Scientific Journal of Informatics
Vol. 11, No. 2, May 2024
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

The Impact of Balanced Data Techniques on Classification Model Performance

Jasman Pardede[1*], Dika Prasetia Pamungkas[2]
[1,2]Department of Informatics, Institut Teknologi Nasional (Itenas) Bandung

Abstract.
Purpose: The aim of this study is to examine the impact of balanced data techniques on the performance of classification models.
Methods: To balance the imbalanced dataset, several resampling techniques are employed: The Synthetic Minority Oversampling Technique (SMOTE), Borderline-SMOTE (B-SMOTE), and SMOTE and Edited Nearest Neighbors (SMOTE-ENN). Classification is then performed using both balanced and unbalanced datasets to evaluate the impact of resampling techniques on classification model performance.
Result: This study proposes the SMOTE, B-SMOTE, and SMOTE-ENN techniques for generating synthetic data. Experimental results showed that re-sampling can improve model performance on KNN, Naive Bayes, and Decision Tree. The best-balanced data technique is the SMOTE-ENN. The second best is B-SMOTE, and the last is SMOTE. If compared to the unbalanced dataset, the SMOTE technique encourages increasing the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 4.79%, 35.89%, 35.32%, 35.63%, 46.94%, and 34.89%, respectively on DT method. The B-SMOTE technique on the DT method improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 5.62%, 36.45%, 35.88%, 36.19%, 47.40%, and 35.46% if compared to the unbalanced dataset. The SMOTE-ENN technique improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 8.11%, 34.53%, 43.25%, 41.63%, 62.85%, and 42.91% if compared to the unbalanced dataset.
Novelty: Based on the experiment results, the best-balanced data technique is the SMOTE-ENN. The SMOTE-ENN technique improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC.

Keywords: Imbalanced data, SMOTE, B-SMOTE, SMOTE-ENN, Performance
Received May 2024 / **Revised** May 2024 / **Accepted** May 2024

INTRODUCTION

Data imbalanced phenomena commonly occur in a variety of real-world situations, including in the world of telecommunications, the web, finance, ecology, biology, medicine, and many other fields [1]. Imbalanced data can result in a more flexible machine learning model for predicting majority classes and less effective for identifying rare instances [2], [3]. Datasets are considered imbalanced when one class has much greater dominance than another. The process of forming a classification model in a minority class is generally less well-trained, so the predictive model built tends to be inaccurate [4].

Techniques used to address imbalanced data include data level, algorithm level, feature level, cost-sensitive level, and deep learning [5]–[7]. In addition, on data-level techniques, there is a method of data resampling where the amount of data in each class is modified to a balance between all classes, this is a proven effective approach [8]. Resampling techniques can be grouped into under-sampling, over-sampling, and combinations [2].

The under-sampling technique removes some data from the dataset that has the majority class. Several under-sampling techniques include: Random Under-sampling, Neighborhood Cleaning Rule, Tomek Links, Edited Nearest Neighbor (ENN), and others [8]. In the over-sampling method, new samples are made based on minority class samples for a more balanced class distribution. Over-sampling techniques include: SMOTE, B-SMOTE, SMOTENS, ADASYN, and others [8]. Combination technique are a combination of under-sampling and over-sampling. Combination techniques include SMOTE-ENN, SMOTE-Tomek, and others [9], [10].

*Corresponding author.
Email addresses: jasman@itenas.ac.id (Pardede)
DOI: 10.15294/sji.v11i2.3649

Under-sampling and over-sampling have an inevitable disadvantage [8]. In under-sampling can cause loss of information, which becomes a problem as it reduces the diversity and representation of majority classes in datasets [1], [11]. While Over-sampling can cause overfitting because it only doubles or synthesizes part of the minority sample [3], [11]. To overcome these constraints, a technique has been proposed that combines under-sampling and over-sampling.

The study aims to reveal the impact of the data-balanced techniques on the performance of the classification model. The study uses several balanced data techniques, namely SMOTE [12], B-SMOTE [13], and SMOTE-ENN [7]. The classification methods proposed in this study, namely Decision Tree, Naïve Bayes, and KNN. The performance of the classification model proposed is measured based on Accuracy, Precision, Recall, F1-score, ROC curve, G-Mean, and Curve-ROC for both imbalanced and balanced data by SMOTE, B-SMOTE, and SMOTE-ENN.

METHODS

Business process model system

To showing the impact of balanced data techniques on classification performance, three balanced data techniques with three types of classification methods are proposed. The balanced data techniques used are SMOTE, B-SMOTE, and SMOTE-ENN. The classification methods implemented are KNN, Naive Bayes, and Decision Tree. The dataset used is an unbalanced dataset, namely Personal Key Indicators of Heart Disease.

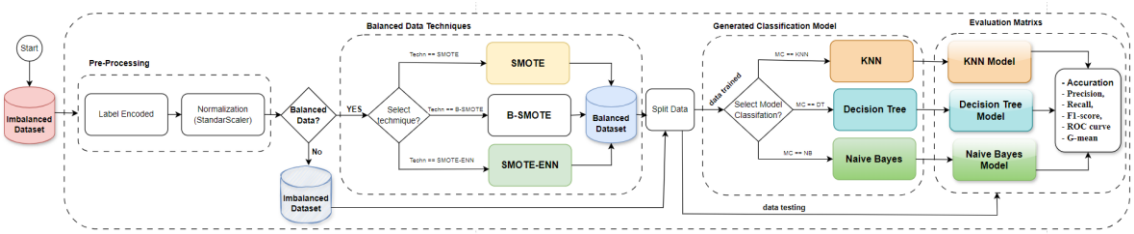

Figure 1. Business process model proposed

Before balancing data on the dataset, a pre-processing process is carried out, namely the label encoded and normalization. Some of the experimental scenarios that were carried out, namely:

1. Generated the classification model on an unbalanced dataset, then measure the performance of the classification model formed.
2. Generated the balanced dataset using the SMOTE, B-SMOTE, or SMOTE-ENN technique. Each dataset generated is used as a dataset to build the classification model. For each dataset that has been balanced, the classification model is generated and the performance of each model is measured.

Before the process of generated the classification model, a data split process is carried out, to separate the training data and the testing data. Meanwhile, model performance is measured using Accuracy, Precision, Recall, and F1-Score, G-Mean, and Curve-ROC values. The Curve-ROC is a graphical representation that visualizes the trade-off between the Recall and the False Positive Rate (FPR). The business proses model used for this study is as shown in Figure 1. The formula for calculating the performance value is as follows [14], [15]:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - score = 2 * \left(\frac{Precision\ x\ Recall}{Precision+Recall}\right) \quad (4)$$

$$FPR = \frac{FP}{TN+FP} \quad (5)$$

$$G - Mean = \sqrt{Precision\ x\ Recall} \quad (6)$$

where:
TP = True Positive
FP = False Positive
FN = False Negative
TN = True Negative

Dataset
The dataset used is a dataset entitled "Personal Key Indicators of Heart Disease" obtained from the Kaggle.com website and processed by Kamil Pytlak. This dataset comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to collect data on the health status of the United States population. The "Personal Key Indicators of Heart Disease" dataset has 18 features and consists of two classes, namely YES with a total of about 27,374 data points, and NO with a volume of about 292,422 data points.

SMOTE
SMOTE is a method used to deal with class imbalances by generating synthetic data samples for a minority class. The SMOTE method works by searching for *k* nearest neighbors (i.e., the nearest incompatibility of data as *k*) for each data point in the minority class, then generating synthesized data as a percentage of the desired minor data duplication (percentage oversampling, N%) and *k*-nearest neighbor randomly selected [12].

B-SMOTE
B-SMOTE is an improved oversampling algorithm based on SMOTE, which uses only a few class samples in the borderline to combine new samples, thereby improving the distribution of sample categories. The borderline sample of SMOTE is divided into three categories: Safe, Dangerous, and Noise [13]. Finally, only the Danger sample will be oversampled. The steps of this algorithm are as follows:
1) If m0 = m, then the samples around x_i all come from a different class, which is noted as noise data. This type of data will have a negative impact on the output of the generation; therefore, it is considered not to be used in the process of making a new sample.
2) If $m/2 \leq m0 < m$, more than half of m samples around x_i come from different classes. Samples that are in the boundary like this are defined as Danger samples.
3) If $0 \leq m0 < m/2$, more than half of the m samples around x_i come from the same class and are quoted as safe samples.

SMOTE-ENN
SMOTE-ENN is a combination of two data processing techniques used in the processing of class imbalances on classification problems, namely SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors) [7]. The commonly used under-sampling algorithm, ENN, removes the sample by checking whether the classes from the majority sample are equal to the classes of the nearest neighbor *k*. In this context, it is assumed that the majority sample is named x_{maj_i}. Then, look for k (usually k values 3) for the closest neighboring sample of x_{maj_i} and evaluate the class x_{maj_i} along with x_{maj_i} for k for the neighboring samples [16]. If a class of x_{maj_i} is not equal to a class from one of the nearest k, then x_{maj_i} will be deleted.

Naïve Bayes
Naïve Bayes (NB) is one of the statistical methods in classification that can capture uncertainties related to models with principles focused on the interpretation of probability. This approach is used to address challenges in areas such as diagnosis and prediction [17]. Another definition says that Naïve Bayes is a method in statistics and machine learning used for classification and prediction. This method is based on the Bayes theorem, which connects the probability of an event with the probabilities of other related events. Naïve Bayes is categorized as a probability-based learning algorithm used mainly in classification.

Decision Tree
A decision tree (DT) is one of the methods of classification that adopts the structure of a tree, where each node represents an attribute, its branch represents the values of attributes, and its leaf represents a class. A decision tree is a well-known and frequently used method of classification, not only because of its efficient construction but also because the resulting model is easy to interpret. In the decision tree, there are three

types of nodes: the root node, the internal node, and the sheet node. To select the attribute as the root, it is based on the highest gain ratio value of the existing attributes [17], [18].

K-Nearest Neighbors

K-Nearest Neighbors (KNN) is an algorithm used in the field of machine learning for classification and regression. The basic principle of KNN is that an object will be classified or predicted based on the majority of its nearest neighbors in the feature space. The K-Nearest Neighbor algorithm is a supervised learning algorithm that requires training data and a specified *k* value to find the nearest *k* data based on distance calculations. If the *k* data has a different class, this algorithm predicts the class of the unknown data to be the same as the majority class [19].

RESULTS AND DISCUSSIONS

To test the methods of re-sampling proposed, there are two approaches in this study, i.e., (1) reducing the number of majority classes before generating synthetic data and (2) generating minority classes to achieve ***n***-percent. The parameters used to generate synthetic data are the same for both proposed approaches. In the SMOTE method use the parameter *k_neighbors* (*k*) = 5. For the B-SMOTE method use the parameters *k* = 5 and *m_neighbors* (*m*) = 10. In the SMOTE-ENN method use the parameters *sampling_strategy* = '*auto'*. Reduce majority classes by a ratio from 10% to 50%. The ratio of 10% states that the majority class dataset is randomly deleted until a ratio of majority to minority data is obtained of only 10%. For 10%, since class **YES** has 27,373 data and class **NO** has 292,422 data then class **NO** was deleted randomly until only 273,730 data. To achieve balanced data, 246,357 were generated synthetic data for the **YES** class. So, the total dataset is 301,103 as shown in Table 1.

Table 1. Majority data reduction

Ratio	NO	YES	Generate	Volume of data
10%	273,730	27,373	246,357	301,103
20%	136,865	27,373	109,492	164,238
30%	91,243	27,373	63,870	118,616
40%	68,432	27,373	41,059	95,805
50%	54,746	27,373	27,373	82,119

In generating minority classes, synthetic data is generated to increase the amount of data in the minority category. Ratio oversampling is 100%, 200%, 300%, 400%, 500%, 600%, 700%, 800%, 900%, and *auto*, based on the amount **YES** class. On *auto*, synthetic data is created as a percentage of the ratio of the minority data to the majority. For 100% oversampling, it states that the number of **YES** class data generated is the same as the amount of original data. The original data for the **YES** class is 27,373 so the synthetic data generated is 27,373. Therefore, the total number of **YES** class data is 54,746. The number of datasets after the balanced data process is 347,168, as shown in Table 2.

Table 2. Oversampling rate on minority data

Oversampling Rate	NO	YES	Generate	Volume of data
100%	292,422	54,746	27,373	347,168
200%	292,422	82,119	54,746	374,541
300%	292,422	109,492	82,119	401,914
400%	292,422	136,865	109,492	429,287
500%	292,422	164,238	136,865	456,660
600%	292,422	191,611	164,238	484,033
700%	292,422	218,984	191,611	511,406
800%	292,422	246,357	218,984	538,779
900%	292,422	273,730	246,357	566,152
auto	292,422	292,422	265,049	584,844

Majority data reduction performance results

The amount of synthetic data generated using the SMOTE technique in majority data reduction is used as a new dataset. The new dataset is split into training data and test data. Training data is used to build a classification model using the KNN, NB, or DT method. Each classification model built is tested for model performance using predetermined test data. Model performance evaluation is considered using Accuracy, Precision, Recall, F1-score, ROC curve, G-Mean, and Curve-ROC. The performance of the classification model in majority data reduction using the balanced data SMOTE technique is as stated in Table 3.

The results of balanced data using the B-SMOTE technique are used as a new dataset. The new dataset resulting from balanced data using the B-SMOTE technique is also used to build the proposed classification model. Then an evaluation of the model performance is carried out. The performance of the classification model in majority data reduction using the B-SMOTE balanced data technique is as stated in Table 4. Whereas, the performance of the classification model in majority data reduction using the SMOTE-ENN technique is as stated in Table 5.

Table 3. Model performance results with majority data reduction on SMOTE

Ratio	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	KNN	0.8670	0.8785	0.8669	0.8660	0.8626	0.8669
10%	NB	0.7114	0.7205	0.7115	0.7085	0.7044	0.7115
	DT	0.9020	0.9020	0.9020	0.9020	0.9019	0.9021
	KNN	0.8264	0.8388	0.8263	0.8248	0.8208	0.8263
20%	NB	0.7063	0.7157	0.7064	0.7031	0.6986	0.7064
	DT	0.8518	0.8519	0.8518	0.8518	0.8518	0.8518
	KNN	0.8052	0.8146	0.8043	0.8034	0.7995	0.8043
30%	NB	0.7074	0.7171	0.7084	0.7047	0.7010	0.7084
	DT	0.8156	0.8156	0.8156	0.8156	0.8156	0.8174
	KNN	0.7868	0.7936	0.7866	0.7855	0.7829	0.7866
40%	NB	0.7112	0.7211	0.7115	0.7081	0.7037	0.7115
	DT	0.7828	0.7828	0.7828	0.7828	0.7828	0.7820
	KNN	0.7743	0.7796	0.7752	0.7736	0.7723	0.7752
50%	NB	0.7083	0.7172	0.7069	0.7043	0.6989	0.7069
	DT	0.7598	0.7597	0.7598	0.7597	0.7598	0.7604

Based on the results of the balanced dataset experiment using the SMOTE technique, it is evident that the larger the majority class data that is reduced before balanced data is carried out, the classification performance decreases for each proposed classification method, as shown in Table 3. The best classification performance is achieved at a ratio of 10%.

Table 4. Model performance results with majority data reduction on B-SMOTE

Ratio	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	KNN	0.8865	0.8942	0.8864	0.8859	0.8836	0.8864
10%	NB	0.7387	0.7438	0.7388	0.7374	0.7353	0.7388
	DT	0.9047	0.9046	0.9046	0.9046	0.9046	0.9046
	KNN	0.8402	0.8544	0.8402	0.8386	0.8342	0.8402
20%	NB	0.7058	0.7117	0.7058	0.7037	0.7008	0.7058
	DT	0.8499	0.8500	0.8499	0.8499	0.8503	0.8499
	KNN	0.8158	0.8312	0.8147	0.8133	0.8074	0.8147
30%	NB	0.6834	0.6905	0.6843	0.6810	0.6780	0.6843
	DT	0.8127	0.8127	0.8127	0.8127	0.8127	0.8127
	KNN	0.7917	0.8074	0.7914	0.7888	0.7832	0.7914
40%	NB	0.6732	0.6819	0.6735	0.6695	0.6648	0.6735
	DT	0.7792	0.7792	0.7792	0.7792	0.7792	0.7800
	KNN	0.7715	0.7861	0.7730	0.7692	0.7650	0.7730
50%	NB	0.6671	0.6757	0.6655	0.6616	0.6548	0.6655
	DT	0.7569	0.7569	0.7569	0.7569	0.7569	0.7569

Table 4 and Table 5 also show the same performance results, namely the best performance at a ratio of 10%. Balanced data techniques with SMOTE and B-SMOTE provide the best performance for the DT method. But for the SMOTE-ENN technique, the best performance is the KNN method. Based on the experimental results, it is stated that the greater the ratio of the majority class that is reduced before balanced data is carried out, the smaller the performance will be.

Table 5. Model performance results with majority data reduction on SMOTE-ENN

Ratio	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	KNN	0.9543	0.9570	0.9497	0.9528	0.9492	0.9497
10%	NB	0.7477	0.7629	0.7645	0.7476	0.7567	0.7645
	DT	**0.9382**	**0.9370**	**0.9363**	**0.9366**	**0.9362**	**0.9363**
	KNN	0.9361	0.9431	0.9351	0.9387	0.9345	0.9351
20%	NB	0.7503	0.7644	0.7750	0.7494	0.7669	0.7750
	DT	0.9165	0.9119	0.9128	0.9123	0.9126	0.9128
	KNN	0.9370	0.9356	0.9294	0.9323	0.9290	0.9294
30%	NB	0.7621	0.7706	0.7876	0.7599	0.7810	0.7876
	DT	0.9037	0.8970	0.8978	0.8974	0.8975	0.8978

	KNN	0.9325	0.9282	0.9259	0.9271	0.9256	0.9259
40%	NB	0.7710	0.7792	0.8000	0.7684	0.7927	0.8000
	DT	0.8940	0.8849	0.8875	0.8862	0.8872	0.8875
	KNN	0.9307	0.9233	0.9227	0.9230	0.9224	0.9227
50%	NB	0.7623	0.7680	0.7973	0.7575	0.7895	0.7973
	DT	0.8913	0.8777	0.8826	0.8801	0.8822	0.8826

Experimental results express that the B-SMOTE technique is better than SMOTE. The SMOTE-ENN technique is better than the B-SMOTE technique. The best performance for SMOTE is using the DT method with Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC values of 0.902, 0.902, 0.902, 0.902, 0.9019, and 0.9021 respectively. The best performance for B-SMOTE is using the DT method with Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC values of 0.9047, 0.9046, 0.9046, 0.9046, 0.9046, and 0.9046, respectively. The best performance for SMOTE-ENN is using the KNN method with Accuracy, Precision, Recall, F1-Score, G-mean, Curve-ROC values of 0.9543, 0.957, 0.9497, 0.9528, 0.9492, and 0.9497, respectively. Meanwhile, the best performance for SMOTE-ENN using the DT method with Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC values are 0.9382, 0.937, 0.9363, 0.9366, 0.9362, and 0.9363, respectively.

Minority data oversampling performance results

Synthetic data generated using the oversampling SMOTE technique on minority data with the amounts presented in Table 2 is used as a new dataset. The new dataset be used for building KNN, NB, and DT classification models. The performance of the classification model based on the balanced data dataset using the SMOTE technique is presented in Table 6.

Table 6. Model performance results with oversampling on SMOTE

Oversampling Rate	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	KNN	0.8619	0.7408	0.7254	0.7326	0.6973	0.7254
100%	NB	0.8096	0.6579	0.6802	0.6673	0.6533	0.6802
	DT	0.8631	0.7435	0.7465	0.7450	0.7268	0.7492
	KNN	0.8483	0.7783	0.7946	0.7858	0.7888	0.7946
200%	NB	0.7839	0.6858	0.6863	0.6861	0.6638	0.6863
	DT	0.8696	0.8098	0.8124	0.8111	0.8060	0.8130
	KNN	0.8463	0.8036	0.8337	0.8156	0.8332	0.8337
300%	NB	0.7660	0.7030	0.6947	0.6985	0.6768	0.6947
	DT	0.8744	0.8406	0.8441	0.8423	0.8414	0.8440
	KNN	0.8485	0.8238	0.8519	0.8336	0.8519	0.8519
400%	NB	0.7493	0.7105	0.6947	0.7009	0.6781	0.6947
	DT	0.8826	0.8647	0.8647	0.8651	0.8641	0.8661
	KNN	0.8510	0.8382	0.8606	0.8443	0.8599	0.8606
500%	NB	0.7393	0.7176	0.7007	0.7065	0.6870	0.7007
	DT	0.8852	0.8750	0.8762	0.8756	0.8756	0.8766
	KNN	0.8535	0.8490	0.8638	0.8509	0.8624	0.8638
600%	NB	0.7284	0.7183	0.7015	0.7061	0.6894	0.7015
	DT	0.8899	0.8845	0.8856	0.8850	0.8854	0.8854
	KNN	0.8583	0.8600	0.8671	0.8578	0.8648	0.8671
700%	NB	0.7240	0.7225	0.7075	0.7104	0.6977	0.7075
	DT	0.8935	0.8910	0.8919	0.8915	0.8918	0.8917
	KNN	0.8619	0.8679	0.8680	0.8619	0.8650	0.8680
800%	NB	0.7190	0.7230	0.7098	0.7107	0.7012	0.7098
	DT	0.8970	0.8963	0.8964	0.8963	0.8964	0.8964
900%	KNN	0.8667	0.8757	0.8693	0.8664	0.8657	0.8693
	NB	0.7156	0.7232	0.7121	0.7108	0.7042	0.7121
	DT	0.9016	0.9015	0.9016	0.9015	0.9016	0.9017
	KNN	0.8665	0.8772	0.8665	0.8656	0.8624	0.8665
auto	NB	0.7083	0.7176	0.7083	0.7052	0.7008	0.7083
	DT	0.9050	0.9050	0.9050	0.9050	0.9050	0.9047

The new dataset resulting from balanced data using the oversampling B-SMOTE technique is also used to build the proposed classification model. Then an evaluation of the model performance is carried out. The performance of the classification model in minority data using the oversampling B-SMOTE balanced data technique is as stated in Table 7. Whereas, the performance of the classification model in minority data using the oversampling SMOTE-ENN technique is as prepared in Table 8.

Based on experimental results, it shows that the three over-sampling techniques on minority data can improve the performance of the classification model. The more balanced the minority data is with the majority data, the performance of the resulting classification model is also greater, which can be seen in oversampling ranging from 500% to 900%. The best performance is in generating synthetic data using *auto* parameters.

Table 7. Model performance results with oversampling on B-SMOTE

Oversampling Rate	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	KNN	0.8705	0.7572	0.7579	0.7576	0.7398	0.7579
100%	NB	0.8138	0.6666	0.6937	0.6779	0.6711	0.6937
	DT	0.8633	0.7440	0.7505	0.7472	0.7321	0.7505
	KNN	0.8658	0.8013	0.8308	0.8142	0.8284	0.8308
200%	NB	0.7912	0.6981	0.7043	0.7010	0.6869	0.7043
	DT	0.8707	0.8113	0.8149	0.8131	0.8088	0.8149
	KNN	0.8665	0.8267	0.8607	0.8403	0.8605	0.8607
300%	NB	0.7767	0.7177	0.7146	0.7161	0.7015	0.7146
	DT	0.8761	0.8425	0.8465	0.8445	0.8441	0.8465
	KNN	0.8682	0.8446	0.8736	0.8551	0.8735	0.8736
400%	NB	0.7674	0.7322	0.7238	0.7276	0.7137	0.7238
	DT	0.8825	0.8642	0.8659	0.8651	0.8647	0.8659
	KNN	0.8717	0.8586	0.8804	0.8654	0.8799	0.8804
500%	NB	0.7595	0.7395	0.7293	0.7335	0.7214	0.7293
	DT	0.8867	0.8765	0.8778	0.8772	0.8773	0.8778
	KNN	0.8753	0.8698	0.8846	0.8728	0.8835	0.8846
600%	NB	0.7535	0.7438	0.7335	0.7373	0.7272	0.7335
	DT	0.8918	0.8863	0.8879	0.8871	0.8877	0.8879
	KNN	0.8789	0.8788	0.8864	0.8783	0.8848	0.8864
700%	NB	0.7486	0.7461	0.7363	0.7391	0.7310	0.7363
	DT	0.8963	0.8939	0.8948	0.8943	0.8948	0.8948
	KNN	0.8831	0.8864	0.8880	0.8831	0.8860	0.8860
800%	NB	0.7441	0.7459	0.7371	0.7386	0.7325	0.7371
	DT	0.8992	0.8984	0.8988	0.8986	0.8988	0.8988
	KNN	0.8859	0.8918	0.8880	0.8858	0.8857	0.8880
900%	NB	0.7419	0.7464	0.7393	0.7391	0.7351	0.7393
	DT	0.9056	0.9055	0.9056	0.9055	0.9056	0.9056
	KNN	0.8882	0.8957	0.8882	0.8877	0.8856	0.8882
auto	NB	0.7413	0.7463	0.7413	0.7400	0.7379	0.7413
	DT	0.9083	0.9083	0.9083	0.9083	0.9083	0.9083

Table 8. Model performance results with oversampling on SMOTE-ENN

Oversampling Rate	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	KNN	0.8619	0.7408	0.7254	0.7326	0.6973	0.7254
100%	NB	0.8096	0.6579	0.6802	0.6673	0.6533	0.6802
	DT	0.8631	0.7435	0.7465	0.7450	0.7268	0.7492
	KNN	0.8483	0.7783	0.7946	0.7858	0.7888	0.7946
200%	NB	0.7839	0.6858	0.6863	0.6861	0.6638	0.6863
	DT	0.8696	0.8098	0.8124	0.8111	0.8060	0.8130
	KNN	0.8463	0.8036	0.8337	0.8156	0,8332	0.8337
300%	NB	0.7660	0.7030	0.6947	0.6985	0.6768	0.6947
	DT	0.8744	0.8406	0.8441	0.8423	0.8414	0.8440
	KNN	0.8485	0.8238	0.8519	0.8336	0.8519	0.8519
400%	NB	0.7493	0.7105	0.6947	0.7009	0.6781	0.6947
	DT	0.8826	0.8647	0.8647	0.8651	0.8641	0.8661
	KNN	0.9407	0.9380	0.9419	0.9397	0.9418	0.9419
500%	NB	0.7697	0.7762	0.7498	0.7550	0.7380	0.7498
	DT	0.9188	0.9168	0.9171	0.9170	0.9170	0.9171
	KNN	0.9438	0.9431	0.9450	0.9437	0.9449	0.9450
600%	NB	0.7615	0.7729	0.7526	0.7538	0.7425	0.7526
	DT	0.9247	0.9242	0.9246	0.9244	0.9246	0.9246
	KNN	0.9497	0.9505	0.9496	0.9496	0.9494	0.9494
700%	NB	0.7593	0.7741	0.7594	0.7560	0.7505	0.7594
	DT	0.9322	0.9322	0.9322	0.9322	0.9322	0.9322
800%	KNN	0.9510	0.9526	0.9495	0.9507	0.9492	0.9495

Oversampling Rate	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	NB	0.7511	0.7672	0.7582	0.7500	0.7497	0.7582
	DT	0.9347	0.9346	0.9344	0.9345	0.9343	0.9344
	KNN	0.9553	0.9576	0.9522	0.9544	0.9518	0.9522
900%	NB	0.7462	0.7617	0.7591	0.7461	0.7513	0.7591
	DT	0.9403	0.9396	0.9394	0.9395	0.9393	0.9394
	KNN	0.9565	0.9589	0.9523	0.9551	0.9518	0.9523
auto	NB	0.7449	0.7601	0.7611	0.7448	0.7534	0.7611
	DT	**0.9412**	**0.9399**	**0.9397**	**0.9398**	**0.9397**	**0.9397**

Table 6 and Table 7 show that the B-SMOTE technique is better than SMOTE also. The SMOTE-ENN technique is better than the B-SMOTE technique. Balanced data techniques with SMOTE and B-SMOTE provide the best performance for the DT method, KNN, and the last NB. Meanwhile, the best performance for the SMOTE-ENN technique is the KNN method, followed by DT, and the end is NB. Based on the experimental results, it is stated that the closer the amount of synthetic data for the minority class is to the number of the majority class, the resulting dataset will produce increasingly improved performance.

The best performance for SMOTE and B-SMOTE technique is using the DT method with Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC values of 0.905, 0.905, 0.905, 0.905, 0.905, and 0.905, respectively. The best performance for SMOTE-ENN is using the KNN method with Accuracy, Precision, Recall, F1-Score, G-mean, Curve-ROC values of 0.9565, 0.9589, 0.9523, 0.9551, 0.9518, and 0.9523, respectively. Meanwhile, the best performance for SMOTE-ENN using the DT method with Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC values are 0.9412, 0.9399, 0.9397, 0.9398, 0.9397, and 0.9397, respectively. The model performance using the dataset generated from Minority Data Oversampling is better than using the dataset generated from Majority Data Reduction data. Finally, the best performance was achieved by the KNN classification model from Minority Data Oversampling data generation.

Combination of model performance results with parameter configuration

In this study, we considered the effect of parameter configuration on the model performance for the best model, i.e. ratios of 10% (R-10%), ratio of 20% (R-20%), Oversampling-800% (O-800%), Oversampling-900% (O-900%), and *auto*. The SMOTE technique using the parameter *k* with {3, 5, and 7}. For the B-SMOTE and SMOTE-ENN technique use the parameters *k* with {3, 5, and 7}, and *m* with {3, 5, and 7}. The performance of the classification model based on the balanced data dataset using the SMOTE technique with parameter configuration is presented in Table 9. The performance of the classification model based on B-SMOTE, and SMOTE-ENN techniques with parameter configuration is presented in Table 10 and Table 11, respectively.

Table 9. Best model performance results with SMOTE parameters configuration

Cases	Parameter	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	k=3	KNN	0.8723	0.8850	0.8722	0.8712	0.8675	0.8722
R-10%	k=7	NB	0.7129	0.7218	0.7130	0.7100	0.7059	0.7130
	k=3	DT	0.9022	0.9022	0.9022	0.9022	0.9022	0.9022
	k=3	KNN	0.8331	0.8473	0.8330	0.8313	0.8269	0.8330
R-20%	k=7	NB	0.7083	0.7173	0.7083	0.7052	0.7009	0.7083
	k=7	DT	0.8518	0.8518	0.8518	0.8518	0.8518	0.8518
	k=3	KNN	0.8685	0.8753	0.8748	0.8685	0.8714	0.8748
O-800%	k=5	NB	0.7190	0.7230	0.7098	0.7107	0.7012	0.7098
	k=3	DT	0.8992	0.8983	0.8987	0.8985	0.8986	0.8983
	k=3	KNN	0.8720	0.8821	0.8747	0.8716	0.8708	0.8749
O-900%	k=7	NB	0.7159	0.7235	0.7125	0.7112	0.7047	0.7125
	k=7	DT	0.9035	0.9033	0.9035	0.9034	0.9035	0.9032
	k=3	KNN	0.8739	0.8859	0.8739	0.8729	0.8694	0.8739
auto	k=7	NB	0.7124	0.7211	0.7124	0.7096	0.7055	0.7124
	k=3	DT	0.9061	0.9061	0.9061	0.9061	0.9061	0.9062

Table 10. Best model performance results with B-SMOTE parameters configuration

Cases	Parameter	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	k=3, m=3	KNN	0.9078	0.9121	0.9078	0.9076	0.9064	0.9078
R-10%	k=7, m=3	NB	0.7470	0.7519	0.7471	0.7459	0.7439	0.7471
	k=3, m=3	DT	0.9112	0.9112	0.9112	0.9112	0.9111	0.9112
	k=3, m=3	KNN	0.8609	0.8692	0.8609	0.8601	0.8576	0.8609
R-20%	k=5, m=3	NB	0.7202	0.7262	0.7202	0.7183	0.7156	0.7202
	k=3, m=3	DT	0.8555	0.8556	0.8555	0.8554	0.8554	0.8555
	k=3, m=3	KNN	0.9037	0.9042	0.9070	0.9036	0.9061	0.9070
O-800%	k=7, m=3	NB	0.7528	0.7545	0.7462	0.7478	0.7421	0.7462
	k=3, m=3	DT	0.9058	0.9049	0.9056	0.9052	0.9056	0.9056
	k=3, m=3	KNN	0.9080	0.9106	0.9095	0.9080	0.9084	0.9095
O-900%	k=5, m=3	NB	0.7503	0.7546	0.7478	0.7478	0.7441	0.7478
	k=3, m=3	DT	0.9116	0.9115	0.9117	0.9116	0.9117	0.9117
	k=3, m=3	KNN	0.9107	0.9143	0.9107	0.9105	0.9095	0.9107
auto	k=5, m=3	NB	0.7468	0.7518	0.7468	0.7455	0.7435	0.7468
	k=3, m=3	DT	0.9141	0.9141	0.9141	0.9141	0.9141	0.9140

The results of the experiments conducted on each resampling method have the best parameters that differ for each classification model. The parameter configuration brings around the performance data generated. In the SMOTE, the best performance in the parameter k = 3 on the DT method, is represented in Table 9. In the B-SMOTE, the best performance in the parameter (k = 3, m = 3) on the DT method, is represented in Table 10. However, in the SMOTE-ENN prepared in Table 11, the trend has different best k parameters in each of its classifications; that is, the KNN is the best performing in the parameters (k = 3, m = 7), and then NB methods are in the parameters (k = 7, m = 7), and the DT method is in the parameters (k = 7, m =7). All of them have similarities in the parameter m = 7, which is that when its value is higher, all the values are better evaluated.

Table 11. Best model performance results with SMOTE-ENN parameters configuration

Cases	Parameter	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
	k=3, m=7	KNN	0.9834	0.9846	0.9800	0.9822	0.9799	0.9800
R-10%	k=5, m=7	NB	0.7616	0.7711	0.7907	0.7591	0.7832	0.7907
	k=5, m=7	DT	0.9566	0.9530	0.9536	0.9533	0.9535	0.9536
	k=3, m=7	KNN	0.9786	0.9775	0.9739	0.9757	0.9738	0.9739
R-20%	k=7, m=7	NB	0.7693	0.7712	0.8105	0.7622	0.8022	0.8105
	k=7, m=7	DT	0.9488	0.9408	0.9422	0.9415	0.9420	0.9422
	k=3, m=7	KNN	0.9813	0.9819	0.9797	0.9807	0.9796	0.9797
O-800%	k=7, m=7	NB	0.7694	0.7849	0.7897	0.7692	0.7815	0.7897
	k=7, m=7	DT	0.9526	0.9510	0.9510	0.9510	0.9509	0.9510
	k=3, m=7	KNN	0.9825	0.9834	0.9799	0.9816	0.9799	0.9799
O-900%	k=7, m=7	NB	0.7664	0.7782	0.7924	0.7651	0.7847	0.7924
	k=5, m=7	DT	0.9555	0.9532	0.9530	0.9531	0.9529	0.9530
	k=3, m=7	KNN	0.9838	0.9846	0.9809	0.9827	0.9807	0.9809
auto	k=7, m=7	NB	0.7634	0.7727	0.7928	0.7608	0.7852	0.7928
	k=7, m=7	**DT**	**0.9579**	**0.9548**	**0.9544**	**0.9546**	**0.9543**	**0.9544**

The Model Performance Combination Result

The best performance classification model for each balanced data technique presented in Table 9, Table 10, and Table 11 are compared with the performance classification model on the imbalanced dataset, as shown in Table 12. The classification performance on the imbalanced dataset is quite good for accuracy, with the best performance using the KNN method. However, if you pay attention to other performance metrics, the performance is less good, i.e. below 0.67. Meanwhile, classification performance on balanced datasets using the balanced data technique proposed improves performance, i.e. always greater than 0.71.

Table 12. Combined performance results of the resampling

Resampling	Parameter	Cases	Method	Accuracy	Precision	Recall	F1-Score	G-mean	Curve-ROC
No Resampling	-	-	KNN	0.9040	0.6446	0.5567	0.5736	0.3643	0.5600
	-	-	NB	0.8431	0.6051	0.6667	0.6232	0.6314	0.6700
	-	-	DT	0.8627	0.5809	0.5861	0.5833	0.4808	0.5900
SMOTE	k = 3	O-900%	KNN	0.8720	0.8821	0.8747	0.8716	0.8708	0.8749
	k = 7	O-900%	NB	0.7124	0.7211	0.7124	0.7096	0.7055	0.7125
	k =7	***auto***	**DT**	**0.9061**	**0.9061**	**0.9061**	**0.9061**	**0.9061**	**0.9062**
B-SMOTE	k =3, m = 3	*auto*	KNN	0.9107	0.9143	0.9107	0.9105	0.9095	0.9107
	k =5, m = 3	O-900%	NB	0.7468	0.7518	0.7468	0.7455	0.7435	0.7478
	k =3, m = 3	***auto***	**DT**	**0.9141**	**0.9141**	**0.9141**	**0.9141**	**0.9140**	**0.9141**
SMOTE-ENN	k =3, m = 7	*auto*	KNN	0.9838	0.9846	0.9809	0.9827	0,9807	0,9809
	k =7, m = 7	*auto*	NB	0.7634	0.7727	0.7928	0.7608	0.7852	0.7928
	k =7, m = 7	***auto***	**DT**	**0.9579**	**0.9548**	**0.9544**	**0.9546**	**0.9544**	**0.9544**

The best classification model on a balanced dataset using the SMOTE technique is the DT method with the parameter k = 7 and *auto* case. The performance of the DT model on balanced data using the SMOTE technique and parameter k = 7 and *auto* case with Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC values are 0.9061, 0.9061, 0.9061, 0.9061, 0.9061, and 0.9062 respectively. While if compared with the performance of the DT model on unbalanced data, the SMOTE technique improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 4.79%, 35.89%, 35.32%, 35.63%, 46.94%, and 34.89%, respectively.

The best classification model on a balanced dataset using the B-SMOTE technique is the DT method in the parameter (k = 3, m = 3) and *auto* case. The performance of the DT model on balanced data using the B-SMOTE technique in the parameter (k = 3, m = 3) and *auto* case with Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC values are 0.9141, 0.9141, 0.9141, 0.9141, 0.914, and 0.9141 respectively. While if compared with the performance of the DT model on unbalanced data, the B-SMOTE technique improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 5.62%, 36.45%, 35.88%, 36.19%, 47.40%, and 35.46%. However, if the DT model on balanced data with the SMOTE technique compare to B-SMOTE technique improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 0.88%, 0.88%, 0.88%, 0.88%, 0.86%, and 0.86%.

The best classification model on a balanced dataset using the SMOTE-ENN technique is the KNN method in the parameter (k = 3, m = 3) and *auto* case. The performance of the KNN model on balanced data using the SMOTE-ENN technique in the parameter (k = 3, m = 3) and auto case with Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC values are 0.9838, 0.9846, 0.9809, 0.9827, 0,9807, and 0,9809 respectively. While if compared with the performance of the KNN model on unbalanced data, the SMOTE-ENN technique improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 8.11%, 34.53%, 43.25%, 41.63%, 62.85%, and 42.91%. However, if the KNN model on balanced data with the SMOTE technique compare to SMOTE-ENN technique improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 11.36%, 10.41%, 10.83%, 11.31%, 11.21%, and 10.81%. Likewise, if the KNN model on balanced data with the B-SMOTE technique compare to SMOTE-ENN technique improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 7.43%, 7.14%, 7.16%, 7.35%, 7.26%, and 7.16%. Thus, it can be revealed that the balanced data techniques always improve performance classification models. The best of the balanced data technique is the SMOTE-ENN.

CONCLUSION

This research reveals that balanced data techniques improve the performance of classification models. In carrying out the balanced data technique, two approaches are used, i.e. reducing the number of majority classes before generating synthetic data and generating minority classes to achieve ***n***-percent. Based on the experimental results, it is revealed that balanced data always improves the performance of the classification model, both by reducing the number of majority classes before generating synthetic data and by generating minority classes. But the best performance is with the generating minority classes approach.

This study proposes the SMOTE, B-SMOTE, and SMOTE-ENN techniques for generating synthetic data. The best techniques are SMOTE-ENN, B-SMOTE, and SMOTE respectively. Based on the results of experiments balanced data using the SMOTE and B-SMOTE techniques, the best performance was

achieved on the DT classification model. Meanwhile, the best performance using the SMOTE-ENN technique is the KNN method. The SMOTE-ENN technique improves the performance of Accuracy, Precision, Recall, F1-Score, G-mean, and Curve-ROC respectively by 8.11%, 34.53%, 43.25%, 41.63%, 62.85%, and 42.91% if compared to the unbalanced dataset. Therefore, every resampling technique impacts how well the classification model performs. The SMOTE-ENN technique has a better model performance improvement than the other.

REFERENCES

[1] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets : A review," *Science (80-.).*, vol. 30, no. 1, pp. 25–36, 2006.

[2] M. Khushi *et al.*, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[3] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," *Comput. Informatics*, vol. 34, no. 5, pp. 1017–1037, 2015.

[4] F. Shi and G. Fan, "An unbalanced data classification method based on improved SMOTE," in *2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, Aug. 2023, pp. 55–59. doi: 10.1109/ICBASE59196.2023.10303096.

[5] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.

[6] D. Ramyachitra and P. Manikandan, "Imbalanced Dataset Classification and Solutions: a Review," *Int. J. Comput. Bus. Res. ISSN (Online*, vol. 5, no. 4, pp. 2229–6166, 2014.

[7] A. Indrawati, "Penerapan Teknik Kombinasi Oversampling Dan Undersampling Hybrid Oversampling and Undersampling Techniques To Handling Imbalanced Dataset," *JIKO(Jurnal Inform. dan Komputer)*, vol. 4, no. 1, pp. 38–43, 2021, doi: 10.33387/jiko.

[8] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Mar. 2018, pp. 1–11. doi: 10.1109/ICCTCT.2018.8551020.

[9] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, "Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data," in *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, May 2016, pp. 225–228. doi: 10.1109/ICOACS.2016.7563084.

[10] M. R. Kumar, N. Natteshan, J. Avanija, K. R. Madhavi, N. S. Charan, and V. Kushal, "SMOTE-TOMEK: A Hybrid Sampling-Based Ensemble Learning Approach for Sepsis Prediction," in *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, Jul. 2023, pp. 724–729. doi: 10.1109/ICECAA58104.2023.10212208.

[11] W. Jindaluang, V. Chouvatut, and S. Kantabutra, "Under-sampling by algorithm with performance guaranteed for class-imbalance problem," in *2014 International Computer Science and Engineering Conference (ICSEC)*, Jul. 2014, pp. 215–221. doi: 10.1109/ICSEC.2014.6978197.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," p. 37, 2002.

[13] Y. Sun *et al.*, "Borderline SMOTE Algorithm and Feature Selection-Based Network Anomalies Detection Strategy," *Energies*, vol. 15, no. 13, 2022, doi: 10.3390/en15134751.

[14] Ž. Đ. Vujović, "Classification Model Evaluation Metrics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 1–8, 2021, doi: 10.14569/IJACSA.2021.0120670.

[15] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, "Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data," *2019 Int. Conf. Comput. Control. Informatics its Appl. Emerg. Trends Big Data Artif. Intell. IC3INA 2019*, pp. 14–18, 2019, doi: 10.1109/IC3INA48034.2019.8949568.

[16] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–14, 2022, doi: 10.1186/s12911-022-02075-2.

[17] A. Saisundar and D. T, "Accurate Human Palm Recognition System in Cybercrime Analysis using Naive Bayes in comparison with Decision Tree," in *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, Jan. 2023, pp. 1–5. doi: 10.1109/ICECONF57129.2023.10083899.

[18] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.
[19] K. Chomboon, P. Chujai, P. Teerarassammee, K. Kerdprasop, and N. Kerdprasop, "An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm," pp. 280–285, 2015, doi: 10.12792/iciae2015.051.