Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 12, No. 4, Desember 2024

ISSN 0216 – 0544
e-ISSN 2301– 6914

IMPLEMENTATION OF BALANCING DATA METHOD USING SMOTETOMEK IN DIABETES CLASSIFICATION USING XGBOOST

[a]Fatwa Ratantja Kusumajati, [b]Basuki Rahmat, [c]Achmad Junaidi

[a,b] Departement of Information Technology, Universitas Pembangunan Nasional "Veteran" Jawa Timur Jalan Raya Rungkut Madya, Surabaya, Jawa Timur, 60294, Indonesia
E-mail: fatwaratantja@gmail.com, basukirahmat.if@upnjatim.ac.id, achmadjunaidi.if@upnjatim.ac.id

Abstract

In this research, XGBoost algorithm and the SMOTETomek approach are employed with the objective of enhancing the accuracy of diabetes classification. The study utilises 2,000 patient data points, comprising demographic and medical information, sourced from Kaggle. The dataset employed in this study comprises a number of variables, including pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, Body Mass Index (BMI), diabetes pedigree function, age, and an outcome variable. The latter is a binary classification label, taking on the values 0 and 1. A value of 0 indicates that the patient is not affected by diabetes, whereas a value of 1 indicates that the patient has diabetes. Diabetes represents a significant public health concern in Indonesia. A significant challenge in this study was the imbalanced nature of the dataset, which included a disproportionate number of non-diabetic samples relative to diabetic samples. To address this class imbalance, the researchers employed the SMOTETomek method. SMOTETomek integrates the SMOTE (Synthetic Minority Over-sampling Technique) and Tomek links algorithms to oversample the minority class and remove borderline samples, thereby balancing the class distributions. The SMOTETomek method achieved higher accuracy (95.01%) than SMOTE and the original data (both 92.13%), highlighting the benefits of combining SMOTE with Tomek Links for XGBoost. During testing, SMOTETomek slightly reduced the minority class accuracy (0.97 vs. 0.99 for SMOTE and original data) but maintained strong F1-score and precision, indicating effective handling of data imbalance despite minor trade-offs.

Key words: Balancing Data, Diabetes Classification, SMOTETomek, XGBOOST.

INTRODUCTION

One of the main health issues that has become more prevalent in Indonesia in recent years is diabetes. The primary causes of this illness are dietary and lifestyle modifications. The International Diabetes Federation (IDF) projects that 537 million people worldwide will have diabetes by 2021. With 19.5 million diabetics in 2021, Indonesia will rank sixth in the world for the highest number of diabetes; by 2045, that figure is predicted to rise to 28.6 million. The extremely high cost of diabetes care is one of the effects of the high rate of diabetes. This covers potential consequences and the cost of medical care. Low-income individuals are more susceptible to problems because they have less access to quality medical treatment. If diabetes is not properly managed,

it can lead to serious side effects like heart disease, stroke, kidney damage, and blindness.

Diabetes has been identified as a significant public health concern in Indonesia, with documented cases dating back to the early 1980s. With a prevalence rate of 6.2% among the population, Indonesia has more than 10 million individuals living with diabetes [1].

It is widely acknowledged that class imbalance presents considerable challenges for classification. This is because imbalanced data can result in an imbalanced ratio of minority classes, which may not result in a significant performance loss. However, the same is not true for majority classes, which may result in a significant performance loss [2].

Data imbalance techniques can be divided into several categories, including oversampling, undersampling, and hybrid sampling. This research will employ the latter approach. Hybrid sampling is a combination of oversampling and undersampling techniques. One example of an oversampling technique that will be used is to The SMOTE technique can overcome the issue of oversampling in data sets. However, it still has the disadvantage that the synthetic data generated in the minority class can become noise in the data set because it does not distinguish between data in different classes. Therefore, it will be combined with an undersampling technique called Tomek-Link. Tomek-Link plays a role in removing noise data contained in the majority class. However, Tomek-Link only plays a role in removing instances defined as Tomek-Links. Therefore, using only Tomek-Link will not be sufficient to achieve a balanced data set. By combining SMOTE and Tomek-Link, it is expected that the resulting accuracy performance will be superior to that achieved by using only one of the data balancing techniques [3].

Related works

In a previous study titled "Classification of Diabetes Disease Patients using the C4.5 Algorithm", the accuracy results were found to be 74.08%. The data used in the study consisted of 768 rows and 10 columns [4].

In a study entitled "Application of Oversampling and Undersampling Combination Techniques to Solve Imbalanced Data Problems," the use of SMOTE, SMOTETomek, and SMOTE-ENN techniques in overcoming imbalanced data is compared [5]. Training and testing data were divided in the study using a 70:30 ratio. SVM modeling is applied to the training data; the precision, recall, and F-measure of the created model are determined using the testing data. The results of this comparison indicate that the SMOTETomek technique can enhance the f-Measure for heart disease data by 0.11.

In the research article entitled "A Contemporary Machine Learning Method for Accurate Prediction of Cervical Cancer," the authors employ the SMOTETomek technique for data balancing and utilize a decision tree to train the model. The decision tree with the selected features and SMOTETomek demonstrated superior performance, with an accuracy of 98%, sensitivity of 100%, and specificity of 97%. The decision tree classifier exhibited remarkable proficiency in classification assignment when the features were reduced and the issue of class imbalance was addressed [6].

The following study, entitled "Median-KNN Regressor-Smote-Tomek Links For Handling Missing And Imbalanced Data In Air Quality Prediction" [7], The study uses information on pollutant measurements and ambient air quality conditions at a specific site from the Air Quality Index (AQI) dataset, which may be used to predict air quality. The research team used the SMOTETomek technique to address the issue of data imbalance. The technique comprises estimating air quality values from the Indian AQI dataset using the Naïve Bayes, KNN, and C4.5 algorithms. Based on the evaluation results of the models developed for the study, the KNN method had an accuracy rate of 96.64%, the C4.5 algorithm had an accuracy rate of 100%, and the Naïve Bayes technique had an accuracy rate of 73.96%. This study suggests utilizing the XGBoost algorithm and the SMOTETomek approach as part of a data balancing method to identify diabetes to increase the accuracy.

In a separate study, entitled "SMOTETomek-Based Resampling for Personality Recognition," the influence of imbalanced and overlapping data distributions across two samples on machine learning classification models is examined. The application of data resampling techniques can enhance the accuracy of classification models. In this study, the classification method employed is particle swarm optimization (PSO), and the resampling technique utilized is SMOTETomek. The findings of the study indicate that the

application of the PSO method and data resampling using the SMOTETomek technique results in enhanced accuracy when compared to scenarios where data resampling is not employed. The results demonstrate that the application of SMOTETomek to the my Personality_text dataset yielded an accuracy rate of 75.82%, a notable improvement over the 68.20% accuracy achieved without SMOTETomek. Similarly, on the my Personality dataset, the use of SMOTETomek led to an accuracy rate of 79.96%, a substantial enhancement over the 71.78% accuracy observed in the absence of SMOTETomek [8]. This study introduces a novel approach by combining the SMOTETomek resampling technique with the particle swarm optimization (PSO) method to address imbalanced and overlapping data in personality recognition tasks. Unlike traditional studies that focus on either resampling or classification methods, this research integrates both to enhance model performance. SMOTETomek reduces class imbalance and overlap, while PSO optimizes feature selection, creating a robust framework for improving accuracy. The approach significantly boosts predictive accuracy for the myPersonality_text and myPersonality datasets, demonstrating its potential for use in other domains with similar challenges.

MATERIAL AND METHODS

Classification

Classification is a machine learning that useful for grouping data into predetermined categories or classes [9]. Classification is a form of supervised learning, whereby the training model employs labeled data to map the input to one of several classes. The process entails learning from a dataset of labeled examples, with the ultimate objective of developing the capacity to predict the correct class label for new data that has not yet occurred or existed previously. Classification is the process of organizing objects into distinct groups according to the values of attributes connected to the objects that are being observed. Since every object has unique qualities, categorization can be used to tell one object apart from another [10].

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) represents a significant advancement in the field of gradient boosting. Gradient boosting is an algorithm that can be employed to identify effective solutions to a range of problems, particularly those pertaining to regression, classification, and ranking [11]. XGBoost represents an effective implementation of the Gradient Boosting Decision Tree (GBDT) algorithm. Decision Tree is a method that transforms facts into decision trees, which represent rules that can be interpreted by a human being [12].

When performing a classification, the most commonly used function is Log Loss. The following equation (1) represents the calculation formula that will be employed in the XGBoost algorithm.

$$L(y, f(x) = -\left({}^{1}/_{n}\right) \sum i = 1^{n} [y_i \times \log\left(f(x_i) + (1 - y_i) \times \log\left(1 - f(x_i)\right)\right]) \quad (1)$$

The explanation provided by equation (1) states that $f(x_i)$ is the predicted probability, y_i is the right label, and n is the number of observations. In regression situations, the loss value is typically represented by the mean square error.

$$L\left(y, f(x)\right) = {}^{1}/_{n} \sum i = 1^{n} (y_i - f(x_i)^2) \quad (2)$$

Unlike other methods, XGBOOST employs a gradient-based optimization approach. To lessen the possibility of overfitting and enhance model generalization, XGBOOST employs a technique called regularization. This is demonstrated by equation (3).

$$L\left(y, f(x)\right) + yT + y \sum w_i^2 \quad (3)$$

The L2 regularization term is one of the best and most popular machine learning methods, and it is based on equation (3), where is the complexity parameter and w_i is the i-th feature weight. It has a decent prediction accuracy and a reasonable training time. It provides a variety of distributed and parallel processing methods that can handle high-dimensional data and category features, as well as reduce training time. [13].

This algorithm combines the capabilities of several weaker learning models, with the objective of creating stronger and more

accurate prediction models. In general, decision trees are used to achieve this objective.

An illustration of the XGBoost algorithm in action can be observed in Fig 1.

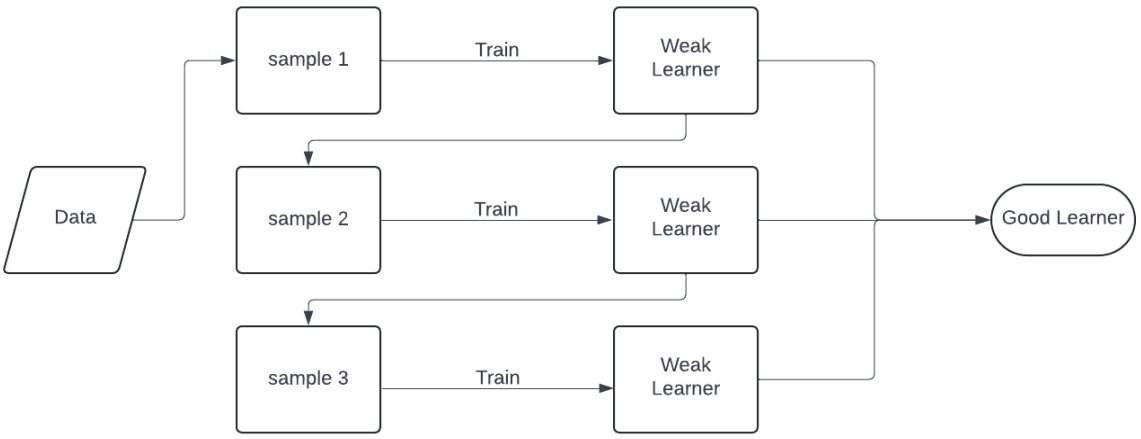

Fig 1. XGBOOST

SMOTE

Machine learning data imbalance is a problem that is addressed by the Synthetic Minority Over-Sampling Technique (SMOTE). The SMOTE method is often used to solve problems related to classification. Examining the minority class samples and then adding additional samples to the dataset that match the minority samples is the basic idea behind the SMOTE approach [14]. While there has been limited discussion of solutions for guided undersampling, oversampling has been the subject of considerable attention due to the success of SMOTE, which has led to the introduction of numerous variants [15]. Finding the K nearest neighbors of the minority sample for x_i is one of the processes in the SMOTE method. [16]. Data imbalance is defined as the phenomenon whereby one class of data exhibits a significantly smaller number of samples than another class. The calculation formula for SMOTE is provided in Equation (4).

$$P_{ij} = x_i + rand(0,1) \times (x_{ij} - x_i) \quad (4)$$

Step 1

For each minority sample (x_i), a distance is calculated from the sample to other samples in the minority set according to a specified rule. This results in the identification of the k nearest neighbors for the given sample.

Step 2

In accordance with the over-sampling magnification, a random subset of k nearest neighbors is selected from each sample (x_i), and these are denoted as (x_{ij}). The pij artificially constructed minority samples are then calculated using the following equation [17].

Tomek-links

Tomek-links One machine learning method for dealing with class imbalance in a dataset is undersampling. The approach is predicated on identifying and removing data from the majority class that is situated close to the class boundary. Once the data from the majority class has been identified, instances from that class are eliminated from each pair. In order to maintain data balance, instances from the minority class are retained. Until no more Tomek-links are found in the data set, this procedure may be done as often as required. There are two methods to use Tomek-Links: as a pre-processing cleaning step or as an undersampling strategy. When this method is used as an undersampling strategy, samples from the majority class are removed. On the other hand, when used as a cleaning step prior to processing, both samples are eliminated. [18].

When majority and minority class instances are eliminated because their borders are not clearly defined, the technique can be used for data cleaning in the context of data pre-processing. Undersampling is one way to apply the approach, It entails getting rid of majority class instances [19].

In essence, the examples that constitute the Tomek-link pair serve to amplify the noise

present within the data distribution. In addition to examples situated at the boundaries and outliers, instances of redundancy also give rise to issues pertaining to class imbalance [20].

The working pattern performed by Tomek-links is as follows: for each sample x in the dataset, Dist(x, y) is the Euclidean distance between sample x and the remaining sample y. [21]. An illustration of the Tomek-Links process can be observed in Fig 2.

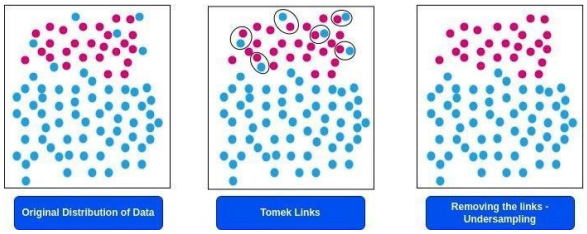

Fig 1. Tomek-Links

To address the issue of overlap caused by oversampling, Tomek links are employed once more in the cleaning stage, with the objective of removing borderline instances from both classes [22]. Consequently, the Tomek-link technique can be employed to eliminate superfluous overlaps between classes subsequent to synthetic sampling, whereby all Tomek-links are eliminated until all nearest pairs are assigned to the same class. The removal of overlapping samples allows for the construction of class clusters that are conducive to training. The formula for calculating the Euclidean distance is provided in equation (5).

$$d(x_i, x_j) = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \qquad (5)$$

SMOTETOMEK

Machine learning models may perform better on minority classes if balancing data techniques are used, such as the synthetic minority oversampling approach combined with Tomek-links (SMOTETomek) [23]. When one class has more samples than the other, there is a data imbalance. This might lead to a machine learning model that is biased in favor of the majority class. SMOTE and Tomek-links data balancing approaches are combined into a single method called SMOTETOMEK. The implementation process for SMOTETOMEK is illustrated in Fig 3.

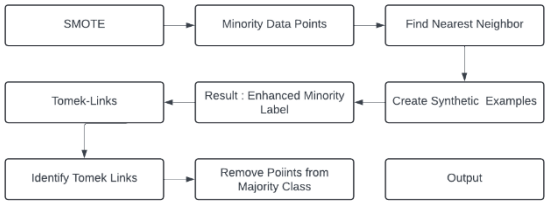

Fig 2. SMOTETOMEK

METHODOLOGY

This research will entail a series of steps, commencing with data acquisition, followed by data preprocessing, model construction, and model evaluation. The methodology employed in this research is illustrated in Fig 4.

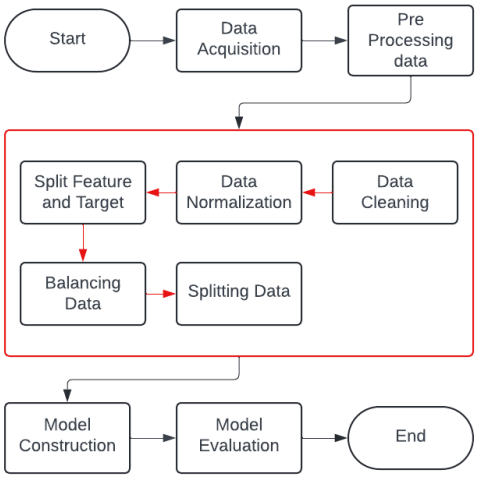

Fig 3. Methodology

Data Acquisition

The The dataset comprises a collection of medical and demographic data pertaining to the patient, along with a binary indicator of diabetes status (whether the patient has diabetes or not). A multitude of features are included in the data set: age, skin thickness (SkinThickness), insulin, blood pressure (BloodPressure), body mass index (BMI), pregnancy history (Pregnancies), diabetes pedigree (DiabetesPedigreeFunction), and blood glucose level (Glucose). The data set utilized in this study was obtained from the Kaggle page, which contains a total of 2,000 data points. We may get the data by clicking on this link: https://www.kaggle.com/datasets/vikasukani/diabetes-data-set. The data utilized in this study are presented in Table 1.

Table 1. The Data Set will be Utilized in The Forthcoming Research Project.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	138	62	35	0	33.6	0.127	47	1
0	84	82	31	125	38.2	0.233	23	0
0	145	0	0	0	44.2	0.63	31	1
0	135	68	42	250	42.3	0.365	24	1
1	139	62	41	480	40.7	0.536	21	0
0	173	78	32	265	46.5	1.159	58	0
4	99	72	17	0	25.6	0.294	28	0
8	194	80	0	0	26.1	0.551	67	0
2	83	65	28	66	36.8	0.629	24	0
2	89	90	30	0	33.5	0.292	42	0

Preprocessing Data

The data pre-processing stage comprises a series of procedures designed to prepare the data for subsequent model training. The data preprocessing procedure employed in this research is illustrated in Fig 5.

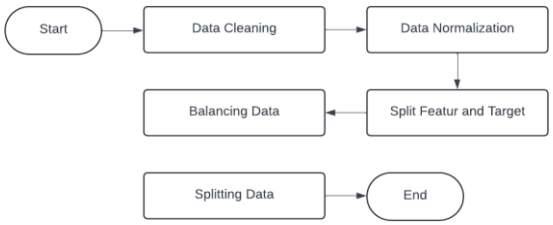

Fig 4. Preprocessing data

Data Cleaning

The preliminary phase of data processing in this study will entail data cleansing. This stage is employed to guarantee the absence of null values, which are commonly designated as missing values and data equating to zero. Furthermore, this procedure entails the elimination of superfluous features, including those pertaining to pregnancy, skin thickness, and diabetes pedigree function. Following the removal of certain features, the data to be utilised in subsequent processes comprises 2000 observations. This is done in order to optimise the performance of the model.

Data Normalization

This stage is employed for the purpose of modifying the value inherent to the feature, which ranges from 0 to 1, through the utilization of the MinMaxScaler. This objective is to prevent the domination of features with higher values than other features, thereby enhancing the performance of the model.

Balancing Data

One of the most important steps in guaranteeing the consistency and uniformity of data inside a dataset is the data balancing stage. Label data or feature data are the sources of the data that needs to be balanced. Outcome data is the target data used in this process. 684 data points are related to label 1 and 1.316 data points are associated with label 0. The SMOTETomek approach will be used by researchers to accomplish data balance.

Splitting Data

The term "split data stage" refers to the process of dividing a dataset into two distinct sections: training data and testing data. The training data is used to train the model, while the testing data is used to assess the model's performance.

Construct a model

This stage will be carried out in several steps, namely the initialization and training of the model. The initialisation stage of the model will be employed to establish certain parameters that will subsequently be utilised during the training process, which will be conducted using the XGBOOST algorithm. Examples of such parameters include the number of decision trees, the learning rate, and the maximum depth. Following the initialization of the model, the researcher will employ cross-validation with a fold of 3. The resulting output will be the accuracy of the classification results. Once the model has been constructed, the evaluation process will commence. This is the stage at which the performance of the model is assessed.

RESULT AND DISCUSSION

Preprocessing Data

Before the modeling process begins, the data pretreatment step is very important. The outcomes of the data preprocessing step are shown in the section that follows. This stage's procedures are helpful for maximizing the model's potential and lowering the overfitting danger.

Cleaning Data

The outcome of the data cleansing process was the removal of features that were not utilized in the classification procedure and the elimination of the value of 0 within the feature. Furthermore, this process guarantees the absence of any missing values, thus ensuring the optimal accuracy of the model. Accordingly, this process represents a pivotal stage in the research process. The outcomes of the data cleansing process are presented in Table 2.

Table 2. The Data Set has been Subjected to a Rigorous Data Cleaning Process.

Glucose	Blood Pressure	BMI	Age	Outcome
138	62	33.6	47	1
84	82	38.2	23	0
145	69.1	44.2	31	1
135	68	42.3	24	1
139	62	40.7	21	0

Data Normalization

By guaranteeing that the data's value range lies between 0 and 1, the data normalization method seeks to lower the danger of overfitting. Table 3 displays the outcomes of the data normalization procedure.

Table 3. The Data Set has been Subjected to a Process Of Normalization.

Glucose	Blood Pressure	BMI	Age	Outcome
0.61	0.387	0.246	0.433	1
0.26	0.592	0.321	0.033	0
0.65	0.461	0.416	0.166	1
0.59	0.449	0.386	0.05	1

Balancing Data

This stage of the research process is very important. The outcomes of the data balancing procedure carried out throughout this study are shown in the section that follows. The results of the data balancing process are presented in Table 4.

Table 4. After Balancing Data

Phase	Data	
	0	1
Before	1316	684
After	1316	1316

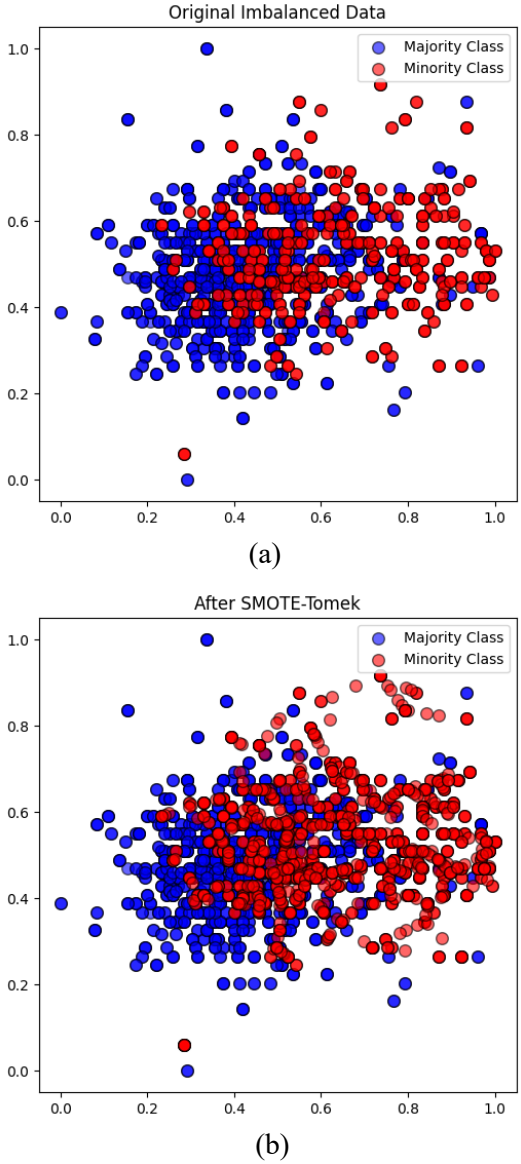

Fig 5. Result SMOTETOMEK: (a) Before SMOTETOMEK, (b) After SMOTETOMEK.

As illustrated in Figure 6, a notable shift in the data is evident. The minority class has been harmonized with the majority class, and the data that exhibited temporal correlations has been excluded.

Splitting Data

The result of the split data process is the division of the data into two distinct parts: the training data, comprising 2105 instances (80% of the total), and the testing data, comprising 527 instances (20% of the total).

Model

In this step, we will commence by declaring the parameters that will be employed for the XGBoost model. There are multiple parameters that will be utilised in the XGBoost model, including max_depth (the maximum depth of each decision tree), n_estimator (the number of decision trees to be generated by the model), and learning rate. Table 5 will illustrate the parameter values that will be utilised in this research.

Table 5. XGBoost Parameter

Method	XGBoost Parameter		
	Max_Depth	LearningRate	n_estimator
SMOTETomek	**3**	**0.01**	**100**
SMOTE	**3**	**0.01**	**100**
Original Data	**3**	**0.01**	**100**

This process constitutes the final stage of the research project. The objective is to create a model that will serve as a tool for classification. The results of the k-cross validation accuracy assessment are presented in Table 5. Fig. 7 shows the steps involved in training the classification model. The following section presents a comparative analysis of the use of K-cross validation in the of SMOTETOMEK, SMOTE, and Original. The results demonstrate that the application of SMOTE and Original techniques yields a distinct level of accuracy compared to that achieved by the SMOTETOMEK approach. In this study, the value of k in the implementation of K-cross validation will be set to k = 3. The outcomes of the K-Croos Validation implementation are presented in Table 6.

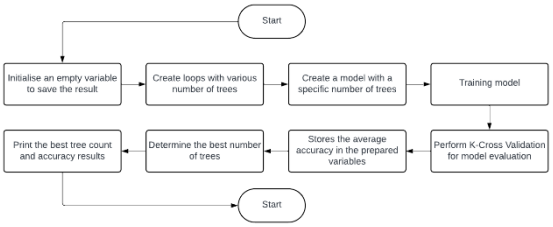

Fig 6. Training model

Table 6. K-Cross Validation

Method	Accuracy			
	Fold 1	Fold 2	Fold 3	Evarage
SMOTE	0.77	0.75	0.77	0.76
SMOTETOMEK	0.83	0.77	0.81	0.80
Original Data	0.77	0.75	0.77	0.76

The outcomes of the K-cross validation process will be employed in the model training procedure. The aforementioned results will be employed as a means of validating the data utilized in the model training process. Upon completion of this phase, the model will undergo training via the XGBoost algorithm. The outcomes of this training will be utilized to assess the model. The subsequent paragraphs present the results of training the model with SMOTETOMEK, SMOTE, and original data.

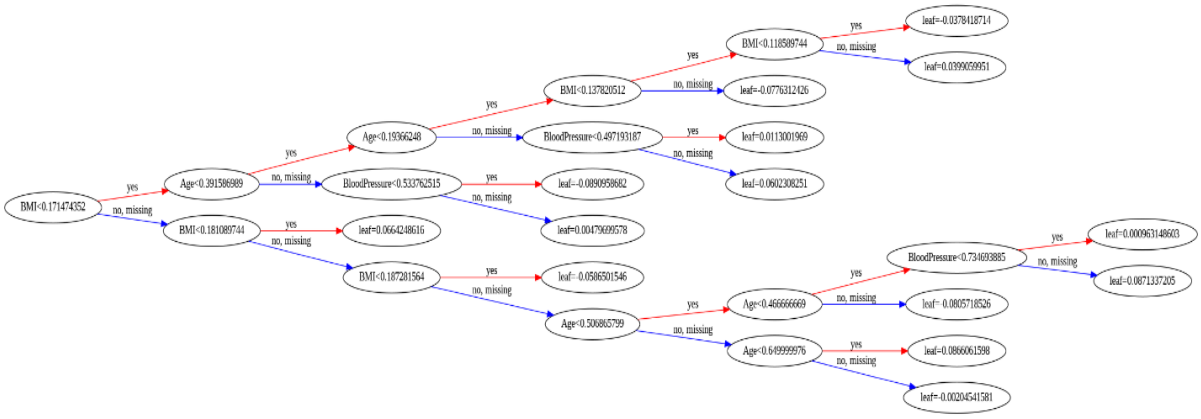

Fig 7. Tree of XGBOOST

Fig. 8 depicts the result of the decision tree generated from the model training. The figure illustrates the optimal decision tree, which has a depth of 99.

Table 7. XGBOOST

Method	Accuracy
SMOTETomek	95.01%
SMOTE	92.13%
Original Data	92.13%

The rationale behind the superior accuracy outcomes achieved through the utilisation of SMOTE lies in its capacity to curtail the influence of noise within the data set. Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic data on the minority class, but this can result in the generation of noise due to the absence of consideration for the surrounding majority class. To address this, Tomek-links can remove samples from the majority class that are in close proximity to the minority class, thereby creating a cleaner and clearer decision boundary. The utilisation of Tomek-links serves to reduce the noise generated from the SMOTE process.

The findings of the training model, as presented in Table 6., indicate that the results of the training model indicate that the SMOTETOMEK technique produces more accurate results than the other two techniques. Therefore, it can be temporarily interpreted that SMOTETOMEK has better performance than the other two techniques.

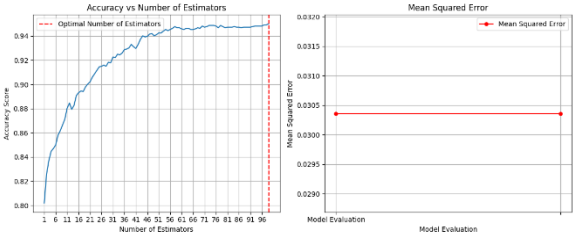

Fig 8. Accuracy and MSE

Fig. 9 illustrates the outcomes of the training model, which achieved an accuracy of 95.01% and a mean squared error (MSE) of 0.0303. These results can be considered indicative of a high degree of accuracy and a low probability of error.

Once this stage is complete, the model evaluation stage will commence. At this stage, a confusion matrix will be produced, which is useful for assessing the prediction results of the model. A comparison of the confusion matrix generated in this study can be seen in Fig 10.

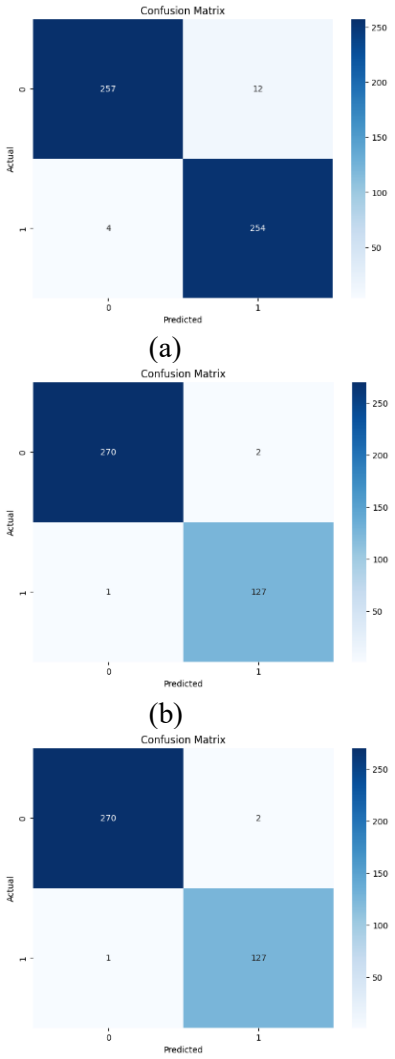

Fig 9. Confusion Matrix: (a) SMOTETOMEK, (b) Original data, (c) SMOTE

The computation of accuracy, F1-score, precision, and recall based on the confusion matrix's findings will constitute the last phase of the investigation. This will provide light on the The model evaluation findings shown in Table 8 can be used to determine the model's effectiveness. The model evaluation results in Table 8 contain several parameters, which represent the prediction results of the model that has been generated using the testing data.

The diminished efficacy of SMOTETomek in the testing phase can be attributed to the potential for Tomek-Link to eliminate data samples that are indispensable for the model to grasp the decision boundary.

Table 8. Model Evaluation

Method	Label	Accuracy	*f1-score*	Precission	Recall
Original	0	0.99	0.99	1.00	0.99
	1		0.99	0.98	0.99
SMOTE	0	0.99	0.99	1.00	0.99
	1		0.99	0.98	0.99
SMOTETOMEK	0	0.97	0.97	0.98	0.96
	1		0.97	0.95	0.98

CONCLUSION

The findings of the research indicate that the utilisation of data balancing techniques, specifically SMOTETOMEK in conjunction with the XGBOOST algorithm, outperforms the use of SMOTE and original data techniques. This is evidenced by the accuracy value, which demonstrates that SMOTETOMEK produces a greater accuracy of 95.01%. Furthermore, the results of the model evaluation provide additional support for this conclusion. The use of SMOTE and the original data techniques yields a considerable discrepancy between the training accuracy and model evaluation results.

This suggests that overfitting may be occurring, but the use of the SMOTETOMEK technique produces a balanced value of training accuracy and model evaluation.

The dataset employed in this study is derived from that used by Kaggle, which may introduce a degree of bias into the data set.

Therefore, it is recommended that future research employ data obtained from official health agencies and utilize a variety of algorithms to achieve enhanced accuracy values.

REFERENCES

[1] T. Ligita, K. Wicking, K. Francis, N. Harvey, and I. Nurjannah, "How people living with diabetes in Indonesia learn about their disease: A grounded theory study," *PLoS One*, vol. 14, no. 2, pp. 1–19, 2019, doi: 10.1371/journal.pone.0212019.

[2] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci. (Ny).*, vol. 513, pp. 429–441, 2020, doi: 10.1016/j.ins.2019.11.004.

[3] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 258–264, 2023, doi: 10.30630/joiv.7.1.1069.

[4] A. A. Robbani, A. M. Siregar, and D. S. Kusumaningrum, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma C4.5," *Sci. Student J. Information, Technol. Sci.*, vol. III, no. 1, pp. 76–82, 2022, [Online]. Available: https://journal.ubpkarawang.ac.id/mahasiswa/index.php/ssj/article/view/424/338

[5] A. Indrawati, "Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 1, pp. 38–43, 2021, doi: 10.33387/jiko.v4i1.2561.

[6] J. Jeremiah Tanimu, M. Hamada, M. Hassan, and S. Yusuf Ilu, "A Contemporary Machine Learning Method for Accurate Prediction of Cervical Cancer," *SHS Web Conf.*, vol. 102, p. 04004, 2021, doi: 10.1051/shsconf/202110204004.

[7] W. Chandra, B. Suprihatin, and Y. Resti, "Median-KNN Regressor-

SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction," *Symmetry (Basel).*, vol. 15, no. 4, 2023, doi: 10.3390/sym15040887.

[8] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.

[9] H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *J. Teknol. Inf. dan Komun.*, vol. 8, no. 2, 2020, doi: 10.30646/tikomsin.v8i2.500.

[10] M. Sholihin, "Classification of Batik Lamongan Based on Features of Color, Texture and Shape," *Kursor*, vol. 9, no. 1, pp. 25–32, 2018, doi: 10.28961/kursor.v9i1.114.

[11] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *J. Math. Theory Appl.*, vol. 4, no. 1, pp. 21–26, 2022, doi: 10.31605/jomta.v4i1.1792.

[12] A. Mariani, R. Siki, N. H. Harani, C. Prianto, and A. Bachelor, "Decision Tree Method, Vendors, Procurement," *J. Ilm. KURSOR*, vol. 10, no. 2, pp. 65–70, 2019.

[13] N. Ahmad, M. J. Awan, H. Nobanee, A. M. Zain, A. Naseem, and A. Mahmoud, "Customer Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques," *IEEE Access*, vol. 12, no. January, pp. 1865–1879, 2024, doi: 10.1109/ACCESS.2023.3334641.

[14] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, "LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM," *Knowledge-Based Syst.*, vol. 196, 2020, doi: 10.1016/j.knosys.2020.105845.

[15] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 9, pp. 6390–6404, 2023, doi: 10.1109/TNNLS.2021.3136503.

[16] V. N. Wijayaningrum, A. P. Kirana, and I. K. Putri, "Student Academic Performance Prediction Framework With Feature Selection and Imbalanced Data Handling," *J. Ilm. Kursor*, vol. 12, no. 3, pp. 123–134, 2024, doi: 10.21107/kursor.v12i3.356.

[17] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-021-03430-5.

[18] R. M. Pereira, Y. M. G. Costa, and C. N. Silla, "MLTL: A multi-label approach for the Tomek Link undersampling algorithm: MLTL: The Multi-Label Tomek Link," *Neurocomputing*, vol. 383, pp. 95–105, 2020, doi: 10.1016/j.neucom.2019.11.076.

[19] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.

[20] D. Devi, S. kr Biswas, and B. Purkayastha, "Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance," *Pattern Recognit. Lett.*, vol. 93, pp. 1339–1351, 2017, doi: 10.1016/j.patrec.2016.10.006.

[21] L. Ai-Jun and Z. Peng, "Research on Unbalanced Data Processing Algorithm Base Tomeklinks-Smote," *ACM Int. Conf. Proceeding Ser.*, pp. 13–17, 2020, doi: 10.1145/3430199.3430222.

[22] Q. Leng, J. Guo, J. Tao, X. Meng, and C. Wang, "OBMI: oversampling borderline minority instances by a two-

stage Tomek link-finding procedure for class imbalance problem," *Complex Intell. Syst.*, vol. 10, no. 4, pp. 4775–4792, 2024, doi: 10.1007/s40747-024-01399-y.

[23] N. A. A. Khleel and K. Nehéz, *A novel approach for software defect prediction using CNN and GRU based on SMOTE Tomek method*, vol. 60, no. 3. Springer US, 2023. doi: 10.1007/s10844-023-00793-1.