Data Mining for Potential Customer Segmentation in the Marketing Bank Dataset

Maulida Ayu Fitriani[1], Dany Candra Febrianto[2]

[1]*Informatics Engineering, Universitas Muhammadiyah Purwokerto, Indonesia*
[2]*Electrical Engineering and Information Technology, Universitas Gadjah Mada, Indonesia*
[1]maulidaayuf@ump.ac.id, [2]dany.candra@mail.ugm.ac.id

Abstract - Direct marketing is an effort made by the Bank to increase sales of its products and services, but the Bank sometimes has to contact a customer or prospective customer more than once to ascertain whether the customer or prospective customer is willing to subscribe to a product or service. To overcome this ineffective process several data mining methods are proposed. This study compares several data mining methods such as Naïve Bayes, K-NN, Random Forest, SVM, J48, AdaBoost J48 which prior to classification the SMOTE pre-processing technique was done in order to eliminate the class imbalance problem in the Bank Marketing dataset instance. The SMOTE + Random Forest method in this study produced the highest accuracy value of 92.61%.

Keywords: data mining, bank marketing, SMOTE

I. INTRODUCTION

A good marketing strategy is needed in the industry to increase profits. The banking industry is no exception, product introduction is one of the marketing efforts that are considered effective because banks can conduct market analysis by utilizing the information technology space that can assist in making decisions [1]. One of the efforts made by banks to introduce products to customers is direct marketing. Direct marketing is the process of identifying possible customers to buy or use a product and promoting other products owned by the bank to predetermined customer segments [2]. Direct marketing can be done via telephone and direct email to prospective customers which allows the prospect to decide whether to take the product being offered or not [3]. Another benefit of direct marketing is to strengthen the relationship between banks and customers. So that it can increase the business continuity of an industry.

The direct marketing process, which is carried out via telephone, sometimes officers have to contact a customer or prospective customer more than once to ascertain whether the customer or prospective customer is willing to subscribe to a time deposit. This activity is deemed inefficient and requires a lot of money. This inefficient marketing process is due to the fact that officers do not know the characteristics of clients who have the potential to subscribe to time deposits.

Technology enables marketing by focusing on maximizing the value of a customer's subscription period through the evaluation of available information and customer metrics, making it possible to build a longer and more closely aligned relationship with business demands [4].

With data sources that can be fully used by banks for customer segmentation based on warehouse and data mining processes [5]. Data warehouse and data mining processes that are effective customer segmentation can help banks find potential customers accurately, and help banks to develop new products that can meet consumer demands. In data mining the classification process is an important job in data mining.

In classification, a combination of input variables is used to build a model, and a good model will provide predictions with accuracy that will produce data output in the form of categorical variables [6]. In other words, classification aims to build a model based on input data that can study unknown basic functions and map several input variables, which characterize an item with one output labeled target eg sales type: "failed" or "success") [4].

Several studies in the area of direct marketing classification at banks have been done before. Ref. [7] resolved the imbalance problem of target data. The imbalance in the target data can reduce predictions in making decisions because it tends to produce predictions for the majority class rather than the minority class. In his research, the SMOTE preprocessing method was carried out to balance the target data, and to classify using several classification methods. This work using Naïve Bayes produced an accuracy value of 88.3%, SVM got 89.68% and Decision Tree got the highest result, which was 92.25%.

Ref. [8] classified potential customers on the Bank Direct Marketing dataset using SVM combined with the AdaBoost algorithm. At the time of pre-processing data, he has selected or compressed the data to be 9280 data, because the comparison of class data had a significant difference in numbers. In conducting training the data is divided into 70% training data and 30% test data. From the proposed method, the results obtained an accuracy of 95% and a sensitivity value of 91.65%.

Ref. [9] implemented a neural network in bank marketing data. Zhang used data encryption to train the model, and compare the results with other classification methods. However, the final result still gets a fairly low accuracy, namely 54%, this is due to the absence of a feature selection process carried out in this study.

The purpose of this paper is to find the most appropriate classification method for classifying customer responses to direct telephone marketing by banks in order to increase customer response from bank marketing officers. Therefore, accuracy is an important factor for determining direct marketing results. Comparison of classification techniques in this paper is expected to help determine models in data mining with the best accuracy for classifying the appropriate targets on the Bank Direct Marketing dataset.

II. METHOD

The dataset in this paper is the Bank Marketing Dataset which is the marketing data for a bank in Portugal [3]. The dataset was obtained from the University of California at Irvine (UCI) Machine Learning Repository. Bank Marketing data contains 17 attribute data, 45,211 instance data and there are 2 classes. The descriptions of the datasets are described in Table I.

This study proposes a comparison of classification methods in data mining for Bank Marketing. The methodology used in this research is depicted in Fig. 1. It is explained that the first thing to do is to collect the direct marketing bank dataset and then extract the data. This study also compares the final results if the training data is pre-processed and not pre-processed. Pre-processing is done for class balancing using the SMOTE method. Furthermore, classification and testing is carried out with test data using 10 fold cross validation. After evaluation, a comparison is made between various classification methods and the effect of pre-processing on classification.

A. Pre-Processing

1) SMOTE: In the Bank Marketing dataset, there is an imbalance of data in the target class y. The effect of using unbalanced data to make the model ineffective on the results obtained. Algorithm processing that ignores data imbalance tends to be dominated by the major class and ignores the minor class[10]. Whereas in fact, in many cases the minority class is more concerned, because the cost of misclassification of the minority class is usually much higher than the majority class [11].

TABLE I
ATTRIBUTES ON THE BANK MARKETING DATASET

No	Attribut	Type	Values
1	Age	Numeric	Real
2	Job	Categorical	Admin, Unknown Unemployed, Management, Housemaid, Enterprenuer, Student, Bluecollar, Self-empoyed, Retired, Technican, Services
3	Marital	Categorical	Maried, Diforced, Single, Widowed
4	Education	Categorical	Secondary, Unknown, Primary, Tertiary
5	Default	Binary	Yes, No
6	Balance	Numeric	Real
7	Housing	Binary	Yes, No
8	Loan	Binary	Yes, No
9	Contact	Categorical	Unknown, Telephone, Cellular
10	Day	Numeric	Real
11	Month	Categorical	Jan, Feb …. Nov, Dec
12	Duration	Numeric	Real
13	Campaign	Numeric	Real
14	Pday	Numeric	Real
15	Previous	Numeric	Real
16	Poutcome	Categorical	Unknown, Failure, Success
17	Y	Binary	Yes, No

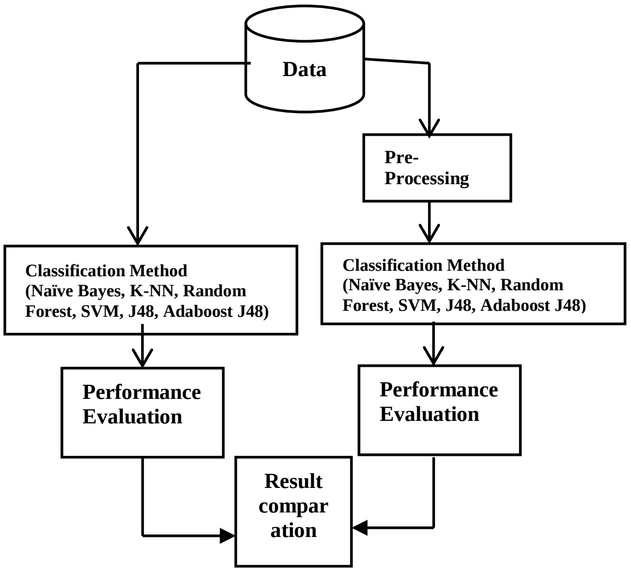

Fig. 1 Research flow

SMOTE (Synthetic Minority Oversampling Technique) is a method to solve class imbalance problems [7]. The principle of the SMOTE Method is to increase the number of data from the minority class so that it is equal to the majority class by generating artificial data. The artificial or synthetic data is made based on the k-nearest neighbor. Generating artificial data with numeric scale is different from categorical. Numerical data are measured for their proximity to Euclidean distances, while categorical data is simpler, namely the mode value [10]. The calculation of the distance between examples of minor classes whose variables are categorical scale is done using the Value Difference Metric (VDM) formula [12] as in (1).

$$\Delta\,(X,Y) = w_x w_y \sum_{i=1}^{N} \delta(x_i, y_i)^r \qquad (1)$$

$\Delta(X,Y)$ is the distance between X and Y, $w_x w_y$ are the weight (negligible), N is: the number of explanatory variables, R is 1 (Manhattan distance) or 2 (Euclidean distance) and $\delta(x_i, y_i)^r$ is the distance between categories, with (2).

$$\delta(V_1, V_2) = \qquad (2)$$

$\delta(V_1, V_2)$ is the distance between the values of V_1 and V_2, while C_{1i} is the number of V_1 yang which belong to i, C_{2i} is the number of V_2 which belongs to class i, I is the number of classes, i=1,2, …. M, C_1 is the number of values 1 occurs. C_2 is the number of values 2 occurs, N is the number of categories and R is a constant (usually 1).

- Procedure of artificial data generation for numeric data
  - Calculate the difference between main vectors and their closest neighbors.
  - Multiply the difference by a random number between 0 and 1.
  - Add this difference to the principal value of the original main vector so that a new principal vector is obtained.
- Procedure of artificial data generation for categorical data
  - Select the majority between the principal vector under consideration and its nearest k-neighbor for nominal values. If there is a similar value, choose randomly.
  - Make the value data as an example of a new artificial class.

B. *Classification Method*

1) Naïve Bayes: Naïve Bayes is a simple probabilistic classification algorithm that calculates a set of probabilities based on the number of frequencies and value combinations from a dataset. This method requires only a small amount of data in the classification process and often gets unexpected results that do not match the reality [13]. In simple terms, the Naïve Bayes grouping assumes the existence of a certain feature in a class is not related to the presence of other features [14]. Bayes' theorem provides a way to calculate the posterior probability (C|X) of P (C), P (X) and P (X|C) using (3).

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \qquad (3)$$

$P(C|X)$ is a posterior probability class (C, target) given a predictor (X, attribute). $P(C)$ is the probability of the previous class. $P(X|C)$ is the probability which is the probability of the predictor of the given class and $P(X)$ is the probability of the previous predictor.

2) K-NN: In short, the KNN is a classification algorithm based on the nearest neighbor to calculate the distance, the Euclidean Distance equation can be used. Euclidean Distance is a formula for finding the distance between 2 points in two-dimensional space, equation 4 shows the calculation of Euclidean Distance.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (4)$$

3) J48: Decision tree with the J48 algorithm is a classification method that uses a tree structure representation where each node represents an attribute, the branches represent the value of the attribute, and the leaves represent the class [7].

The node at the top of the decision tree is known as the root. Where the following steps are taken to build a decision tree;

- Forming a decission system consisting of condition attributes and decision attributes. Shows an example of a decision system in this study. It only consists of n objects, E1, E2, E3, E4, ..., En and attribute conditions, namely sales, purchases, warehouse stock, and operating expenses. Meanwhile, profit is a decision attribute.
- Calculate the amount of column data, the amount of data based on the results attribute members with certain conditions. For the first process, the conditions are still empty.
- Select attributes as Node 4. Create a branch for each member of the Node.
- Check whether the entropy value of any Node member is zero. If present, determine which leaves were formed. If all the entropy values.
- If there is a Node member that has an entropy value greater than zero, repeat the process from the beginning with Node as a condition until all members of the Node are zero.

Node becomes the attribute with the highest gain value from the existing attributes. Eq. (5) is used to calculate the gain value of an attribute.

$$\mathrm{Gain\ (S, A) = Entropy\ (S)} - \sum_{\mathrm{i=1}}^{\mathrm{n}} \left|\frac{\mathrm{si}}{\mathrm{s}}\right| \times \mathrm{Entropy\ (Si)} \quad (4)$$

S	: Case Collections
A	: Attribute
N	: The number of partitions attribute A
|Si|	: The proportion of Si to S
|S|	: Number of cases in S.

Meanwhile, to calculate the Entropy value with (6).

$$\mathrm{Entropy(S)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} -\mathrm{pi} \times \log_2 \mathrm{pi} \qquad (5)$$

S	: Case Collections
N	: Number of Partitions S
Pi	: The proportion of Si to S

4) Random Forest: The random forest is an ensemble method of learning used for classification, regression, and other tasks. Random forest performance was adapted from a decission tree, with each tree being developed from a bootstrap sample based on training data. When developing the tree, a subset of attributes is drawn randomly from the best attributes to be selected split. The final model of the random forest is based on the results of the entire subset tree that has been developed. Tree is a simple algorithm that divides data from node to node based on class division. Trees are found earlier than random forest.

5) SVM: Support Vector Machine (SVM) is a machine learning technique that separates attribute space from hyperplane, thereby maximizing the margin between instances of a class and class values [16]. SVM can work well on high dimensional data. But SVM training times tend to be slow, even though SVM is very accurate for handling complex nonlinear models. Weakness SVM is prone to overfitting when compared to other methods [17]. The maximum margin separation hyper plane was found by solving the optimization problem of quadratic programming (QP).

The biggest advantage of SVM comes when data is separated nonlinearly. In this case, SVM makes data linearly separable with the help of kernel functions. The kernel function is the mapping of data input patterns to several high dimensional spaces so that the data points are linearly separated. In practice, it does not define the mapping of data points implicitly, but it is explicitly defined as the inner product between data points according to being separated in high dimension space [7].

6) AdaBoost: Adaptive boosting (AdaBoost) is one of several variants of the boosting algorithm [18]. AdaBoost is an ensemble learning that is often used in the AdaBoost algorithm. AdaBoost and its variants have been successfully applied to several fields due to their solid theoretical basis, accurate predictions, and great simplicity with the following steps:

- Input: A collection of research sample sets with lable {(xi, yi), …, (Xn, Xn)}, a component learn algorithm with a number of turns T.
- Initialization: Weight of a training sample $W_i^{1=}1/N$, for all *i=1, ..., N*
- Do for *t=1, ..., T*
  - Use the component learn algorithm to train an h_t classification component, on the training weight sample.
  - Calculate the training error with $h_t\colon \varepsilon_t = \sum_{i=1}^{w} w_i^t, y_i \neq h_t(x_i)$
  - Set a weight for the component classifier $h_t =$ $= \ a_t = \frac{1}{2} In\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
  - Update training sample weights $w_i^{t+1} = \frac{w_i^t \exp\{-a_t y_t h_t(x_i)\}}{c_t}, i = 1, \dots, N$
    C_t is a normalization constant
- Output $f(x) = sign(\sum_{t=1}^{T} a_t\, h_t(x))$

C. Evaluation

The evaluation model used is the use of 10-fold cross validation and confusion matrix by comparing the results of the classification carried out by the system with the actual classification results. Accuracy measurements with confusion matrix can be seen in Table II.

1) Accuracy: Accuracy as in (7) is the level of closeness between the predicted value and the actual value.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

2) TPR (True Positive Rate): True Positive Rate as in (8) is the value of true positives (correctly classified data).

$$TPR = \frac{TP}{TP+FN} \qquad (7)$$

3) Recall: Recall as in (9) is the success rate of the system in recovering information.

$$Recall = \frac{TP}{TP+FN} \qquad (8)$$

4) Precision: Precision as in (10) is the level of accuracy between the information requested by the user and the answer given by the system.

$$Precision = \frac{TP}{TP+FP} \qquad (9)$$

5) F-Measure: The F-measure as in (11) is to combine recall and precision scores into one measure of performance.

$$F - Measure = \frac{2*recall*precision}{(recall+precision)} \qquad (10)$$

III. RESULTS AND DISCUSSION

After extracting the direct marketing bank dataset, to build a classification algorithm using WEKA. In each attribute, no missing value was detected, but it is clear that there is some information that is not known (unknown). If the unknown data is included in the missing value category to deal with existing missing values, treatment can be done to predict it with statistical methods such as mode, association, or other methods. However, in this dataset, most unknown values occur in attributes of nominal type, so the use of the classification method is considered more appropriate. Then SMOTE was carried out to eliminate the imbalance of target attributes in the dataset. The previous data amounted to 45,211 instance data with attribute Y which has a value of yes 5,289 and no. 39,922 after SMOTE is done to yes 39,922 and no 26,455. The comparison is shown in Fig. 2.

TABLE II
CONFUSION MATRIX

		Class Prediction	
		Yes	**No**
Class Actual	Yes	TP	FN
	No	FP	TN

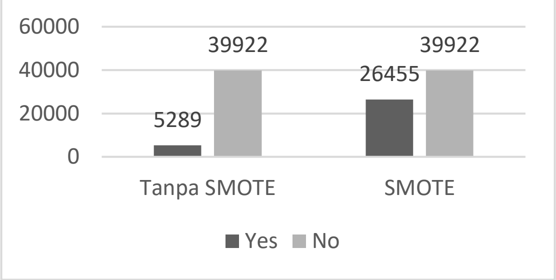

Fig. 2 Comparison of the number of attributes Y

Then do the test with test data using the concept of 10-fold cross validation and confusion matrix by comparing the results of the classification carried out by the system with the actual classification results. The data that has been pre-processed, the classification process is carried out against the target from the Bank Marketing dataset using the existing model in the WEKA software then comparisons between various classification methods and the effect of pre-processing on the classification.

The first experiment for classification without using SMOTE pre-processing using the Naïve Bayes method obtained the results of scoring an accuracy value of 88%, TPR 88%, Recall 88%, Precision 88.40% and F-Measure 88.20%. The KNN method continued with an accuracy value of 86.9%, TPR 87%, Recall 87%, Precision 86% and F-Measure 86.40%. The Random Forest method produces an accuracy value of 90.38%, TPR 90.40%, Recall 90.40%, Precision 89.30% and F-Measure 89.60%. Then the J48 method in this study obtained an accuracy value of 90.30%, TPR 90.30%, Recall 90.30%, Precision 89.50% and F-Measure 89.80%. The PART method obtained the results of scoring an accuracy value of 89.10%, TPR 89.10%, Recall 89.10%, Precision 88.70% and F-Measure 88.90%. Then continued the SVM method in this study to get an accuracy value of 89.20%, TPR 89.30%, Recall 89.30%, Precision 87.20% and F-Measure 86.60% and finally in the classification test without pre-processing. Using the AdaBoost method, the accuracy value was 89.36%, TPR 89.40%, Recall 89.40%, Precision 87.30% and F-Measure 87.30%. From all tests without pre-processing, it is found that the Random Forest method has the highest scoring accuracy. Comparison of the scoring of each classification method without using pre-processing can be seen in Table III.

The next experiment was to conduct Bank Marketing dataset training by conducting the SMOTE process before classification. So that we get a comparison of the target number of data instances as shown in Figure 2.Using the Naïve Bayes method, the results of the scoring are 82.17%, TPR 82.20%, Recall 82.20%, Precision 82.60% and F-Measure 82. , 30%. Followed by the KNN method resulted in an accuracy value of 86.76%, TPR 86.8%, Recall 86.8%, Precision 86.8% and F-Measure 86.8%. The Random Forest method produces an accuracy value of 92.61%, TPR 92.60%, Recall 92.60%, Precision 92.70% and F-Measure 92.60%. Then the J48 method in this study obtained an accuracy value of 90.52%, TPR 90.50%, Recall 90.50%, Precision 90.60% and F-Measure 80.30%. Then proceed with the SVM method in this study to get an accuracy value of 86.73%, TPR 86.70%, Recall 86.70%, Precision 86.70% and F-Measure 86.70% and finally in testing the AdaBoost method the accuracy value is 92.35%, TPR 92.40%, Recall 92.40%, Precision 92.40% and F-Measure 92.40%. From the whole test, it is obtained that the Random Forest method gets the highest scoring in all evaluations both accuracy, TPR, Recall, Precision, F-Measure. Comparison of the scoring of each combination of the SMOTE pre-processing method and the classification method can be seen in Table IV.

In the Bank Marketing dataset the use of SMOTE and tree-based classification methods can increase the scoring which is quite good, but in the SVM and Naïve Bayes methods there is a decrease in the scoring value. The comparison of the results of the accuracy of each method is shown in Fig. 3.

The SMOTE + Random Forest method gets an accuracy value of 92.61% with confusion matrix as in Table V and the ROC (Receiver Operating Characteristic) curve as in Fig. 4.

IV. CONCLUSION

In this study, the use of the SMOTE method with Random Forest obtained fairly reliable results applied to the Bank Marketing dataset. The problem of imbalance data instances which has a significant difference in the Bank Marketing dataset can be solved effectively using the SMOTE method and can increase the accuracy value which is quite significant compared to the classification without the SMOTE method first, but in this study the increase in the accuracy value is more in the classification method, tree-based, whereas for the K-NN, Naïve Bayes and SVM methods in this study the use of SMOTE actually reduces the accuracy value. However, in this study, the longest computation time to run the model is SVM, and Random Forest, because the use of SMOTE increases the number of instance data which has an impact on increasing computation time, so in further research it can add attribute selection methods to reduce computation time and increase the accuracy value. at the time of data training.

TABLE III
TEST RESULTS WITHOUT SMOTE

	Accuracy	TPR	Recall	Precision	FMeasure
Naïve Bayes	88,00%	88,00%	88,00%	88,40%	88,20%
KNN	86,90%	87,00%	87,00%	86,00%	86,40%
R Forest	90,38%	90,40%	90,40%	89,30%	89,60%
J48	90,30%	90,30%	90,30%	89,50%	89,80%
SVM	89,20%	89,30%	89,30%	87,20%	86,60%
Adaboost	89,36%	89,40%	89,40%	87,30%	87,30%

TABLE IV
TEST RESULTS WITH SMOTE

	Accuracy	TPR	Recall	Precision	FMeasure
Naïve Bayes	82,17%	82,20%	82,20%	82,60%	82,30%
KNN	86,76%	86,80%	86,80%	86,80%	86,80%
RForest	92,61%	92,60%	92,60%	92,70%	92,60%
J48	90,52%	90,50%	90,50%	90,60%	80,30%
SVM	86,73%	86,70%	86,70%	86,70%	86,70%
Adaboost	92,35%	92,40%	92,40%	92,40%	92,40%

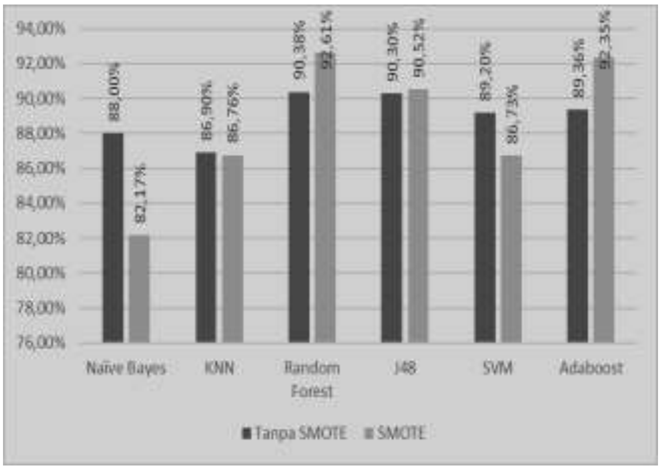

Fig. 2 Comparison of Accuracy Values

TABLE V
CONFUSION MATRIX FOR SMOTE + RANDOM FOREST

Yes	No	Total
37197	2725	39.922
2177	**24268**	26.445

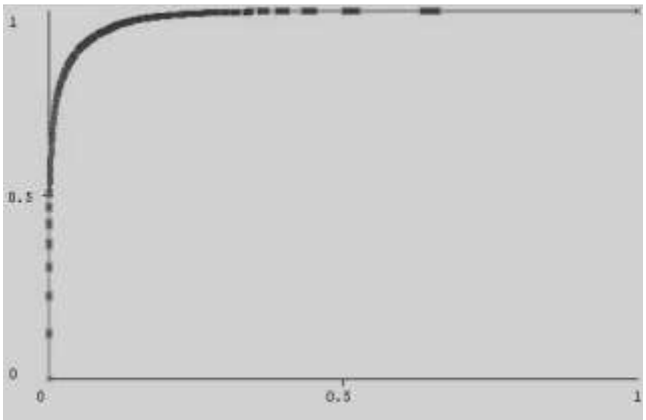

Fig. 3 ROC curve for SMOTE + random forest

REFERENCES

[1] S. Abbas, “Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset,” *Int. J. Comput. Appl.*, 2015.

[2] T. Parlar and S. K. Acaravci, “Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data,” *Int. J. Econ. Financ. Issues*, vol. 7, no. 2, pp. 692–696, 2017.

[3] S. Moro, R. M. S. Laureano, and P. Cortez, “Using data mining for bank direct marketing: An application of the CRISP-DM methodology,” *ESM 2011 - 2011 Eur. Simul. Model. Conf. Model. Simul. 2011*, no. Figure 1, pp. 117–121, 2011.

[4] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” *Decis. Support Syst.*, vol. 62, pp. 22–31, 2014.

[5] Maghfirah, T. B. Adji, and N. A. Setiawan, “Menggunakan Data Mining Untuk Segmentasi Customer Pada Bank Untuk Meningkatkan Customer Relationship Management (CRM) Dengan Metode Klasifikasi (Agoritma J-48, Zero-R Dan Naive Bayes),” *SNST ke-6*, pp. 65–70, 2015.

[6] R. Vaidehi, “Predictive Modeling to Improve Success Rate of Bank Direct Marketing Campaign,” *Int. J. Manag. Bus. Stud.*, vol. 6, no. 1, pp. 22–24, 2016.

[7] A. N. Rais, “Integrasi SMOTE Dan Ensemble AdaBoost Untuk Mengatasi Imbalance Class Pada Data Bank Direct Marketing,” *J. Inform.*, vol. 6, no. 2, pp. 278–285, 2019.

[8] A. Lawi, A. A. Velayaty, and Z. Zainuddin, “On identifying potential direct marketing consumers using adaptive boosted support vector machine,” *Proc. 2017 4th Int. Conf. Comput. Appl. Inf. Process. Technol. CAIPT 2017*, vol. 2018-Janua, pp. 1–4, 2018.

[9] J. Zhang, “Analysis of Neural Network on Bank Marketing Data Dataset Pre-processing,” *Coll. Comput. Sci.*, 2018.

[10] W. P. K. N. V. C. K. W. B. Lawrence O. Hall, “SMOTE: Synthetic Minority Over-sampling Technique Nitesh,” *J. Artif. Intell. Res.*, no. Sept. 28, pp. 321–357, 2002.

[11] J. Sun, H. Li, H. Fujita, B. Fu, and W. Ai, “Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting,” *Inf. Fusion*, vol. 54, no. December 2018, pp. 128–144, 2020.

[12] F. M. A. Rossi Azmatul Barro, Itasia Dina Sulvianti, “Penerapan Synthetic Minority Oversampling Technique (Smote) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu,” *Xplore J. Stat.*, vol. 1, no. 1, pp. 1–6, 2013.

[13] S. Muthuselvan, S. Rajapraksh, K. Somasundaram, and K. Karthik, “Classification of Liver Patient Dataset Using Machine Learning Algorithms,” *Int. J. Eng. Technol.*, vol. 7, no. 3.34, p. 323, 2018.

[14] C. Anam and H. B. Santoso, “Perbandingan Kinerja Algoritma C4 . 5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa,” *J. Ilm. Ilmu-Ilmu Tek.*, vol. 8, no. 1, pp. 13–19, 2018.

[15] G. Rahangdale, M. Ahirwar, and M. Motwani, “Application of k-NN and Naïve Bayes Algorithm in Banking and Insurance Domain,” *Int. J. Comput. Sci. Issues*, vol. 13, no. 5, pp. 69–75, 2016.

[16] I. Oktanisa *et al.*, “Perbandingan Teknik Klasifikasi

Dalam Data Mining Untuk Bank a Comparison of Classification Techniques in Data Mining for," vol. 5, no. 5, pp. 567–576, 2018.

[17] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Comparative study on data mining classification methods for cervical cancer prediction using pap smear results," *Proc. 2016 1st Int. Conf. Biomed. Eng. Empower. Biomed. Technol. Better Futur. IBIOMED 2016*, pp. 1–5, 2017.

[18] H. Liu, H. Q. Tian, Y. F. Li, and L. Zhang, "Comparison of four Adaboost algorithm based artificial neural networks in wind speed predictions," *Energy Convers. Manag.*, vol. 92, pp. 67–81, 2015.