Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2024. Vol. 3. No. 3; e-ISSN: 2808-4519

Synthetic Minority Oversampling Technique (SMOTE) for Boosting the Accuracy of C4.5 Algorithm Model

Wiwi Rahayu[1], Deny Jollyta[2*], Alyauma Hajjah[3], Johan[4], Gusrianty[5], Gustientiedina[6], Yulvia Nora Marlim[7], Yenny Desnelita[8]

[1,2,3,4,5,6,7,8]*Institut Bisnis dan Teknologi Pelita Indonesia*
rahayuwiwi224@gmail.com[1], deny.jollyta@lecturer.pelitaindonesia.ac.id[2], alyauma.hajjah@lecturer.pelitaindonesia.ac.id[3], johan@lecturer.pelitaindonesia.ac.id[4], gusrianty@lecturer.pelitaindonesia.ac.id[5], gustientiedina@lecturer.pelitaindonesia.ac.id[6], yulvia.nora@lecturer.pelitaindonesia.ac.id[7], yenny.desnelita@lecturer.pelitaindonesia.ac.id[8]*

Abstract

The low accuracy of the classification model may be caused by dataset imbalance. In reality, low-accuracy models are unacceptable. The purpose of this research is to address data imbalances in an employee performance dataset identified using the C4.5 method. SMOTE is the approach for addressing data imbalance. SMOTE is utilized to generate a large amount of data in the majority or minority class, which has an initial classification accuracy of just 17%. The C4.5 algorithm classifies the new dataset created by SMOTE, which consists of 11 attributes divided three times between training and testing data. The research found that with a 60:40 data split, the classification model had a 69% accuracy. Model accuracy climbed to 76% at 70:30 data splitting, and 86% at the final splitting, which was 80:20. The model's output matches the evaluation findings obtained using the confusion matrix. The research findings indicate that SMOTE may improve classification model accuracy by boosting data in imbalanced classes.

Keywords: *Accuracy; Boosting; C4.5 Algorithm; Confusion Matrix; SMOTE*.

1. Introduction

Data imbalance is a common challenge while attempting to solve classification problems. Unbalanced data happens when the number of observations in the training data for each class label is not balanced, in which case the amount of data in one class is much more than in other classes [1].

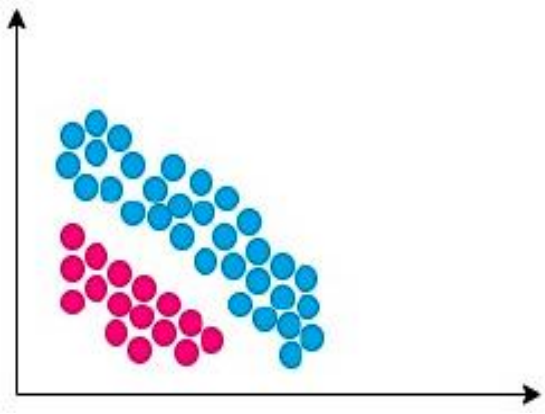

Fig. 1: Class imbalance illustration

Figure 1 shows an example of an imbalance in the amount of data in a class. The class with the most data is termed as the majority class, whereas the class with the least data is termed as the minority class. In general, machine learning classification methods improve accuracy by lowering mistakes without regard for class distribution [2]. Decision Trees and Logistic Regression algorithms are biased in favor of the majority class while ignoring the minority. In its implementation, data imbalance still affects the accuracy of the classification model [3].

In this research, a classification algorithm is utilized to assess employee performance potential as well as the employee's ability to improve in order to deliver greater performance in the future. The classified data is an employee dataset. Employee Identification, Age, Department, Location, Education, Recruitment Type, Job Level, Onsite, Awards, Certification, and Salary are the 11 attributes used to evaluate employee performance. This research purpose to address the imbalance in employee performance statistics in classification results using the C4.5 algorithm by focusing on attribute ratings. SMOTE, or Synthetic Minority Oversampling Technique, is the approach used to overcome this imbalance problem. SMOTE is a technique for balancing the number of sample data distributions in the minority class by picking samples until the total number of samples equals the number of samples in the majority class [4].

A number of previous studies have demonstrated the role of SMOTE in data balancing using classification algorithms. According to research [5], SMOTE improved the accuracy of C4.5, K-Nearest Neighbor, and Naïve Bayes algorithms in classifying new student data by balancing data where the majority class has more data. The accuracy values for each algorithm were 80.72%, 80.46%, and 74.49%, respectively. The same research was carried out by [6] that utilizing SMOTE on a 70-30 data split resulted in increased accuracy using the Naïve Bayes algorithm. In addition, SMOTE was successful in enhancing the accuracy of the Random Forest algorithm's model on heart failure patient data to 84.9% and 90% [7], [8].

Several research have shown that the SMOTE approach impacts the increase in model accuracy achieved by the C4.5 algorithm. Even though it is not statistically significant, SMOTE has been shown to be effective in resolving data imbalances handled by the C4.5 algorithm. The partition of training and testing data was carried out three times, which distinguishes present research from previous studies. SMOTE is used to split training and testing data into 60:40, 70:30, and 80:20 ratios. The findings of this research demonstrate not only the accuracy value of the C4.5 algorithm, but also SMOTE's capacity to increase the quantity of data required to attain the maximum accuracy. Academically, this research leads to a better understanding of SMOTE's performance on the C4.5 algorithm, and practically, the findings assist users address data imbalances more rapidly and correctly.

2. Material

2.1. C4.5 Algorithm

The C4.5 algorithm transforms data into a decision tree that conveys rules. Using the c4.5 algorithm involves many main phases, including transforming the shape of data in the table into a tree model, then converting the tree model into rules, and lastly simplifying the rules [9]. To answer classification issues, the C4.5 approach contains calculation variables that must be satisfied in steps [10], such as:

1. Entropy
   Entropy is the estimated number of bits needed to extract a class from a random quantity of data in a sample space using the formula:

   $$Entrophy\ (s) = -\sum_{i=1}^{m} pi\ log2\ (pi) \quad (1)$$

   Definition (1):
   S = sample space
   p+ = number of positive solutions (supported)
   p- = number of negative solutions (not supported)

2. Gain
   Gain or Information Gain is used for selecting attributes that will serve as branches in a decision tree. The Gain formula is stated as the following equation:

   $$Gain(S, A) = Entropy(S) - \sum_{Values(A)}^{n} \frac{|Sv|}{|S|} Entropy(Sv) \quad (2)$$

   Definition (2):
   S = sample space
   A = attribute
   V = a possible value for attribute A
   Value(A) = possible set for attribute A
   |Sv| = number of samples for V value
   |S| = the total number of data samples
   Entropy (Sv) = entropy for a sample that has a V value

3. Split Info
   Split Info is a method for normalizing a collection of attributes that form branches in a tree structure. Split Info represents Entropy or prospective information using the following equation:

   $$SplitInfo(S, A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (3)$$

   Definition (3):
   S = sample space
   A = attribute
   S_i = number of samples for attribute i

4. Gain Ratio
   Split Info will separate attributes and compare the Gain for each attribute. The property with the greatest Gain Ratio value will be chosen as the test attribute for the node, also known as the internal node. The gain ratio is stated using the following equation:

$$gainratio(a) = \frac{gain(a)}{split(a)} \quad (4)$$

   Definition (4):
   a = attribute
   gain(a) = information gain on attribute a
   Split(a) = split information on attribute a

2.2. SMOTE

SMOTE goals to balance class distribution by randomly boosting and duplicating minority class samples. SMOTE creates new minority instances by mixing existing ones. It employs linear interpolation to generate virtual training records for the minority class. These synthetic training records are constructed randomly selecting one or more of the k-nearest neighbors for each example in the minority class [11]. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data. The following pseudocode demonstrates how SMOTE works:

> Step 1: Setting the minority class set A, for each [Tex]$x \in A$[/Tex], the k-nearest neighbors of x are obtained by calculating the Euclidean distance between x and every other sample in set A.
>
> Step 2: The sampling rate N is set according to the imbalanced proportion. For each [Tex]$x \in A$[/Tex], N examples (i.e x1, x2, …xn) are randomly selected from its k-nearest neighbors, and they construct the set [Tex]A_1[/Tex] .
>
> Step 3: For each example [Tex]$x_k \in A_1$[/Tex] (k=1, 2, 3…N), the following formula is used to generate a new example: [Tex]$x' = x + rand(0, 1) * \mid x – x_k \mid$[/Tex] in which rand(0, 1) represents the random number between 0 and 1.

Fig. 2: SMOTE pseudocode

SMOTE balances data in a variety of domains, including health [7], economics [12], and transportation [13]. In current research, SMOTE generates employee performance data in three splitting data. This follows in order to obtain the highest accuracy for the C4.5 algorithm model.

2.3. Confusion Matrix

The Confusion Matrix is an approach for determining model accuracy in data mining methods. The matrix confusion approach results in accuracy, precision, and recall numbers. Accuracy is defined as the percentage of correctly categorized data that has been validated against the classification findings. Precision is defined as the proportion of cases expected to be positive despite the fact that the actual data is likewise positive. The fraction of accurately anticipated positive instances is known as recall or sensitivity [14]. The classification rule with the best accuracy is the optimal classification rule [15]. The accuracy metric displays the overall number of correct predictions made by the predictive model out of all predictions. It is computed as follows:

$$Accuracy = \frac{TPR+TNR}{TPR+FPR+TNR+FNR} \quad (5)$$

$$Precision = \frac{TPR}{TPR+FPR} \quad (6)$$

$$Recall = \frac{TPR}{TPR+FNR} \quad (7)$$

Definition (5)-(7):
TPR (True Positive Rate) = positive predictive model is true, FPR (False Positive Rate) = positive predictive model is false, TNR (True Negative Rate) = negative predictive model is true, FNR (False Negative Rate) = negative predictive model is false.

3. Research Method

The research stages were organized in the following order to meet the research purposes:

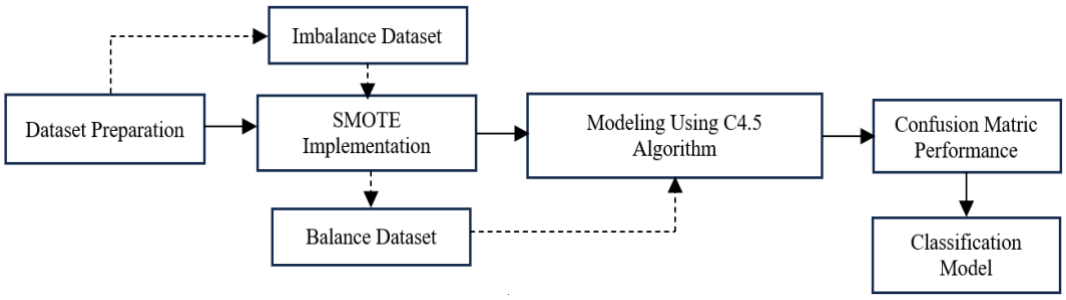

Fig. 3: Research stages

Figure 3 displayed that the research begins by preparing the dataset for classification. The secondary data utilized was obtained from the Kaggle.com website. Data is chosen after it has been processed in the Knowledge Discovery Database (KDD). Next, SMOTE is used to do a data balance check. Classes having an unequal quantity of data will be balanced by creating a specific amount of data until the balance is obtained. In this research, SMOTE was separated into three categories of training and testing data, with percentages of 60-40, 70-30, and 80-20. The goal of releasing this data is to assess the gain in accuracy caused by adding datasets using SMOTE. The C4.5 algorithm is used to classify the freshly produced dataset. The generated model was then evaluated with a confusion matrix. The model with the best accuracy is chosen as the classification model. Python is the computer language used to process the research data.

4. Result and Discussion

4.1. Initial Classification Process

The first stage is classifying the employee performance dataset obtained from Kaggle. There are 500 datasets. Datasets are selected through processing stages. The dataset consists of 11 attributes, namely Employee Identification (emp_id), Age (age), Department (dept), Location (location), Education (education), Recruitment Type (recruitment_type), Job Level (job_level), Onsite (onsite), Awards (awards), Certification (certification), and Salary (salary), with 5 ratings becoming classes, as shown in Table 1.

Table 1: Class of Dataset

Class	Classification	Information
1	Very Poor	Employees have very poor performance
2	Poor	Employees have poor performance
3	Moderate	Employees have modest performance.
4	Good	Employees have good performance
5	Excellent	Employees have excellent performance

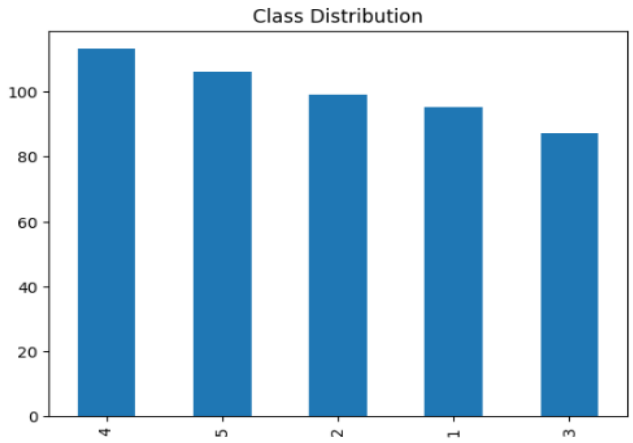

Fig. 4: Class distribution

Table 1 shows the rating assigned to measure employee performance. Employee performance evaluations are based on ratings that take into consideration the assessment qualities, and the data is delivered in different quantities to each class. Figure 4 illustrates the data distribution for each class. The C4.5 algorithm, which relates to equations (1)-(4), was then used to classify 500 datasets, resulting in a model assessment with a very low value. The evaluation results, which are processed using Python, are shown in the form of a confusion matrix with a calculation method according to equations (5) to (6), as follows:

```
print("\nEvaluasi Model:")
print(f"Accuracy: {accuracy:.2f}")
print("Classification Report:")
print(classification_rep)
```

```
Evaluasi Model:
Accuracy: 0.17
Classification Report:
              precision    recall  f1-score   support

           1       0.27      0.18      0.22        22
           2       0.07      0.04      0.05        23
           3       0.16      0.15      0.15        20
           4       0.10      0.27      0.14        11
           5       0.29      0.25      0.27        24

    accuracy                           0.17       100
   macro avg       0.18      0.18      0.17       100
weighted avg       0.19      0.17      0.17       100
```

Fig. 5: Initial confusion matrix

According to Figure 5, the initial classification yields a model accuracy of 17%, whereas the precision and recall for each class are less than 30%. This demonstrates that the present dataset's percentage and sensitivity are quite low, and there is a chance of data imbalance between the majority and minority class. The model was not approved due to the low evaluation scores.

4.2. SMOTE Implementation

SMOTE was used to balance data in order to improve model accuracy. At this point, the SMOTE experiment was repeated three times using minority oversampling to produce data on the initial dataset, splitting the training and testing data into 60:40, 70:30, and 80:20 ratios. The outcomes of the three trials were immediately assessed with the C4.5 algorithm to check whether there was an improvement in data and class balance based on the accuracy findings. In the first trial, SMOTE was able to add 65 data points, with each class contributing equally. SMOTE was successful in increasing data to 199 datasets. Furthermore, during the third phase, SMOTE recorded 699 data points. Total data training and testing after the SMOTE process can be shown in Table 2.

Table 2: SMOTE Results

Testing	Splitting Data	Initial Dataset	Additional Dataset
1	60:40	500	65
2	70:30	500	199
3	80:20	500	699
Total			**963**

According to Table 2, the dataset split is achieved by combining the Additional Dataset and the Initial Dataset, resulting in a total of 1463 datasets to be evaluated with the C4.5 method. This sum was then divided into three identical data splits, as shown in Table 2, and evaluated with the C4.5 algorithm, yielding accuracy of 69%, 76%, and 86%, respectively. Table 3 shows a comparison of the performance of the new dataset.

Table 3: New Dataset Performances

Testing	Splitting Data	New Dataset	Data Training: Data Testing	Accuracy
1	60:40	1463	878 : 585	69%
2	70:30	1463	1024 : 439	76%
3	80:20	1463	1170 : 293	86%

In addition to accuracy, the confusion matrix displays the whole model evaluation based on the tests performed.

```
Evaluasi Model untuk Pembagian Data 60:40:
Accuracy: 0.69
Classification Report:
              precision    recall  f1-score   support

           1       0.60      0.70      0.65        74
           2       0.67      0.60      0.63        87
           3       0.62      0.58      0.60        86
           4       0.80      0.77      0.78       205
           5       0.64      0.69      0.67       108

    accuracy                           0.69       560
   macro avg       0.66      0.67      0.66       560
weighted avg       0.69      0.69      0.69       560
```

(a)

```
Evaluasi Model untuk Pembagian Data 70:30:
Accuracy: 0.76
Classification Report:
              precision    recall  f1-score   support

           1       0.62      0.83      0.71        48
           2       0.76      0.53      0.62        70
           3       0.82      0.71      0.76        69
           4       0.85      0.82      0.83       152
           5       0.70      0.86      0.77        81

    accuracy                           0.76       420
   macro avg       0.75      0.75      0.74       420
weighted avg       0.77      0.76      0.76       420
```

(b)

```
Evaluasi Model:
Accuracy: 0.86
Classification Report:
              precision    recall  f1-score   support

           1       0.80      0.77      0.79        31
           2       0.86      0.79      0.82        47
           3       0.80      0.89      0.84        45
           4       0.89      0.89      0.89        93
           5       0.88      0.88      0.88        64

    accuracy                           0.86       280
   macro avg       0.85      0.84      0.84       280
weighted avg       0.86      0.86      0.86       280
```

(c)

Fig. 6: Confusion matrix for SMOTE tests. (a) illustrated a 60:40 data split, (b) illustrated a 70:30 data split, (c) illustrated an 80:20 data split

Figure 6 shows a confusion matrix for each model, which is made up of three separate data divisions. The quantity of datasets used in the SMOTE experiment boosts model accuracy. The data distribution is even for the majority and minority classes. The problem of data and class imbalance can be attributed to the fact that the data is not ideal. Three trials proved that the maximum and acceptable model accuracy was reached. A data split of 80:20 yields the maximum accuracy. The following is an explanation of the model accuracy findings for employee performance classification with the greatest accuracy, which is 86%:

1. Class 1: 80% of employees who were predicted to have very poor performance, turned out to have very poor performance, with 77% of the total employees who were supposed to have very poor performance successfully identified.
2. Class 2: 86% of employees who were predicted to perform poorly, turned out to have poor performance, 79% of the total employees who should have performed poorly were identified.
3. Class 3: 80% of employees predicted to have average performance have average performance. As many as 89% stated that of the total employees who should have moderate performance, the model succeeded in identifying them.
4. Class 4: 89% of employees who are predicted to have good performance, perform well. As many as 89% stated that of the total employees who should have good performance, the model succeeded in identifying them.
5. Class 5: 88% of employees who were predicted to have excellent performance turned out to have excellent performance, identified 88% of the total employees who should have excellent performance.

To see the data distribution and evaluation coming from the SMOTE procedure over three tests, the SMOTE data findings were pooled and re-evaluated to produce a comparative visualization of precision and recall.

```
import pandas as pd
```

```
df_asli = pd.read_excel("shuffled_dataset.xlsx")
```

```
df_smote = pd.DataFrame(X_resampled, columns=X.columns)
df_smote['rating'] = y_resampled
```

```
df_gabung = pd.concat([df_asli, df_smote], ignore_index=True)
```

```
df_gabung.to_excel("path_ke_dataset_gabungan 4.xlsx", index=False)
```

Fig. 7: SMOTE results merged into new dataset

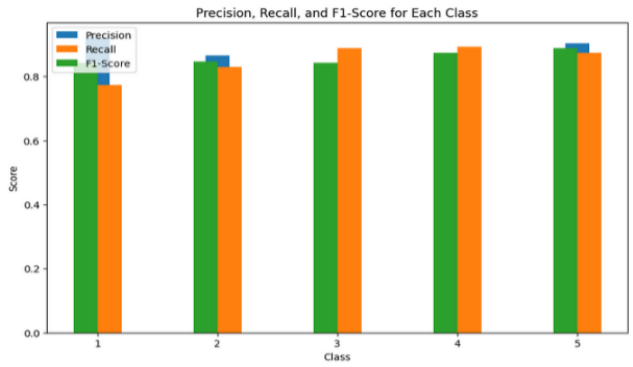

Fig. 8: Classification model evaluation with new datasets

According to the tests, increasing the amount of data increases the accuracy of the C4.5 algorithm model. Figures 7 and 8 provide visualizations of Figure 6 (c), which represents the SMOTE result. The C4.5 algorithm model's accuracy improves as the amount of data generated by the SMOTE process increases. In this research, SMOTE also performed swiftly, with an average training time of less than 1 second.

5. Conclusion

This research demonstrated that SMOTE may overcome data imbalances in employee performance datasets identified with the C4.5 algorithm utilizing various data splitting methods. SMOTE generated data can quickly improve classification model accuracy. This research's findings led to the development of a free testing approach that involves dividing data until it reaches the required accuracy. This research may still be extended in the pre-processing stages of attribute selection to improve model accuracy, such as employing the Principal Component Analysis (PCA) method. Furthermore, it is expected that evaluating the SMOTE dataset using Exploratory Data Analysis (EDA) will verify data correctness in unbalanced classes.

Acknowledgement

We would like to thank Institut Bisnis dan Teknologi Pelita Indonesia in Pekanbaru for supporting the completion and publication of this research.

References

[1] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, “Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks,” *Appl. Sci.*, vol. 13, no. 4006, p. 34, 2023, doi: 10.3390/app13064006.

[2] A. S. Hussein, T. Li, C. W. Yohannese, and K. Bashir, “A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving

SMOTE," *Int. J. Comput. Intell. Syst.*, vol. 12, no. 2, pp. 1412–1422, 2019, doi: 10.2991/ijcis.d.191114.002.
[3] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 54, p. 15, 2023, doi: 10.3390/info14010054.
[4] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 196–201, 2019, doi: 10.29207/resti.v3i2.945.
[5] R. A. Nurdian, Mujib Ridwan, and Ahmad Yusuf, "Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 1, pp. 24–32, 2022, doi: 10.28932/jutisi.v8i1.4004.
[6] A. A. Arifiyanti and E. D. Wahyuni, "Smote: Metode Penyeimbang Kelas Pada Klasifikasi Data Mining," *SCAN - J. Teknol. Inf. dan Komun.*, vol. 15, no. 1, pp. 34–39, 2020, doi: 10.33005/scan.v15i1.1850.
[7] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, "An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction," *Sci. Program.*, vol. 2021, no., p. 12, 2021, doi: 10.1155/2021/6621622.
[8] F. P. Arifianti and A. Salam, "Automated Maintenance System For Freshwater Aquascape Based On The Internet Of Things (Iot)," *Adv. Sustain. Sci. Eng. Technol.*, vol. 6, no. 1, pp. 02401025-01 ~ 02401025-08, 2024, doi: 10.26877/asset.v6i1.17951.
[9] D. Jollyta, P. Prihandoko, A. Hajjah, E. Haerani, and M. Siddik, *Algoritma Klasifikasi Untuk Pemula Solusi Pyhton dan RapidMiner*, Pertama. Yogyakarta: Deepublish, 2023.
[10] Y. Fakir, M. Azalmad, and R. Elaychi, "Study of The ID3 and C4.5 Learning Algorithms," *J. Med. INFORMATICS Decis. Mak.*, vol. 1, no. 2, pp. 29–43, 2020, doi: 10.14302/issn.2641.
[11] A. Muneer, R. F. Ali, A. Alghamdi, S. M. Taib, A. Almaghthawi, and E. A. Abdullah Ghaleb, "Predicting customers churning in banking industry: A machine learning approach," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 1, pp. 539–549, 2022, doi: 10.11591/ijeecs.v26.i1.pp539-549.
[12] M. H. Kotb and R. Ming, "Comparing SMOTE Family Techniques in Predicting Insurance Premium Defaulting using Machine Learning Models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 9, pp. 621–629, 2021, doi: 10.14569/IJACSA.2021.0120970.
[13] D. B. Prakash, C. Narendar, D. Sumalatha, and A. Professor, "Handling Class Imbalance Problem in Machine Learning Using Synthetic Minority Oversampling Technique (Smote)," *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. 03, no. 03, pp. 1863–1868, 2021, [Online]. Available: www.irjmets.com.
[14] H. Yun, "Prediction model of algal blooms using logistic regression and confusion matrix," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 3, pp. 2407–2413, 2021, doi: 10.11591/ijece.v11i3.pp2407-2413.
[15] K. A. Abbas *et al.*, "Unsupervised machine learning technique for classifying production zones in unconventional reservoirs," *Int. J. Intell. Networks*, vol. 4, no. October 2022, pp. 29–37, 2023, doi: 10.1016/j.ijin.2022.11.007.