Multi-Class Imbalance Classification of Diabetes Cases Using Light Gradient Boosting Machine

Indah Manfaati Nur[1,2], *Dedi* Rosadi[1*] and Abdurakhman[1]

[1]Department of Mathematics, Universitas Gadjah Mada, Sleman, Yogyakarta, Indonesia
[2]Department of Statistics, Universitas Muhammadiyah Semarang, Semarang, Central of Java, Indonesia

Abstract. Diabetes is the third leading cause of death in Indonesia. Diabetes is considered a silent killer because it kills slowly and triggers various complications of chronic diseases in the body of the sufferer. Early detection of diabetes is very important to reduce the risk of more serious health problems and reduce the country's socio-economic losses in diabetes management. Machine learning classification is an alternative method that can be used for early detection of diabetes by predicting category labels from observed data. This study aims to classify diabetes using the Light Gradient Boosting Machine (LGBM) method with Synthetic Minority Oversampling Technique of Nominal and Continuous (SMOTENC). The SMOTENC oversampling method is used to handle the imbalance problem in the dataset used, while the LGBM method is used for multi-class classification of diabetes. The results showed that by applying the SMOTENC technique, a more balanced data distribution was obtained, so that when used in the classification process using LGBM, it resulted in high model performance. Based on the confusion matrix, the accuracy value is 90%.

1 Introduction

Diabetes or also called diabetes mellitus (DM) is a chronic disease caused by metabolic disorders with symptoms of hyperglycaemia or increased blood sugar levels due to abnormalities in insulin secretion or decreased insulin sensitivity or both [1]. Diabetes is considered a silent killer because it is the root of all diseases in the human body that kill slowly. Hyperglycaemia that occurs in diabetes results in complications of serious health problems such as coronary arteries, vascular disease, stroke, diabetes neuropathy, kidney failure, blindness, amputation, to reduced life expectancy and can take the lives of sufferers [2]. Diabetes has become a global health emergency due to its rapidly increasing prevalence in the 21st century. In Indonesia, diabetes is a serious health threat as it is the third leading cause of death.

Based on data from the International Diabetes Federation (IDF), Indonesia ranks fifth in the world and first in ASEAN as the country with the most people with diabetes. In addition, Indonesia also occupies the sixth position with the highest number of deaths due to diabetes, reaching 236 thousand people. The number of people with diabetes in Indonesia has increased

*Corresponding author: dedirosadi@ugm.ac.id

rapidly in the last ten years with a total of 19,47 million people with diabetes and a diabetes prevalence of 10,6% in 2021, an increase of 167% compared to the total number of sufferers in 2011 of 7,29 million. It was recorded that in early 2023, people with diabetes reached 35 million people or as much as 13% of the total 270 million Indonesian population [3].

To reduce the risk of greater losses and prevent more fatal complications of the disease, it is necessary to diagnose and prevent diabetes as early as possible. Testing for diabetes using a more accurate clinical examination is costly and time-consuming because the diabetes test procedure needs to go through a series of screening and several tests that must be done periodically. Therefore, a more efficient support effort is needed by using a machine learning classification algorithm approach to classify and perform early detection of diabetes based on important information extracted from the data of previous diabetes.

Classification is a technique in data mining where machine learning models work to predict the correct class label for each observation. The application of machine learning algorithms for diabetes classification has been done [4, 5] using Naïve Bayes, Decision Tree, SVM, and XG Boost methods. Based on the results of research by juniors, using the decision tree algorithm and SVM, the comparison of accuracy values for Decision Tree was higher of 85,28%, while the accuracy value for SVM of 83,85%. Then another research for diabetes classification was carried out by Nasution using the XGBoost method for diabetes classification, it produces an accuracy of 90,10%, while the Naïve Bayes algorithm only produces an accuracy of 79,68%. However, in terms of efficiency and scalability, some of these methods are still unsatisfactory when used for large data and high feature dimensions. To overcome these problems. Therefore, we consider using the Light Gradient Boosting Machine (LGBM) method for diabetes classification.

Research using the LGBM method has been conducted by Michael [6] for breast cancer classification, where from the classification results obtained accuracy of 99.86%, precision of 100%, recall of 99,60% and F1-Score of 99,80%. Another study using LGBM was conducted by Zhang and Gong [7] for the classification of pre-diagnosis of acute liver failure, where the accuracy value of LGBM was 75,81% higher than XG Boost-FA of 67,78%. Although the LGBM method is able to produce very high accuracy performance, it is possible that the classification results obtained tend to achieve the highest level of accuracy because they only take information from the majority class. This happens if there is a case of imbalance in the data used for classification.

In the real world, data sets are often highly skewed, for example in medical data sets such as data on diabetes. Unbalanced data can lead to poor prediction models and accuracy because the resulting model predictions tend to the majority group so that the contribution of minority groups to the model is small [8]. One way to overcome unbalanced data is to perform resampling techniques. We consider using the SMOTENC (SMOTE for Nominal and Continuous) algorithm which can be used to handle data sets with categorical features. Research using SMOTENC and the Random Forest classification algorithm with Gradient Boosting imputation has been conducted by Gök and Olgun [9] for the classification of Covid-19 patient severity predictions using blood samples. Based on the classification results using the SMOTE-NC resampling technique produces a more balanced distribution of target values with a model accuracy result of 98%.

This research is organized as follows. The first section summarizes the motivation for this research, and the second section describes the research methodology used in this study which includes an explanation of the dataset used and the steps in conducting the research. The objectives of this study are given in the third section, which provides the classification results and accuracy rates. The last section will summarize all the findings of this research.

2 Research Methodology

In this section, we will explain the scope of the research and the research procedure. The scope of the research consists of explaining the variables, data, and data sources used. While the research procedure explains the methods and steps used in this research.

2.1 Dataset

The data used in this study is quantitative data in the form of secondary data, namely CDC Diabetes Health Indicators obtained through the UC Irvine Machine learning Repository website source. The original dataset consists of 21 predictor variables (X) and response variables (Y) with 253.680 instance data. The amount of data used was 100.000 data samples taken using simple random sampling. In more detail, the research variables are presented in Table 1.

Table 1. Research Variables.

Variables	Name	Data Types	Label
Y	Diabetes	Category	0 = not diabetes 1 = pre diabetes 2 = diabetes
X_1	High Blood Pressure	Category	0 = low 1 = high
X_2	High cholesterol	Category	0 = not 1 = high
X_3	Check cholesterol in 5 years	Category	0 = not 1 = high
X_4	BMI	Numeric	Body Mass Index
X_5	Smoking	Category	0 = not 1 = yes
X_6	Stroke	Category	0 = not 1 = yes
X_7	Heart Disease or Attack	Category	0 = not 1 = yes
X_8	Physical Activity	Category	0 = not 1 = yes
X_9	Fruit Consumption per Day	Category	0 = not 1 = yes
X_{10}	Vegetable Consumption per Day	Category	0 = not 1 = yes
X_{11}	Consuming Alcohol in Large Amounts	Category	0 = yes 1 = not
X_{12}	Having Health Insurance	Category	0 = not 1 = yes
X_{13}	Did not go to the doctor because of the cost	Category	0 = not 1 = yes
X_{14}	General Health	Category	1 = excellent 2 = very good 3 = good 4 = enough 5 = bad
X_{15}	Mental Health	Numeric	Scale 1-30

Variables	Name	Data Types	Label
X_{16}	Physical Health	Numeric	Scale 1-30
X_{17}	Difficulty Walking	Category	0 = yes 1 = not
X_{18}	Gender	Category	0 = female 1 = male
X_{19}	Age	Category	1 = 18-24 years old 2 = 25-29 years old ⋮ 12 = 75-79 years old 13 = ≥ 80 years old
X_{20}	Education	Category	1 = never went to school 2 = elementary school 3 = junior high school 4 = senior high school 5 = undergraduate 6 = master
X_{21}	Income (USD)	Category	1 = < 10.000 2 = 10.000-15.000 3 = 15.000-20.000 4 = 20.000-25.000 5 = 25.000-35.000 6 = 35.000-50.000 7 = 50.000-75.000 8 = > 75.000

source: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators.

2.2 Research Methods

2.2.1 Synthetic Minority Oversampling Technique of Nominal and Continuous (SMOTE-NC)

SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) is a developmental oversampling algorithm of SMOTE designed to generate synthetic data [10] for datasets containing both continuous and categorical features by generating synthetic data based on k-nearest neighbours using Euclidean distance. SMOTE-NC is specifically designed to handle datasets with both categorical and continuous features by generating synthetic samples for minority classes by considering both continuous and categorical data features. The generation of new synthetic data for continuous features is different from the generation of data on categorical features. The difference is that in continuous features data generation uses Euclidean distance while in categorical data it uses mode values. The steps of generating new synthesis data from continuous and categorical features are presented in Figure 1.

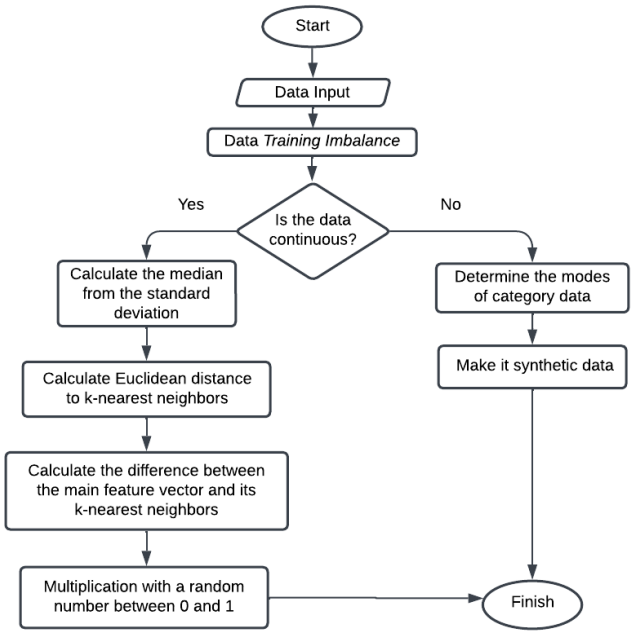

Fig. 1. SMOTE-NC flowchart.

2.2.2 Light Gradient Boosting Machine (LGBM)

Light Gradient Boosting Machine or LGBM is a machine learning framework based on the Gradient Boosting Decision Trees (GBDT) algorithm that can be used for regression, binary classification, multi-class classification, and other machine learning tasks [11, 12]. LGBM can directly handle categorical features without the need to convert them into numerical representations, making it easier when used for datasets with a mixture of numerical and categorical features.

The LGBM architecture uses a leaf-wise technique that allows it to build trees vertically, while other classification algorithms, use a level wise technique that builds trees horizontally. The leaf-wise technique allows LGBM to focus on the most important branches in each iteration, resulting in faster training speed especially for handling large-scale problems. Due to its speed, LGBM can easily handle large datasets with complex features without compromising accuracy. The flowchart for the SMOTENC-LGBM algorithm is shown in Figure 2.

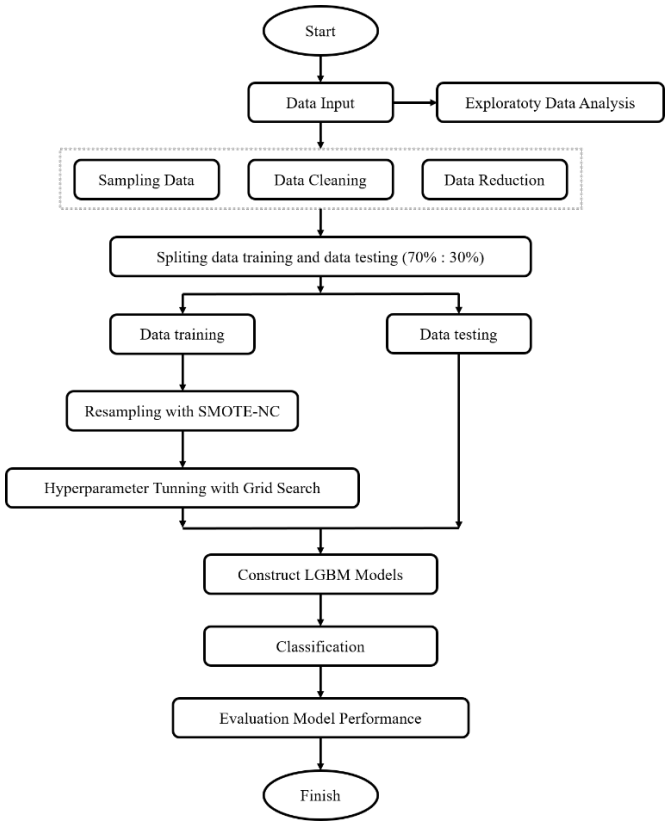

Fig. 2. Flowchart of SMOTENC-LGBM.

Hyperparameter tuning is the process of finding the optimal value combination and identifying the optimal value of the hyperparameters of a machine learning model to improve the performance of the model's results [13]. Hyperparameters are parameters used to control model behaviour, such as model complexity or learning speed. Hyperparameters in machine learning algorithms must be optimized because they can improve the performance of the model and produce more accurate results [14, 15]. Some hyperparameter tuning techniques that can be used for LGBM include grid search. The hyperparameter tuning step with grid search is shown in Figure 3.

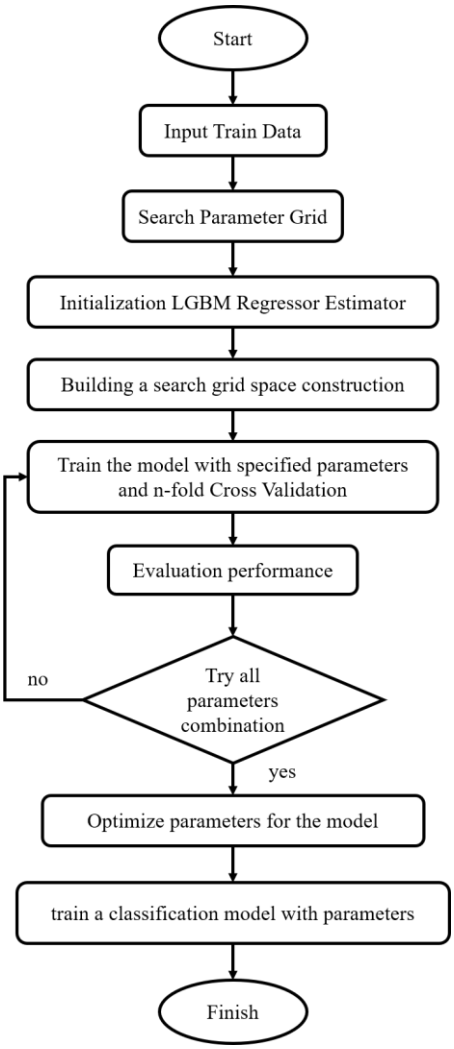

Fig. 3. Flow chart of Hyperparameter Tuning with Grid Search.

3 Results and discussion

In this section, we divide into several parts, namely data pre-processing, determining tuning hyperparameters, and classification results of the SMOTENC-LGBM method.

3.1 Pre-processing data

This research conducts testing with a public dataset, namely the CDC Diabetes Health Indicator data. The data set consists of 253.680 data with 22 variables, consisting of 21 independent variables and 1 dependent variable, namely Diabetes Status. From 253.680 data, this research only uses 100.000 data samples taken using simple random sampling. At this stage, the data reduction stage is carried out, which is to reduce the number of features in the data. Techniques that can be used in data reduction include feature selection. The method that

can be used in feature selection is filter methods using statistics with the Chi-squared test to test the independence and estimated dependence of a class on a feature. The higher the Chi-Square value the feature is more dependent on the response and can be selected for model training. Based on the results of the Chi-Square calculation, we chose to discard 6 variables, namely X_3, X_9, X_{10}, X_{12}, X_{13} and X_{18}, so that the variables used for the next stage are only 15 variables.

Table 2. Chi-Square Results Variables.

Variable	Chi-Square	Variable	Chi-Square	Variable	Chi-Square
X_1	3614,86	X_8	289,97	X_{15}	469,23
X_2	2366,68	X_9	36,05	X_{16}	10614,81
X_3	19,54	X_{10}	40,12	X_{17}	3431,43
X_4	5689,32	X_{11}	373,52	X_{18}	49,72
X_5	115,32	X_{12}	3,11	X_{19}	3876,52
X_6	885,29	X_{13}	59,59	X_{20}	234,08
X_7	2462,45	X_{14}	3466,16	X_{21}	1672,55

In preprocessing, data cleaning was also carried out which included removing duplicate data so that the total amount of data obtained from cleaning was 93981 data. The next stage is to carry out visualization to find out the percentage comparison of the amount of data from each class on the response variable, namely diabetes. This visualization aims to detect cases of imbalance in the data used. The observation results in Figure 4 show that there are indications of imbalance in the class on the diabetes variable.

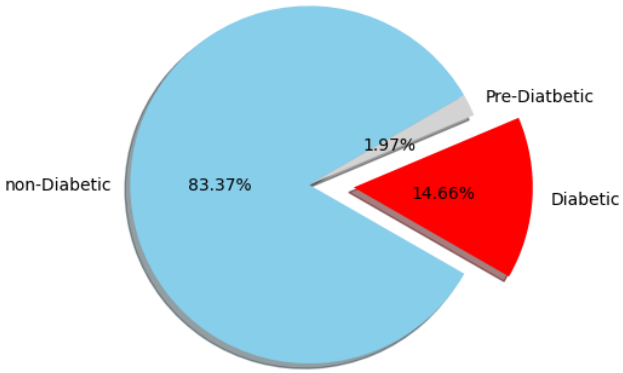

Fig. 4. Visualization of the distribution of the response variable (Diabetes).

Figure 4 shows that diabetes data is still imbalance, where diabetes with category 0 (no diabetes) is 78352 or 83,37%, category 1 (pre-diabetes) is 1847 or 1,97%, and category 2 (diabetes) is 13782 or 14,66%. To overcome this, this study uses the SMOTENC algorithm. The results of handling imbalance variables with the SMOTENC method on the response variable can be seen in Table 3. The same thing is also done on the predictor variables, so that the number of samples on the response and predictor variables is 235.056 samples of data used.

Table 3. SMOTENC Results on Response Variables.

Category	Before Oversampling	After Oversampling
0	78352	54832
1	13782	54832
2	1847	54832

3.2 Selection of Hyperparameter Tuning

The hyperparameter tuning method that can be used for LGBM in this research is grid search. Grid search is an algorithm used to find the combination of hyperparameters that performs best in a machine learning model. This study uses several main parameters that are often used in classification using LGBM. The parameters used in this research and the definitions can be seen in table 4.

Table 4. Definition of LGBM Parameters Used.

Parameter	Definitions
λ_{l1}	Parameters for L1 regulation used to reduce model complexity and overfitting.
λ_{l2}	Parameters for L2 regulation used to control model complexity and prevent overfitting
num_leaves	Parameter that determines the maximum number of leaves in a tree by building the tree in a level-wise manner
feature_fraction	Parameters used to reduce the number of features used in each iteration (tree), reduce overfitting and speed up the learning process
bagging_fraction	Parameter used to control bagging by speeding up training and overcoming overfitting
bagging_freq	Parameters that control the bagging frequency
min_child_samples	Parameter used to set the amount of data needed to form a leaf in the tree

In the hyperparameter tuning technique using grid search, the first step is to initialize the model by entering several values for each parameter that will be used. From the entered values, the grid search algorithm will determine which value is the most optimal. The results of the hyperparameter grid search can be seen in Table 5

Table 5. Hyperparameter Tuning Grid Search Results.

Parameter	Entered parameter values	Best Parameter
λ_{l1}	0,1; 0,3; 0,5	0,1
λ_{l2}	0,1; 0,3; 0,5	0,1
num_leaves	64, 150, 247	247
feature_fraction	0,1; 0,5; 1,0	1,0
bagging_fraction	0,1; 0,5; 1,0	1,0
bagging_freq	5, 10, 20	5
min_child_samples	8, 14, 20	14

Based on the resulting best parameter values, a combination of parameters will be formed for $\lambda l1 = 0.1$, $\lambda l1 = 0.1$, num_leaves = 247, feature_fraction = 1.0, bagging_fraction = 1.0, bagging_freq = 5, and min_child_samples = 14. The combination of parameters formed will be used in the modeling process using the LGBM classification algorithm.

3.3 Classification Results of SMOTENC-LGBM

Based on the parameters that have been obtained previously, the SMOTENC-LGBM analysis is then carried out. Figure 5 shows the Confusion Matrix of the diabetes data classification process using SMOTENC-LGBM.

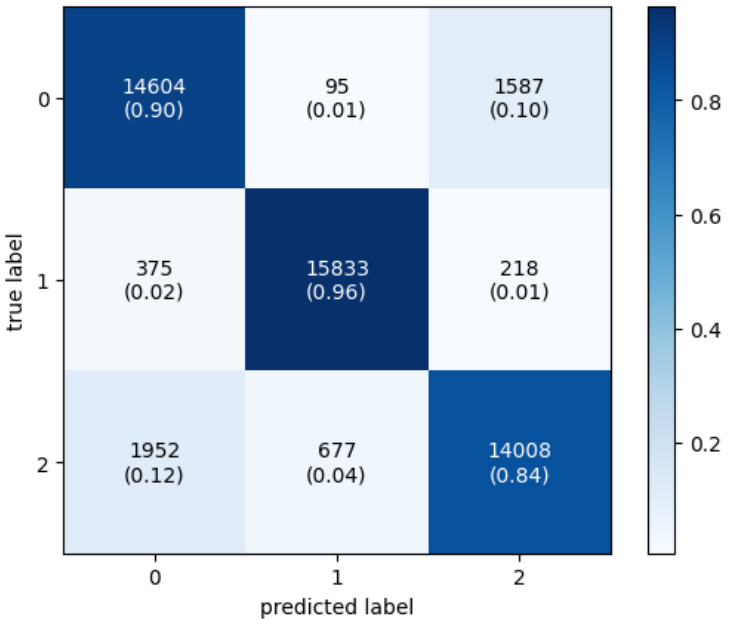

Fig. 5. Confusion matrix of SMOTENC-LGBM algorithm classification results.

Based on the results of the confusion matrix in Figure 5, it can be seen that the classification prediction results obtained for patient data classified as not diabetes are 16931 (with details

of 14604 patients classified as not diabetes, 375 patients classified as pre-diabetes, and 1952 classified as diabetes), 16605 patients classified as pre-diabetes (with details of 95 patients classified as not diabetes, 15833 patients classified as pre-diabetes, and 677 classified as diabetes), and 15813 patients classified as diabetes (with details of 1587 patients classified as not diabetes, 218 patients classified as pre-diabetes, and 14008 classified as diabetes). Based on the results of the confusion matrix in Figure 5, we can also determine the value of precision, recall, f1-Score, and accuracy. For the value of each precision, recall, f1-score, and accuracy can be seen in Table 6.

Table 6. Evaluation Results for SMOTENC-LGBM Classification.

Category	Precision	Recall	F1-Score
Not Diabetes	0,86	0,90	0,88
Pre-Diabetes	0,95	0,96	0,96
Diabetes	0,89	0,84	0,86
Accuracy			0,90

Table 6 shows the results of evaluating the classification model using SMOTENC-LGBM, the precision value obtained which is used to measure the accuracy of the classification model in making positive predictions, the precision value obtained in the Not Diabetes class category is 86%, in the Pre-Diabetes class it is 95%, and in the diabetes class is 89%. The recall value used to measure how well a model correctly identifies the positive class, the recall value obtained in the Not Diabetes class category is 90%, pre-diabetes of 96%, and diabetes is 84%. The F1-Score value used gives an idea of how well the model is at accurately classifying positive and negative reviews. The F1-Score value for the Not Diabetes class category is 88%, pre-diabetes is 96%, and diabetes is 86%.

Accuracy is used to measure how well the classification model predicts the correct class. The higher the accuracy value, the better the model's performance in classification. In Table 6, an accuracy value of 90% is obtained, so it can be concluded that the classification model with SMOTENC-LGBM can predict the diabetes category correctly based on the 15 predictor variables used is 90%.

Table 7. Comparison of model accuracy without and with SMOTENC

	LGBM	SMOTENC-LGBM
Accuracy	0,84	0,90

Table 7 shows a comparison of accuracy results between the LGBM classification model without using the SMOTENS oversampling technique and using SMOTENS to overcome imbalances in the dataset. Based on the comparison, the results show that the LGBM model with SMOTENS produces higher accuracy when compared to the regular LGBM model. Which in the LGBM model obtained an accuracy of 84%.

4 Conclusion

In this study, the application of Light Gradient Boosting Machine (LGBM) as a classification algorithm to produce a model that can predict diabetes. The dataset used has a different amount of data for the target variable (response), so to overcome the problem of imbalance in the amount of data, the SMOTENC technique is used. Then, the variable selection process used in the study selected 15 of the 21 variables based on the variable Chi-Square value. Based on the classification results obtained for patient data classified as not diabetes as many as 16931, patients classified as pre-diabetes as many as 16605, and classified as diabetes as many as 15813. The prediction accuracy result obtained using the SMOTENC-LGBM classification algorithm is 90%. For further research, other classification algorithms can be used to compare with the SMOTENC-LGBM classification algorithm.

The author would like to thank Beasiswa Pendidikan Indonesia (BPI), and The Ministry of Finance, Republic of Indonesia, for their contributions to financing doctoral studies.

References

1. L. Lisna, R. Widiyanti, Jurnal Dinamika Kesehatan jurnal kebidanan dan keperawatan, **11**(1), 147 − 158 (2020)
2. IDF (International Diabetic Federation). Diabetic Atlas (2nd ed. Delice Gan. Brussels, Belgium (2015)
3. Kemenkes RI. Diabetic: Penderita di Indonesia bisa mencapai 30 juta orang pada tahun 2030. https://p2ptm.kemkes.go.id/tag/diabetic-penderita-di-indonesia-bisa-mencapai-30-juta-orang-pada-tahun-2030 (Accessed on 9 Desember 2023.
4. I. M. K. Karo, Hendriyana, J. Teknologi Terpadu **8**(2), 94 – 99 (2022)
5. M.K. Nasution, R. Saedudin, V.P. Widharta, *Perbandingan Akurasi Algoritma Naïve Bayes dan Algoritma Xgboost Pada Klasifikasi Penyakit Diabetic*, in eProceeding of Engineering **8**(5), 9765–72 (2021)
6. E. Michael, H. Ma, H. Li, S. Qi, BioMed. Res. Intl. 2022 (2022)
7. D. Zhang, Y. Gong, IEEE Access **8**, 220990 − 221003 (2020)
8. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, J. Artif. Intell. Res., **16**, 321 – 357 (2002)
9. E.C. Gök, M.O. Olgun, Neural. Comput. Appl., **33**(22), 15693 – 15707 (2021)
10. T. Wongvorachan, S. He, O. Bulut, Information, **14**(1), 54 (2023)
11. A.R. Lubis, S. Prayudani, Y. Fatmi, O. Nugroho, *Classifying News Based on Indonesian News Using Light GBM*, in 2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), 162 − 166 (2022)
12. A. Wibowo, S.R. Purnama, C. Pratama, L.S. Heliani, D.P. Sahara, S.T. Wibowo, Geod. Geodyn., **14**(2), 150 – 162 (2023)
13. S. Khomsah, N.H. Cahyana, A.S. Aribowo, Int. J. Adv. Comput. Sci. Appl. **14**(9), 250 – 256 (2023)
14. J. Wu, X.Y. Chen, H. Zhang, L.D. Xiong, H. Lei, S.H. Deng, J. Electron. Sci. Technol., **17**(1), 26 – 40 (2019)
15. D. Mishra, B. Naik, J. Nayak, A. Souri, P. B. Dash, S. Vimal, Digit. Commun. Netw., **9**(1), 125 – 137 (2023)