TIN: Terapan Informatika Nusantara
Vol 3, No 8, January 2023, page 286-293
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v3i8.4124

The Application of The Neighborhood Cleaning Rule in Conjunction with Random Forest, K-Fold Cross-Validation, and Grid Search for Addressing Imbalanced Datasets

Laila Qadrini[1,*], Muh Hijrah[1], Laelatul Hikmah[2], Handayani[1]

[1] Statistics Department, FMIPA, Universitas Sulawesi Barat, Sulawesi Barat, Indonesia
[2] Statistics Department, FMIPA, Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang, Semarang, Indonesia
Corresponding author's e-mail: laila.qadrini@unsulbar.ac.id

Abstract−Finding a model that explains and separates data classes is the process of classification in data mining, which is used to guess the class of an item with an unknown class. Numerous strategies have been developed since categorization can be applied in a wide range of applications. But a common issue with classification is class imbalance. Data predictability suffers as a result of the issue of unbalanced classes. There are typically not an equal number of examples in each class in real-world categorization datasets. Class imbalance is not a problem when there are not significant differences in how the classes are distributed. Due to class imbalance, prediction models may skew in favor of the majority class, with the minority class contributing little to the model. One often used strategy for addressing class imbalance is the resampling technique. This study's objective is to put the Resampling Algorithm into practice. Neighborhood Cleaning Rule Random Forest K-Fold Tune Grid Search was carried out on a dataset that includes cases of Low Birth Weight Infants (BBLR) in Majene Regency and breast cancer diagnoses, which was posted on the UCI website. The Neighborhood Cleaning Rule (NCL), a data processing method, eliminates noise or other disturbances from datasets used for modeling or analysis. The F1-Score, G-Mean, Accuracy, and Sensitivity values from the model are good.

Keywords: NCL; BBLR; Random Forest; Kfold; Tune Grid Search

1. INTRODUCTION

Indonesia, along with other nations, has demonstrated its dedication to attaining the Sustainable Development Goals (SDGs), which encompass a specific objective of lowering the Neonatal Mortality Rate (NMR) to a minimum of 12 per 1,000 live births by the year 2030. The Neonatal Mortality Rate (NMR) is a measure that quantifies the number of newborns who pass away within the initial 28 days of life, presented as a ratio per 1,000 live births (Bappenas, 2020). According to the findings of the Indonesian Demographic and Health Survey (IDHS) conducted in 2017, the Neonatal Mortality Rate (NMR) observed during the preceding five-year period amounted to 15 deaths per 1,000 live births. This indicates that almost 1 in every 67 newborns experienced mortality within the initial month of their existence (Kemenkes, 2019). Low birth weight is identified as the principal factor contributing to neonatal death (Astuti & Lenti, 2021.).

According to the World Health Organization (WHO), low birth weight (BBLR) is characterized by infants who are born with a weight below 2500 grams. According to the IDHS 2017 findings, a total of 94 percent of live births throughout the preceding five-year period were reported to have birth weight information available. Additionally, it was observed that 7 percent of these births were classified as having low birth weight. Classification in the field of Data Mining refers to the systematic procedure of identifying and constructing a model that effectively elucidates and distinguishes several classes of data. The primary objective of this method is to accurately forecast the class of an item whose class is unknown. Classification is a widely applicable concept that has been the subject of extensive algorithmic development. Nevertheless, the issue of class imbalance frequently arises in the context of classification (Ihfa & Harsanti, 2020).

The presence of imbalanced class problems has a detrimental impact on the accuracy of data predictions. The issue of imbalanced classes has emerged as a significant obstacle for the effectiveness of numerous classification methods (Siringoringo, 2018). The concept of class imbalance pertains to a situation in which there exists an unequal distribution or substantial disparity in the number of instances across several classes (Pangestika et al., 2021). In practice, it is common for classification datasets to exhibit imbalanced class distributions, wherein the number of examples in each class is not equal. Nevertheless, the issue of imbalance does not arise when there is not a significant disparity in class ratios. The issue of imbalance arises as a concern when the presence of a substantial ratio significantly impacts the outcomes of evaluations (Qadrini et al., 2022). The aforementioned phenomenon has the potential to result in classification conclusions that are less than optimum (Wasono, 2022).

These concerns are present in almost all datasets. The issue of class imbalance can provide challenges if left unattended, since it often leads to biased models that prioritize the majority class, thereby reducing the significance of minority classes. The Resampling technique is a frequently employed algorithm for addressing imbalanced classes (Lestari et al., 2020). Resampling strategies encompass the process of achieving equilibrium in the initial dataset by modifying sample counts across distinct classes through the utilization of diverse sampling algorithms. The aforementioned balanced data is thereafter utilized for training purposes through the implementation of classification algorithms. Resampling techniques can be categorized into three main approaches: Oversampling, Undersampling, and Hybrid.

The Neighborhood Cleaning Rule (NCL) method is employed as a means to tackle the issue of class imbalance within datasets. The NCL algorithm addresses the issue of imbalanced data by employing a resampling technique that

TIN: Terapan Informatika Nusantara
Vol 3, No 8, January 2023, page 286-293
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v3i8.4124

involves the removal of certain instances from the majority class, specifically those examples that are located in close proximity to the boundary separating the majority and minority classes. The primary objective of NCL is to improve the performance of the model on minority classes by removing examples that have the potential to create noise or confusion into the model.

Prior studies have involved the utilization of Machine Learning techniques to classify the likelihood of Low Birth Weight incidents in Indonesia. The results indicate that the utilization of resampling strategies in classification modeling on imbalanced data and big datasets might enhance classification accuracy, particularly for minority classes. This may be observed through the notably elevated sensitivity levels in comparison to the unaltered dataset (prior to resampling). In addition, it was shown that the random forest model demonstrated superior performance compared to the other four classification models in terms of sensitivity, specificity, F1-Score and G-mean,

Moreover, a study conducted on the categorization of instances of Low Birth Weight (BBLR) in newborns through the application of the Learning Vector Quantization (LVQ) technique revealed that the system attained a mean accuracy of 60.5% when employing optimal values for the learning rate (0.1), learning rate decrement (0.1), and maximum epoch. In the context of k-fold cross-validation, the maximum accuracy attained was 58.3%, while the mean accuracy was calculated to be 46.85% (Suryani Agustin et al., 2019). The author intends to apply NCL resampling and employ random forest cross-validation to categorize unbalanced data in instances of low birth weight (BBLR) occurrences in Kabupaten Majene, based on prior research. Furthermore, the author employs simulation data modeling techniques in order to ensure methodological consistency, utilizing Breast Cancer data obtained from the UCI dataset. The process of grid search cross-validation involves the automatic validation of each combination of models and hyperparameters.

2. RESEARCH METHODS

2.1 Neighbourhood Cleaning Rule (NCL)

One of the undersampling strategies for addressing imbalanced class distribution, involving the reduction of data by cleaning, was identified by J. Laurikkala. One notable feature of NCL is its strong emphasis on the quality of data removal, encompassing both data reduction and data cleaning processes. The data cleaning procedure is designed to address samples belonging to both the majority and minority classes. The theory of NCL is primarily grounded in the utilization of one-sided selection (OSS), which is a technique employed for the purpose of instance-based data reduction in order to achieve class balance. The primary objective of this approach is to meticulously minimize extraneous information. In the NCL framework, the data cleaning procedure is implemented independently for both the majority and minority samples. The NCL framework employs the Edited Nearest Neighbor (ENN) technique as a means of data cleansing inside the dominant class (Choirunnisa, 2019).

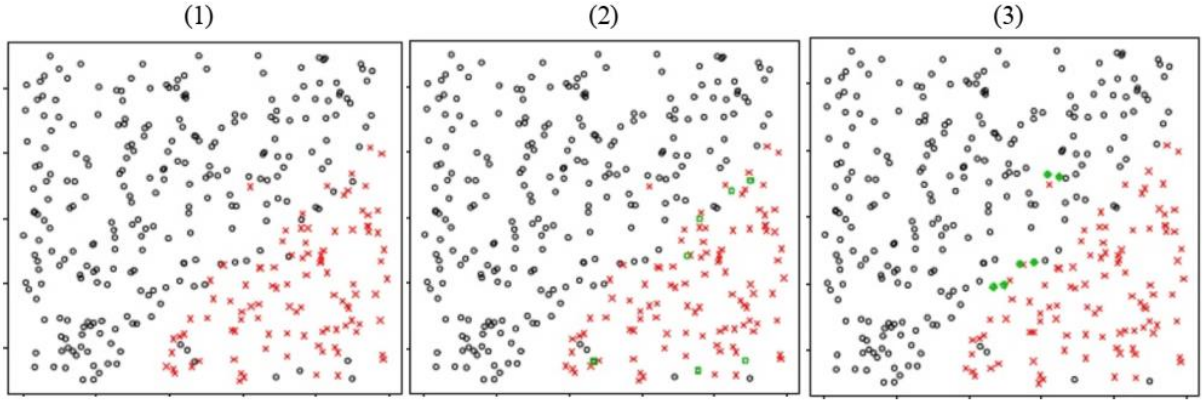


Figure 1. Illustration of data: (1) Original, (2) After ENN process, (3) After NCL process

Figure 1 depicts the ENN and NCL processes, wherein the majority class is represented by black dots, the minority class is represented by red dots, and deleted dots are visually represented in green.

2.2 The Random Forest

The Random Forest algorithm is a machine learning technique that combines many decision trees to make predictions or classifications. It is known for its ability to handle complex datasets and reduce overfit. The Random Forest (RF) algorithm utilizes a recursive binary splitting technique to arrive at terminal nodes inside a tree structure, which is built upon classification and regression trees. The Random Forest algorithm offers various advantages, such as its capacity to produce relatively low error rates, excel in classification tasks, easily handle enormous training datasets, and effectively estimate missing data. The approach generates multiple autonomous trees by employing bootstrap sampling from the training samples and input variables at each node to randomly select subsets. The term refers to a concept or

TIN: Terapan Informatika Nusantara
Vol 3, No 8, January 2023, page 286-293
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v3i8.4124

phenomenon that is being explored or The K-Fold Cross-Validation Algorithm is a widely used technique in machine learning for evaluating the performance of a predictive model (Devella et al., 2020).

It involves partitioning the available data into K The K-Fold Cross-Validation algorithm involves the partitioning of data into k equally sized subsets. During the course of the procedure, a single partition is chosen for the purpose of training, whereas the remaining partitions are utilized for validation. The aforementioned technique is iterated k times, guaranteeing that each partition is utilized for training precisely once. The aggregate error is determined by adding together the individual errors of each of the k operations.

2.3 Hyperparameter Tuning

Hyperparameter tuning is a crucial aspect in machine learning and statistical modeling. It involves the process of selecting the optimal values for hyperparameters, which are parameters that are not learned
The optimization of machine learning (ML) algorithm performance is heavily influenced by the process of hyperparameter tuning (Lujan-Moreno et al., 2018). The determination of hyperparameter values is independent of the data and is established prior to the commencement of the model's learning phase. Hyperparameters are adjustable parameters that exert influence on the output of a model. The methodology known as Grid Search Cross-Validation encompasses the combined utilization of Grid Search and Cross-Validation methodologies. The process involves conducting a systematic examination and verification of every possible combination of models and hyperparameters, proceeding through each combination individually in an iterative manner (Nugraha & Sasongko, 2022).

2.4 Evaluation of the Model

The process of choosing a suitable performance metric for the evaluation of an algorithm is a crucial step. When a classifier is trained with imbalanced data, it has the potential to demonstrate elevated accuracy rates, despite the presence of bias towards the majority class. The appropriate selection of performance measures facilitates the effective evaluation of the algorithm's ability to adapt. The main goal is to optimize the number of True Positives (TP) and True Negatives (TN) while decreasing the occurrence of False Negatives. The measure of Accuracy is a comprehensive indicator of the classifier's performance. However, it may yield deceptive results in the context of imbalanced data, since it tends to prioritize the majority class. Additional performance criteria encompass Recall/Sensitivity, which quantifies the accuracy of the positive class, and Specificity, which quantifies the accuracy of the negative class. Precision, as an additional criterion for evaluating performance, measures the accuracy of the model. According to Turlapaty, classifiers that exhibit high precision values are indicative of effective performance (Turlapati & Prusty, 2020).

In addition to the aforementioned performance measures, combined performance metrics are employed to effectively manage the trade-off between False Positive (FP) and False Negative (FN) rates, exemplified by the F1-score. The F1 score quantifies the equilibrium between precision and sensitivity. A higher F1 score indicates improved accuracy in minority classes. The F1-score is equal to 0 in cases where both precision and sensitivity have values of 0. This study will utilize evaluation metrics that have been employed in prior studies conducted Erlin (Erlin et al., 2022). The Confusion Matrix is a widely utilized tool for assessing Accuracy, Precision, and Recall. The Confusion Matrix is a matrix-based measuring tool that is commonly utilised to assess the classification accuracy of classes based on a given method (Qadrini L et al., 2021). The format of the Confusion Matrix will be displayed in Table 1. Nevertheless Nevertheless, the primary emphasis will be placed on the evaluation metrics of Accuracy, Sensitivity, Specivicity, F1-score, and G-Mean, as represented by equations (1), (2), (3),(4) and (5) respectively.

Table 1. The Confusion Matrix Format for Two Classes

Confusion Matrix		Actual	
		TRUE	FALSE
Prediction	TRUE	TP (True Positive) Correct result	FP (False Positive) Unexpected result
	FALSE	FN (*False Negative*) Missing result	TN (True Negative) Correct absence of result

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \quad (1)$$

$$\text{Re}call\ /\ Sensitivity = \frac{tp}{tp + fn} \quad (2)$$

$$Specivicity = \frac{tn}{tn + fp} \quad (3)$$

$$F1 - Score = \frac{2(recall\ precision)}{recall + precision} \quad (4)$$

$$G - Mean = \sqrt{Sensitivity + Specivity} \quad (5)$$

TIN: Terapan Informatika Nusantara
Vol 3, No 8, January 2023, page 286-293
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v3i8.4124

2.5 Steps for Research

The Resampling NCL Random Forest K-Fold Tune Grid Search classification technique will be used for this study's analysis. This algorithm's technique will be used to process the data, and the accuracy results will be assessed. The steps for this stage's data testing are broken down below.

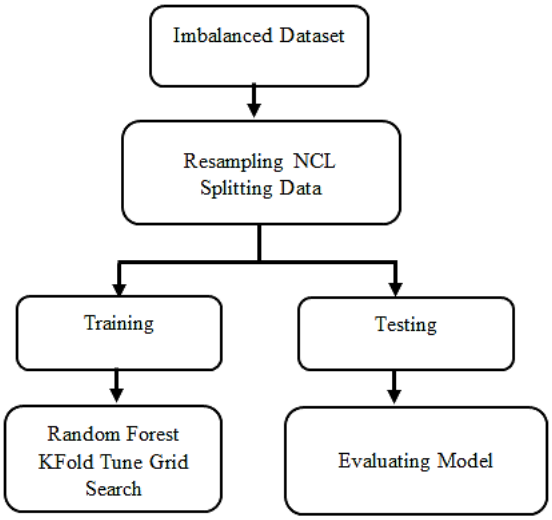


Figure 2. Step For Research

The study procedure is shown in Figure 2, starting with Resampling NCL and dividing the data into 80% training data and 20% test data. The training data are subjected to the Random Forest K-Fold Tune Grid Search algorithm, and the model is evaluated using the test data. Performance indicators, such as the Confusion Matrix, F1 Score, and G-mean - measures of algorithmic goodness - are used to obtain and assess classification results.

3. RESULTS AND DISCUSSION

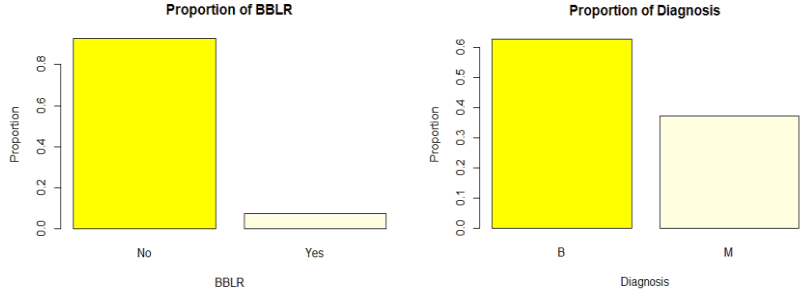
3.1 Dataset

Breast cancer and unbalanced BBLR data that can be downloaded from the UCI Machine Learning Repository were used in this work. Table 2 provides details on the data.

Table 2. Details of Dataset

Dataset	Number of Observation	Number of Variable	Class
BBLR	407	5	No, Yes
Breast Cancer	569	31	Benign, Malignant

Age, Gravida, Parity, Abortus, and BBLR with classes Non-BBLR and BBLR are the variables in the BBLR dataset for the Breast Cancer dataset. variable ID number, Diagnosis (M = malignant, B = benign, Ten real-valued features are computed for each cell nucleus: Radius (Mean of Distances From Center To Points On The Perimeter), Texture (Standard Deviation Of Gray-Scale Values), Perimeter, Area, Smoothness (Local Variation In Radius Lengths), Compactness (Perimeter^2 / Area - 1.0), Concavity (Severity of Concave Portions Of The Contour), Concave Points (Number Of Concave Portions Of The Contour), Symmetry, Fractal Dimension ("Coastline Approximation" - 1) . Class distribution: 357 benign, 212 malignant. The proportions of the dataset are depicted in Figure 3.



(a) (b)

Figure 3. (a) Proportion of BBLR, (b) Proportion of Diagnosis

TIN: Terapan Informatika Nusantara
Vol 3, No 8, January 2023, page 286-293
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v3i8.4124

3.2 Eksploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) constitutes an integral component inside the data mining process (Arifiyanti & Wahyuni, 2020) Exploratory Data Analysis (EDA) assumes a crucial role prior to engaging in the modeling process, as it serves as a preliminary step in comprehending the data and its underlying structure. Data exploration refers to the initial examination of data in order to identify patterns, investigate anomalies, evaluate hypotheses, and verify assumptions. This process involves the use of summary statistics and graphical representations. The present study incorporates several steps of Exploratory Data Analysis (EDA). These steps encompass the following: verification of data dimensions, examination of data structure, creation of class proportion tables, visualization of class distribution through bar charts, identification of outliers, generation of data correlation plots, conversion of character variables into numeric and categorical formats, and assessment of data statistics.

3.3 Preprocessing data

The procedure of preprocessing for imbalanced data classification using random forest with k-fold cross-validation and tune grid search encompasses multiple processes aimed at ensuring the efficacy of the model. The following is a comprehensive framework of the procedure:
The process of loading and exploring data.

The dataset BBLR and Breast Cancer, which includes unbalanced data, should be loaded. Examine the structural composition, characteristics, and distribution of the dataset's attributes and target variable.
Data Resampling:

In order to address the issue of class imbalance, resampling techniques are employed. Specifically, a random forest model is utilized in conjunction with k-fold cross-validation. Additionally, a grid search is performed to modify the model by undersampling the majority class.
Data Splitting: The dataset should be partitioned into two distinct components: the characteristics (X) and the target variable (Y). The data should be divided into training and testing sets with a ratio of 80:20.
Feature Engineering: Conduct any requisite feature engineering tasks, such as scaling or normalization, to enhance the quality and suitability of the features. The management of missing values can be addressed by either imputation or removal techniques.

3.4 Splitting Data on Training and Testing Data

The division of data into training and testing sets is an essential and pivotal stage in the development of a classification model. This stage is crucial in order to ensure that the model that has been constructed has the ability to generalize effectively to data that it has not been previously exposed to. In order to facilitate accurate predictions, it is important to partition the data into distinct training and testing sets. The present study partitions the complete dataset into an 80% portion for training purposes and a 20% portion for testing purposes. The partitioning of data into distinct subsets is crucial for evaluating the efficacy of a model on previously unobserved data and for verifying its capacity to generalize.

3.5 The outcome obtained from employing the Random Forest algorithm with K-Fold Cross-Validation and Tune Grid Search.

The "Confusion Matrix and Statistics" table is derived from the evaluation of a classification model's performance on a given dataset, either through testing or prediction. The table quantifies the performance of the model in classification tasks by evaluating the concordance between the projected outcomes and the ground truth values.

Tabel 3. Confusion Matrix Random Forest K-Fold 10 Tune Grid Search for BBLR Dataset

	Actual	
Prediction	No	Yes
No	76	4
Yes	0	0

Table 3 displays four potential prediction results alongside their respective actual values. Current, The term "No" refers to the specific value that has been categorized as negative. The term "Yes" refers to the specific value that is categorized as positive. Prediction: The model's prediction is rated as "No" (negative). The aforementioned statement indicates that the model's prediction has been classed as "Yes" (positive). Within the presented tabular representation, The algorithm accurately predicted "No" for a total of 76 cases that indeed had an actual value of "No". The model inaccurately classified four instances as "Yes" when their true value was "No." The model does not forecast any instances with an actual value of "Yes" as "No". The model did not accurately predict any instances with a true value of "Yes" as "Yes".

TIN: Terapan Informatika Nusantara
Vol 3, No 8, January 2023, page 286-293
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v3i8.4124

Tabel 4. Confusion Matrix Random Forest K-Fold 10 Tune Grid Search for Breast Cancer Dataset

	Actual	
Prediction	B	M
B	77	3
M	3	31

Table 4 presents four distinct forecast results together with their associated actual values. Reference : The reference value classed as "B" (representing the negative class, namely Benign) is denoted as the actual value. The result classed as "M" represents the real malignant kind, which is considered positive. Prediction: The model's prediction is categorised as "B" (representing the negative class, namely Benign). The classification "M" represents the model's prediction of malignancy, indicating the positive kind. Based on the data presented in the table, numerous inferences can be inferred. The model accurately predicted the value "B" in 77 cases. The model made inaccurate predictions of "M" for three situations that really had a value of "B". The model accurately predicted the actual value of "M" in 31 occasions. The model inaccurately predicted the value "B" for three cases that really had a value of "M". From this point, it is possible to compute a range of evaluation metrics in order to evaluate the performance of the model. These metrics include accuracy, sensitivity (recall) for the "M" class, specificity for the "B" class, positive predictive value, and negative predictive value. The table presented herein offers essential data pertaining to the model's ability to accurately differentiate between class "B" and "M" in terms of its predictions. The outcomes of this assessment will provide insights into the degree to which the model can effectively classify tumor types as either "B" or "M."

3.6 Evaluation of Classification Results

The Confusion Matrix table is an essential tool for evaluating the performance of a model, particularly in terms of its classification accuracy across several classes. The table provides the necessary data to compute several assessment measures, including accuracy, sensitivity (recall), specificity, positive predictive value, and negative predictive value. These metrics are used to assess the classification model's performance and determine its quality. Table 5 presents an evaluation of the categorization outcomes on the BBLR and Breast Cancer datasets.

Table 5. An Evaluation of The Categorization Outcomes on The BBLR and Breast Cancer Datasets.

Data	Evaluation Metric			
	Accuracy	F1-Score	Sensitivity	Gmean
BBLR	0,95	0,97	1	0
Breast Cancer	0,94	0,96	0,96	0,93

Based on the Table 5, The classification results of the BBLR data using random forest Kfold Tune Grid Search yielded an accuracy of 0.95. This value signifies that the model accurately classified 95% of the data instances. The accuracy metric evaluates the model's ability to make accurate predictions for both positive and negative classifications. The F1-Score, with a value of 0.97, is a metric that combines precision and recall. A value of 0.97 indicates that the model achieves a favorable equilibrium between precision and recall, facilitating precise predictions for both positive and negative categories.

The sensitivity value is equal to 1. A value of 1 indicates that the model effectively detects all instances of the intended class, which are positive cases. The absence of incorrect negative predictions suggests that the model exhibits a high degree of sensitivity towards positive cases. A Gmean score of 0 indicates a poor geometric mean between sensitivity and specificity. This observation suggests a disparity in the model's performance with regards to sensitivity and specificity. The model may demonstrate proficiency in accurately recognizing one class, while encountering difficulties in accurately identifying the other class.

The classification results of the BBLR data using random forest Kfold Tune Grid Search yielded an accuracy of 0.94. This value signifies that the model accurately identified 94% of the total data. The accuracy metric evaluates the model's ability to make accurate predictions for both positive and negative classifications. The F1-Score, with a value of 0.96, is a performance measure that combines precision and recall. The model's score of 0.96 indicates a commendable equilibrium between precision and recall, enabling it to make precise predictions for both positive and negative classifications.

The sensitivity of the model is 0.96, indicating its ability to correctly detect 96% of all positive occurrences, which corresponds to the intended class. The model exhibits a significant degree of sensitivity towards positive data values. The Geometric Mean (Gmean) metric, with a value of 0.93, quantifies the equilibrium between sensitivity and specificity. A result of 0.93 signifies a favorable equilibrium in the model's capacity to accurately discern positive and negative instances. The classification results of the Breast Cancer data were obtained using the random forest algorithm with Kfold Tune Grid Search. The accuracy of the model is 0.94. This finding suggests that the model accurately identified 94% of the dataset.

The accuracy metric evaluates the model's ability to correctly predict outcomes for both positive and negative classifications. Accuracy is a metric that quantifies the ratio of accurate forecasts to the overall number of predictions

TIN: Terapan Informatika Nusantara
Vol 3, No 8, January 2023, page 286-293
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v3i8.4124

made. Nevertheless, in the present investigation, where the positive class (cancer) may constitute a minority within the dataset, relying solely on accuracy is inadequate. The presence of an imbalanced negative class can introduce bias into the accuracy metric, perhaps resulting in an inflated perception of the model's performance.

The F1-Score, with a value of 0.96, is a performance measure that combines precision and sensitivity. A score of 0.96 signifies that the model achieves a harmonious trade-off between precision and recall, enabling it to deliver precise and reliable predictions for both positive and negative classifications. The F1 score can be defined as the harmonic mean of precision and sensitivity. The utilisation of this measure is appropriate in situations where there exists an asymmetry between the presence of positive and negative classes, such as in the context of breast cancer. The F1 score is a metric that successfully balances the sensitivity of recognising positive situations and the precision of preventing false positive errors. In situations of breast cancer, it is of utmost importance to prioritise high sensitivity in cancer detection and minimise the occurrence of misdiagnosis.

The Geometric Mean (G-Mean) is a mathematical measure used to calculate the central tendency of a set of numbers. It is computed by taking the nth root of The G-Mean measure is considered appropriate for evaluating the equilibrium between sensitivity and specificity. This paper presents a comprehensive analysis of the model's classification performance, taking into account the distribution of classes within the dataset. When considering BBLR and breast cancer, the utilisation of G-Mean can provide a more comprehensive assessment of the model's efficacy in accurately identifying both positive and negative cases in a balanced manner.

The aforementioned statistics suggest that the model exhibits exceptional performance in classification, characterized by a notable level of accuracy, a well-balanced compromise between precision and recall, and robust sensitivity and specificity.

4. CONCLUSION

One of the health studies in which it is preferred that False Positives occur rather than False Negatives is this one. The percentage of accurate forecasts among all predictions is known as accuracy. Accuracy alone, however, is insufficient in this investigation because the positive class (M) is underrepresented in the sample. Sensitivity is therefore regarded as a crucial evaluation criterion to guarantee that the model can identify breast cancer cases, identify low birth weight infants (BBLR), and avoid missing genuine positive instances. High sensitivity in identifying cancer and BBLR and avoiding misdiagnosis are critical. The F1 score strikes a compromise between effectively detecting positive cases (Sensitivity) and avoiding false positive mistakes (Precision). The G-Mean can provide a more thorough understanding of the model's performance in balancing positive and negative events in the context of BBLR and breast cancer. This study has effectively used the resampling NCL Random Forest KFold Tune Grid Search because it produces high levels of accuracy, sensitivity, F1-Score, and G-Mean.

REFERENCES

Arifiyanti, A. A., & Wahyuni, E. D. (2020). SMOTE: Metode penyeimbang kelas pada klasifikasi data mining. *Scan: Jurnal Teknologi Informasi Dan Komunikasi*, *15*(1), 34–39.

Astuti, F. D., & Lenti, F. N. (2021.). *Implementasi SMOTE untuk mengatasi Imbalance Class pada Klasifikasi Car Evolution menggunakan K-NN*.

Bappenas, S. (2020). Metadata Indikator Tujuan Pembangunan Berkelanjutan (TPB). *Sustainable Development Goals (SDGs) Indonesia Pilar Pembangunan Ekonomi*.

Choirunnisa, S. (2019). *Metode hibrida oversampling dan ketidakseimbangan data kegagalan*.

Devella, S., Yohannes, Y., & Rahmawati, F. N. (2020). Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *7*(2), 310–320.

Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., & Zoromi, F. (2022). Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang. *MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *21*(3), 677–690.

Ihfa, R., & Harsanti, T. (2020). Komparasi Teknik Resampling Pada Pemodelan Regresi Logistik Biner. *Seminar Nasional Official Statistics*, *2020*(1), 863–870.

Kemenkes, R. I. (2019). Profil Kesehatan Indonesia Tahun 2021. Kementerian Kesehatan Republik Indonesia. *Jakarta: Kementerian Kesehatan Republik Indonesia*.

Lestari, A., Mariati, E., & Widiatry, W. (2020). Model Klasifikasi Kepuasan Mahasiswa Teknik Terhadap Sarana Pembelajaran Menggunakan Data Mining. *Jurnal Teknologi Informasi: Jurnal Keilmuan Dan Aplikasi Bidang Teknik Informatika*, *14*(2), 112–118.

Lujan-Moreno, G. A., Howard, P. R., Rojas, O. G., & Montgomery, D. C. (2018). Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. *Expert Systems with Applications*, *109*, 195–205.

Nugraha, W., & Sasongko, A. (2022). Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search. *SISTEMASI : Jurnal Sistem Informasi*, *11*(2), 391–401.

Pangestika, M. P., Sumertajaya, I. M., & Rizki, A. (2021). Penerapan Synthetic Minority Oversampling Technique pada Pemodelan Regresi Logistik Biner terhadap Keberhasilan Studi Mahasiswa Program Magister IPB. *Xplore:*


TIN: Terapan Informatika Nusantara
Vol 3, No 8, January 2023, page 286-293
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v3i8.4124

Journal of Statistics, *10*(2), 152–166.

Qadrini, L., Hikmah, H., & Megasari, M. (2022). Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 386–391. https://doi.org/10.47065/josyc.v3i4.2154

Qadrini L, Sepperwali A, & Aina A. (2021). Decision Treedan Adaboostpada Klasifikasi Penerima Program Bantuan Sosial. *Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial*, *2*(7), 1959–1966.

Siringoringo, R. (2018). Klasifikasi data tidak seimbang menggunakan algoritma SMOTE dan k-nearest neighbor. *Journal Information System Development (ISD)*, *3*(1).

Suryani Agustin, Budi Darma Setiawan, & Mochammad Ali Fauzi. (2019). Klasifikasi Berat Badan Lahir Rendah (BBgustin, Suryani Setiawan, Budi Darma Fauzi, Mochammad AlLR) Pada Bayi Dengan Metode Learning Vector Quantization (LVQ). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(3), 2929–2936. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/4831/2254/

Turlapati, V. P. K., & Prusty, M. R. (2020). Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19. *Intelligence-Based Medicine*, *3*, 100023.

Wasono, R. (2022). *Perbandingan Metode Random Forest dan naive bayes untuk Klasifikasi Debitur Berdasarkan Kualitas Kredit*.