation, the Private Employees category dominated the number of cases for both Wife-Initiated Divorce and Husband-Initiated Divorce. Defendants from this category contributed to more than 1000 cases of Wife-Initiated Divorce and more than 300 cases of Husband-initiated divorce. In addition, the Unemployed category also accounted for a significant number of cases, with over 400 cases for Wife-Initiated Divorce and over 500 cases for Husband-Initiated Divorce. Defendants from the Self-Employed and Laborers categories also dominated in the Wife-Initiated Divorce cases, with more than 500 cases. Meanwhile, respondents from the Civil Servant and Armed Forces categories appear to be less involved in divorce cases than the other categories. Overall, these two graphs show that divorce cases are more prevalent among the Unemployed and Private Employees, both as plaintiffs and defendants. In addition, Wife-Initiated Divorce cases were far more prevalent than Husband-Initiated Divorce cases in almost all occupational categories, indicating that women were more likely to file for divorce, especially among the Unemployed and Private Employees.

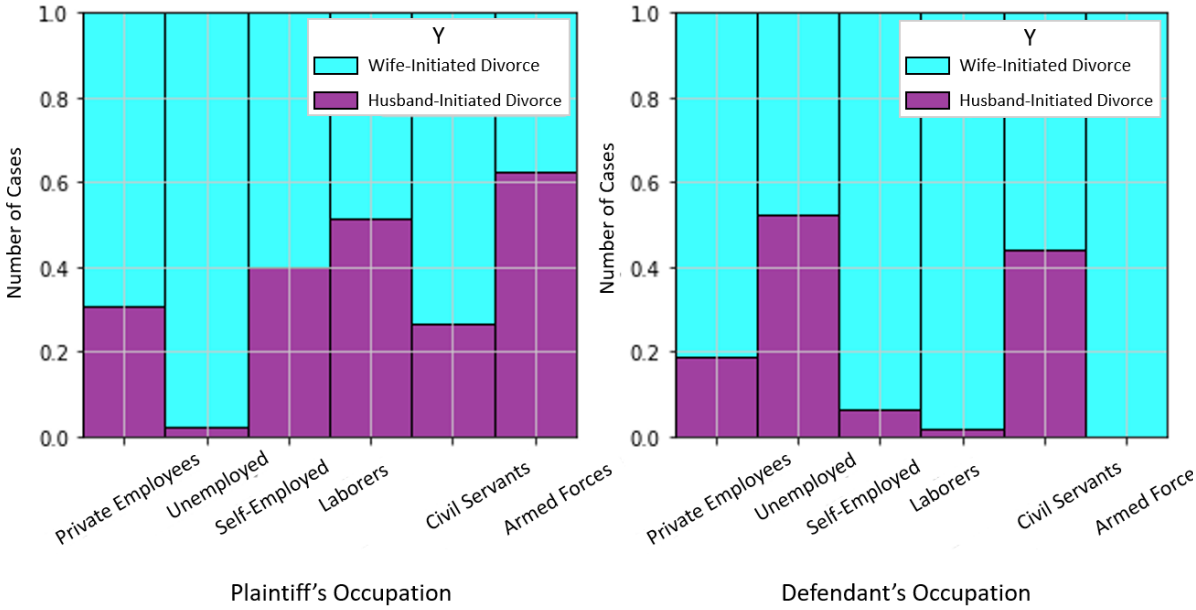

Figure 9. Proportion of Plaintiff's Occupation (X_3) and Defendant's Occupation (X_4)

The graph in Figure 9 displays the proportion of divorce cases categorized by the occupations of the plaintiff and the defendant across two types of divorce under Islamic law: Wife-Initiated Divorce, denoted in light blue, and Husband-Initiated Divorce, denoted in dark purple. The left-hand graph, representing the plaintiff's occupation, shows that in the categories of Unemployed, Private Employees, Self-Employed, and Civil Servants, Wife-Initiated divorce accounts for a higher proportion compared to Husband-Initiated Divorce. However, in the categories of Laborers and Armed Forces, the proportion of Husband-Initiated Divorce is higher. The right-hand graph, which illustrates the defendant's occupation, indicates that the proportion of defendants in the categories of Armed Forces, Laborers, Self-Employed, Private Employees, and civil servants is higher in Wife-Initiated Divorce compared to Husband-Initiated Divorce. The only exception is the Unemployed category, where the proportion of Husband-Initiated divorce slightly exceeds that of Wife-Initiated Divorce. This finding suggests that the proportion of unemployed wives being divorced by their husbands is greater than that of unemployed husbands being divorced by their wives. Additionally, the graph reveals an interesting phenomenon. Despite being unemployed, wives tend to initiate divorce more frequently, as shown by the dominance of Wife-Initiated Divorce, compared to unemployed husbands who initiate divorce, reflected in Husband-Initiated Divorce. Moreover, among defendants in the Armed Forces category, Wife-Initiated Divorce shows a higher proportion than Husband-Initiated divorce, indicating that husbands in the Armed Forces are more likely to be sued for divorce by their wives.

5. CONCLUSION

The study demonstrates that while data imbalance does not significantly impact the initial Random Forest model's performance, utilizing SMOTE improves the balance between precision and recall, making it more suitable for applications requiring stable evaluation metrics. Both models identify five key explanatory variables for divorce type classification: plaintiff's age, defendant's age, their occupations, and the duration of marriage, highlighting consistent factors across imbalanced and balanced data. These findings underscore the importance of addressing data imbalance for enhanced model robustness while affirming the critical variables influencing divorce classifications.

6. ACKNOWLEDGMENT

The authors extend their gratitude to the Faculty of Mathematics and Natural Science, Universitas Negeri Jakarta which has funded this research.

7. REFERENCES

[1] Breiman, L., "Random forests," *Machine learning*, 45, 5-32, 2001.

[2] Barakat, N. H., Barakat, S. H., & Ahmed, N., "Prediction and staging of hepatic fibrosis in children with hepatitis c virus: A machine learning approach," *Healthcare Informatics Research,* 25(3), 173-181, 2019.

[3] Bhagat, R. C., & Patil, S. S., "Enhanced SMOTE algorithm for classification of imbalanced big-data using random forest," *in 2015 IEEE international advance computing conference (IACC)*, pp. 403-408, June 2015.

[4] K. Polat, "A Hybrid Approach to Parkinson Disease Classification Using Speech Signal: The Combination of SMOTE and Random Forests," *in 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pp. 1-3, 2019.

[5] Adhi, M. D., "Menteri Agama: Angka Perceraian di Indonesia Meningkat," *Kumparan*, Dec. 07, 2018. [Online]. Available: https://kumparan.com/kumparannews/menteri-agama-angka-perceraian-di-indonesia-meningkat-1544179658506355359/full

[6] Handayani, L., "Regresi Probit untuk Analisis Variabel-Variabel yang Mempengaruhi Perceraian di Sulawesi Tengah," *Jurnal Aplikasi Statistika & Komputasi Statistik,* 12(1), 13-21. 2020.

[7] Sari, D. L., Saputra, M., & Gemasih, H., "Penerapan Data Mining Dalam Proses Prediksi Perceraian Menggunakan Algoritma Naive Bayes Di Kabupaten Aceh Tengah," *Jurnal Teknik Informatika dan Elektro,* 4(1), 23-35, 2022.

[8] J. Bolhari, Fatemeh, R. Z., Nasrin, A., M. M. Naghizadeh, Hajar, P., & Mehdi S., "The Survey of Divorce Incidence in Divorce Applicants in Tehran," *Journal of Family and Reproductive Health*, 6(3), 129-137, Sep. 2012.

[9] Akter, M., & Begum, R., "Factors for divorce of women undergoing divorce in Bangladesh," *Journal of Divorce & Remarriage*, 53(8), 639-651, 2012.

[10] Badan Pusat Statistik Provinsi DKI Jakarta, 2021. [Online]. Available: https://jakarta.bps.go.id/indicator/27/602/1/nikah-talak-dan-cerai-serta-rujuk-di-provinsi-dki-jakarta.html

[11] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J., "Random forests for classification in ecology," *Ecology*, 88(11), 2783-2792, 2007. doi:10.1890/07-0539.1

[12] Biau, G., & Scornet, E., "A random forest guided tour," *Test*, 25(2), 197-227, 2016. doi:10.1007/s11749-016-0481-7

[13] Liaw, A., & Wiener, M., "Classification and regression by randomForest," *R News*, 2(3), 18-22, 2002.

[14] Zhang, H., & Singer, B., "Recursive partitioning and applications," *Springer Series in Statistics*, 2010.

[15] Chen, T., & Guestrin, C., "Xgboost: A Scalable Tree Boosting System," *in Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, August 2016.

[16] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 16, 321-357, 2002. doi:10.1613/jair.953

[17] Han, H., Wang, W.-Y., & Mao, B.-H., "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," *in Proceedings of the 2005 International Conference on Intelligent Computing*, 3644, 878-887, 2005. doi:10.1007/11538059_91

[18] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C., "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29, 2004. doi:10.1145/1007730.1007735

[19] Blagus, R., & Lusa, L., "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, 14(1), 106, 2013. doi:10.1186/1471-2105-14-106

[20] Brown, I., & Mues, C., "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications*, 39(3), 3446-3453, 2012. doi:10.1016/j.eswa.2011.09.033

[21] Fernández, A., et al., "Learning from imbalanced data sets: Metrics and strategies toward accurate performance evaluation," *IEEE Transactions on Neural Networks and Learning Systems*, 29(11), 5042-5059, 2018. doi:10.1109/TNNLS.2017.2771290

[22] Buda, M., Maki, A., & Mazurowski, M. A., "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, 106, 249-259, 2018. doi:10.1016/j.neunet.2018.07.011

[23] Douzas, G., & Bacao, F., "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, 465, 1-20, 2018. doi:10.1016/j.ins.2018.06.056

[24] Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F., "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics*, 42(4), 463-484, 2012. doi:10.1109/TSMCB.2011.2168600

[25] Provost, F., & Kohavi, R., "Glossary of terms," *Machine Learning*, 30, 271-274, 1998.

[26] Sokolova, M., & Lapalme, G., "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, 45(4), 427-437, 2009. doi:10.1016/j.ipm.2009.03.002

[27] Powers, D. M. W., "Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness, and correlation," *Journal of Machine Learning Technologies*, 2(1), 37-63, 2011.

[28] Chinchor, N., "MUC-4 evaluation metrics," *in Proceedings of the 4th conference on Message understanding*, pp. 22-29, Association for Computational Linguistics, 1992.

[29] He, H., & Garcia, E. A., "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284, 2009.

[30] Brownlee, J., *Master Machine Learning Algorithms Discover How They Work and Implement Them from Scratch (1.12).* Machine Learning Mastery, 2017.

[31] Hastie, T., Tibshirani, R., & Friedman, J., *The Elements of Statistical Learning (2nd ed.)*. Springer, 2009.

[32] Jordan, J., *Learning from Imbalanced Data*, 2018. [Online]. Available: https://www.jeremyjordan.me/imbalanced-data/ [Accessed: November 15, 2024].

[33] Chawla, N.V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P., "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research, 16*, 321-357, 2002.

[34] McKinney, W., "Data structures for statistical computing in python" *in Proceedings of the 9th Python in Science Conference*, pp. 51–56, 2010.

[35] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., & others, "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research 12*, 2825–2830, 2011.

[36] Lemaître, G., Nogueira, F., & Aridas, C. K., "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research, 18*(1), 559-563, 2017.

[37] Waskom, M. et al., mwaskom/seaborn: v0.8.1, Zenodo, 2017. https://doi.org/10.5281/zenodo.883859

[38] Hunter, J. D., "Matplotlib: A 2D graphics environment," *Computing in Science Engineering, 9*(3), 90–95, 2007.