The 4ᵗʰ International Seminar on Science and Technology
August 9th 2018, Postgraduate Program Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

1

Accounts Receivable Seamless Prediction for Companies by Using Multiclass Data Mining Model

Ferry Irawan[1], Febriliyan Samopa[2]

Abstract—Most companies find themselves in highly competitive markets nowadays. As a result, many companies struggle to manage their financial obligation to pay their supplier on time. Delayed payments to suppliers can create all kinds of issue with the supplier's cash flow. Finding a way to reduce or avoid any potential losses because of this delay is needed. Currently, data mining techniques have been widely applied to the assessment or prediction of credit scores for customers in the banking industry (credit scoring), and the most commonly used method is classification. Based on previous studies, research has been conducted to develop a data mining model to produce the best classification model to predict a customer's payment capabilities. With the application of data mining approaches using oversampling, feature selection (FS) algorithm, including Relief, Information Gain Ratio, PCA, and multiclass algorithm, including Random Forest, SVM, One-versus-all, All-versus-all and Error Correcting Output Coding (ECOC), is expected to produce good accuracy to predict the ability of these payments. As a result of this research, the model proposed can provide the best classification model with 84.24% accuracy and AUC value of 95.3% using sample dataset of manufacturing industry within three years period.

Keywords—Credit scoring, Data mining, Payment, Prediction, Receivables.

I. INTRODUCTION

In highly competitive markets nowadays, many companies struggle to manage their financial obligation on time. This condition can create an issue with a supplier's cash flow because of the delay in their payment. Therefore, finding a way to reduce or avoid any potential losses because of this delay is needed. As the company and the number of customers grow, then the current accounts receivable (A/R) control is not sufficient anymore. Because increasingly difficult to analyze and inaccurate to determine customer priorities, especially customers who need special attention.

Various researches of data mining have provided many useful products and applications in various fields, including applications in business, i.e. credit scoring of banking customers in purpose to estimate whether the customer needs to be given credit facilities or not [1], prediction of customer payment behavior that shows whether the customer will pay on time or not [2], etc. There are many approaches in previous research provided in the following: (1) single classifier model, (2) multiple classifier model, (3) ensemble classifier model and (4) hybrid model [3]. Continuing previous studies, the development of data mining model for credit scoring of banking customers by using several classification algorithms, then developed a new model to predict the seamless of customer receivables (smooth, average, and not smooth) with multiclass approach. The proposed model includes a combination of stages, including oversampling, feature selection and three classification approaches covering single, ensemble and multiclass classifications. Each classification algorithm will be compared to get the best algorithm based on ACC and AUC measures. For experimental results, dataset related to customer receivables of a private company in Surabaya in 2014 to 2016 has been recorded.

The contribution of this research in the form of the proposed data mining model is as follows: (1) provides the development of previous credit scoring research, (2) is expected to be a reference for similar researches and adds knowledge in studying in the field of data mining classification, (3) and the best generated classification model can be applied to a presentation tool that provides support to improve the performance of the company's receivable control.

II. METHOD

The basic idea of data mining model research to predict the seamless of customer receivable payment is taken from previous studies of credit scoring on credit card customers. Both relate to the ability to pay their obligations. Thomas defines credit scoring as a set of decision models that help lenders in providing credit to consumers [4]. In practice, credit decisions are often still based on subjective or qualitative.

Huang et al. (2007) conducted research using two credit card datasets from the UCI database and compared several algorithms. Accuracy was obtained by the combination of genetic algorithms and SVM classifiers [1]. Tsai and Wu (2008) used three credit datasets from Australia, Germany, and Japan. This study compared the accuracy between singles and multiple classifiers using neural networks algorithms. In theory, multiple classifiers should produce better accuracy, but the results of this study were the

[1]Ferry Irawan is with Departement of Management of Technology, Institut Teknologi Sepuluh Nopember, Surabaya, 60264, Indonesia. E-mail: ferry.irawan16@mhs.mmt.its.ac.id

[2]Febriliyan Samopa is with Departement of Information System, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. E-mail: iyan@its.ac.id

opposite [5]. Koutanaei et al. (2015) conducted a study using a credit dataset from Iran's Export Development Bank through a hybrid approach, combining FS algorithms and ensemble learning classifiers. The combination of PCA algorithm and ANN-AdaBoost classification had the best accuracy (ACC 91%, AUC 91.2%) [3]. Chen et al. (2013) designed a framework for predicting payment behavior from telecom service subscribers, to forecast whether customers will pay on time or not. The developed framework included association mining, clustering analysis and ensemble of Decision Tree algorithm. The average accuracy was 78.53% [2]. From many previous credit scoring studies, clearly data mining has techniques and methodologies for credit assessment, and it can be applied in a broader context [6].

A. Oversampling

The use of data mining often faces problems of imbalanced data. Imbalanced data refers to a dataset where one or more classes have more data samples than other classes [7]. If this imbalanced data is used as a sample of classification learning, then the resulting classification model will not be able to predict each class optimally. The common technique used to solve this problem is resampling, and oversampling is one of the resampling methods, which produces the better result in classification accuracy than using the undersampling method in previous credit scoring research [8]. This study uses the SMOTE algorithm because of the performance and effectiveness [9].

B. Feature selection algorithms

A data object needs to know that its features are recognizable and differentiated from other data objects. The optimum features that can be known from data objects will facilitate and accelerate the identification process of its data object. A large number of features does not necessarily guarantee the accuracy of the predicted model generated, since not all features have a significant impact in the formation as a classification function, and quite possibly some features even reduce the accuracy result. Therefore, feature selection (FS) being used to select features that have an impact in the making precise classification function. There are several FS algorithms used i.e. Relief, Information Gain Ratio (IGR) and Principal component analysis (PCA).

Relief was first proposed by Kira and Rendell in 1992. The basic idea is to measure the quality of features based on the ability of its values to distinguish classes from objects that are close to one another. The smaller the "near hit" and the larger "near miss", then the more significant the weight of that feature [10]. IGR algorithm is used to handle the deficiencies in the Information Gain (IG) algorithm that is having problems or bias for attributes that have highly variable values (the algorithm tends to prefer features with high variations in certain instances) [11]. This algorithm works based on the value of IG and Split Information. PCA was invented by Karl Pearson in 1901 as a method for reducing data. The algorithm can find the linkage between attributes in an initial dataset by making a linear combination of many attributes into several new attributes, while retaining as much as possible of the variation present in the initial dataset. The result is a new dataset with new attributes, which are uncorrelated [12].

C. Multiclass algorithm

The general approach of multiclass classification can be divided into two types [13]:
1) Binary (two-class) classification algorithms that can be naturally extended to handle multiclass problems directly, i.e. Naive Bayes, CART decision tree, SVM, ANN and KNN [14].
2) Decomposition of multiclass problems into several two-class classification works that can be solved using general two-class classification algorithms, called the hypothetical function or h: X → Y and Y = {1, … , k} [15], then the classification results are combined. Several methods have been proposed for decomposition: one-versus-all (OVA), all-versus-all (AVA), and Error Correcting Output Coding (ECOC).

Naive Bayes, this algorithm applies Bayes theory to calculate the probability and predicts the class with the highest posterior probability. Decision tree is the most commonly used algorithm by researchers in scoring credit scoring. CART is a classification method that uses historical data to form a decision tree that can classify new data, developed by Breiman, Freidman, Olshen, Stone in 1984 in their paper "Classification and Regression Trees" [16]. Support Vector Machine or abbreviated with SVM, introduced and developed by Boser, Guyon, Vapnik in 1992 as a classification technique for linear and nonlinear problems [17]. The concept of SVM is simply to find the hyperplane by maximizing the distance between classes (margin). Artificial Neural Network (ANN) or commonly called Neural Network is a network that modeling the nerves of the human brain (called neurons) in implementing the pattern recognition process. The basis of this modeling is the ability of the human brain to organize neurons in order to recognize patterns effectively and solve problems. K-Nearest Neighbors or abbreviated K-NN is a simple type of algorithm in which an object is classified according to the majority class of its nearest neighbor class.

Many studies are applying ensemble classifier algorithms to credit scoring, such as Random Forest, Bagging, Boosting and Stacking [18]. Random Forest is a collection of un-pruned decision trees through random FS, which have been trained through bootstrapping of training data. This algorithm is based on the voting tree procedure from the most popular class (Brown and Mues, 2012). Bagging is short for Bootstrap Aggregation. The algorithm was developed by Breiman [19] and is based on the concept of majority voting, where the different subset of training data is used randomly for the same base classifier training. Similar to Bagging algorithm, Boosting also runs an ensemble classification, but the sample data has been given the weight from the previous classification training so that the next classification training process can be more

The 4ᵗʰ International Seminar on Science and Technology
August 9ᵗʰ 2018, Postgraduate Program Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

3

accurate. Stacking combines several different learning algorithms to achieve higher prediction accuracy. It combines predictions from some base-level base learners with a meta-level base learner [18].

OVA, this method is also referred to as one-against-all in some literature. The basic idea is to train k classifications between one class with the other classes (k is the number of classes). Known training dataset S then formed k number of two-class training dataset, i.e. $S_1,...,S_k$, where S_i is a collection of instances labeled 1 if the instance is class i and -1 if otherwise. Then run binary predictor training or hypothesis $h_i: X \rightarrow \{\pm 1\}$ based on S_i. With the expectation of $h_i(x)$ is worth 1 if and only x belong to class i. On the basis of $h_i,...,h_k$, multiclass predictions are formed based on the biggest of h values [15]. AVA, also known as one-against-one. The basic idea is to train $k(k-1)/2$ classifications to differentiate each pair of class. Known training dataset S, for each $1 \leq i \leq j \leq k$ a training dataset $S_{i,j}$ is formed containing all samples S labeled i or j. The label +1 is set if the label in the dataset S is i and -1 if the label in the dataset S is j. The next step is to apply the classification algorithm to each training dataset $S_{i,j}$ to obtain the hypothesis $h_{i,j}$ and the prediction result is determined from the biggest of h values. ECOC, this method works by training N binary classifier to differentiate K classes. Each class is assigned a codeword of length N corresponding to the binary matrix M. Each row of M points to a particular class [20]. Evaluating N binary classifier with the test data and its codeword results compared to codeword from K classes (matrix M), and a class with the smallest hamming distance will be selected as the class label for the test data.

III. EXPERIMENTAL DESIGN

A. Dataset description and data pre-processing

Kennedy (2013) used the attribute collection of arrears and payment behavior [21], while Yeh and Lien (2009) used historical data of credit card payments from a bank in Taiwan as research attributes [22]. Based on previous studies, the authors determine the attributes that will be used to predict the seamless of accounts receivables using receivables historical data of a private company from 2014 to 2016. After going through the preparation steps, the final dataset has 27 features in the following structure: credit limit, risk category, payment status of due A/R (the period h until h-5, -1 = no delay, 1 = average delay of payment is one month, 2 = average delay of payment is two months, ... and above), ending balance of A/R (the period h until h-5), the amount of payment in the period (period h until h-5), ending balance of due A/R (the period h until h-5) and label of class.

B. The proposed model

Figure 1 shows a block diagram of the proposed model. It consists of (1) data gathering and pre-processing, (2) oversampling, (3) FS, and (4) modeling. The first stage is data gathering and dataset preparation, and then the second stage is to ensure that the classes on the dataset are in balance by performing oversampling in minority classes using SMOTE algorithm. The objective is to find the sampling parameter that produces a True Positive Rate (TPR) in balance for each class using the Random Forest and SVM classification algorithms.

After obtaining the balanced dataset then proceed to the third stage. This stage is the selection of the FS algorithm that will generate a set of the selected optimal features to simplify and accelerate the process of class identification. Experiments are performed using several FS algorithms: Relief, IGR, and PCA. Each FS algorithm is tested several times using different parameter values, and for each resulted feature set will be the new dataset for classification testing using 10 folds cross validation with Random Forest and SVM algorithms. This experiment will be repeated until getting the FS algorithm with the optimum parameters that produce no multicollinearity dataset and the finest average of ACC, AUC and TPR measures.

After getting the dataset with the best features, the fourth stage is a series of steps to get a classification model that produces the best accuracy. In the first step, some tests are conducted using a two-class algorithm with the combination of parameters corresponding to each algorithm, i.e. Naive Bayes, CART, SVM, ANN, and KNN. The results of the first step are some of the best classification models generated from each algorithm and its best parameters. Furthermore, the second step will be done many tests using ensemble classification algorithms with the combination of parameters on each algorithm, i.e. Random Forest, Bagging, Boosting and Stacking. In the ensemble algorithm generally, there are parameters of classifier (classification algorithm) which roles as an algorithm that will run on multiple iterations and generate predictive values. The classifier parameter in this step will use the combination of algorithms and its parameter that produces the best classification models from the first step. The purpose of the ensemble algorithm using each best two-class algorithm as its parameter is expected to provide more accurate models. In the third step, it will be tested using multiclass classification algorithms with the approach of AVA, OVA, and ECOC. In this algorithm, there are also classifier parameters as in the ensemble classification algorithm. The classifier parameter will use each algorithm and its best parameters from the result of the first and second step. After completing the tests in all three steps, where each step produces the best classification models of each algorithm, the best classification model can be selected based on the average of ACC and AUC measures.
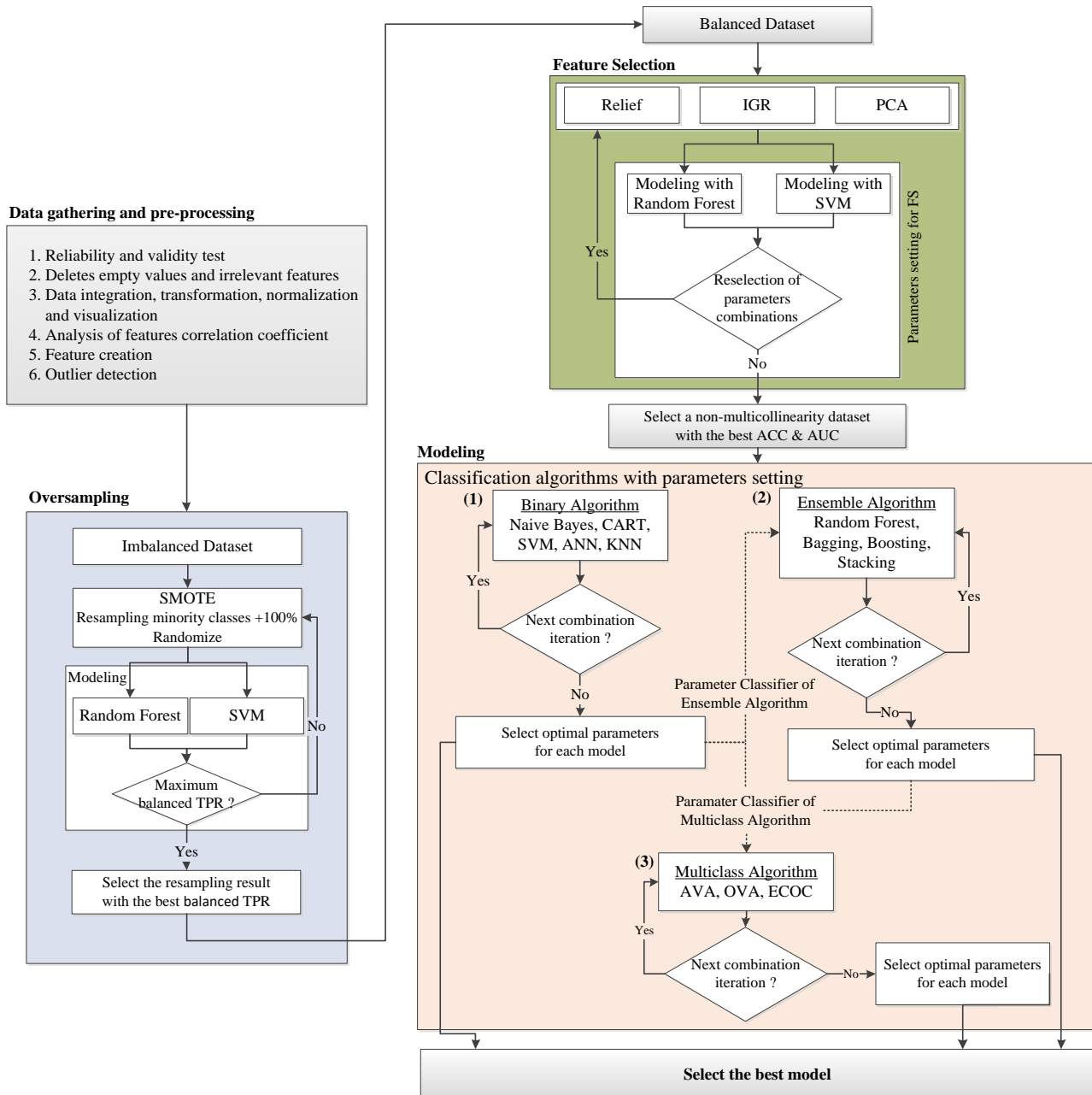
Figure 1. The block diagram of the proposed model.

IV. RESULTS AND DISCUSSION

After data collection conducted and preparation resulted in 10,230 rows of data, the next step would be labeling by using expert judgment method, and this caused data imbalance. The majority class was smooth (74%), the minority classes were average (14%) and not smooth (12%). After some oversampling tests using SMOTE algorithm with percentage parameter from 100% to 600%, there was a significant change in the balance of TPR using the combination of oversampling parameter 500% and

SVM algorithm, where TPR of class A decreased drastically from 95% to 28.5%, and TPR class B raised from 11.9% to 77.9% (see Figure 2). Therefore, this research would use the oversampling parameter of 400% that gave the best-balanced TPR on Random Forest (average TPR 88.8%) and SVM algorithm (average TPR 47.4%). From this process, the resulted dataset contained 20,934 rows of data and the composition of each class was 30% approximately. To shorten the label name, the author uses label A representing the smooth class, B for the average class and C for the not smooth class.
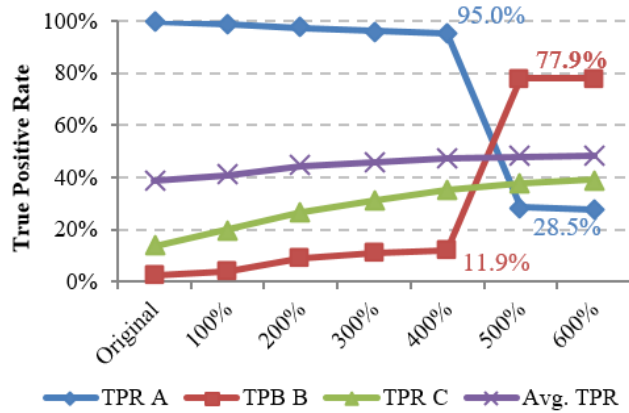
The 4ᵗʰ International Seminar on Science and Technology
August 9ᵗʰ 2018, Postgraduate Program Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

5

Figure 2. The chart of balanced TPR uses 100 - 600% oversampling and SVM algorithm.

Continuing with the selection of the FS algorithm. Relief algorithm used the combination of "number of features" (10, 15, 17 and 20) and "nearest neighbors" parameters (5, 10, 15, 20 and 25). While IGR used the combination of "number of features" parameter (10, 13, 16, 19, 22 and 25), and PCA used the combination of "variance covered" parameter from 0.95 to 1. New generated datasets from each FS algorithm would be tested for accuracy using Random Forest and SVM, and the best results were obtained from testing using the generated dataset of PCA algorithm with "variance covered" parameter 0.99. The result was superior on each assessment criterion, i.e. average correlation 0%, average ACC 79.4%, average AUC 88.3%, and average TPR 79.5% (see Table 1). In addition, the resulting dataset was the only dataset that does not experience multicollinearity.

After getting the dataset with the optimal features, modeling stages were executed. The first step used two-class algorithms that can be extended to handle multiclass problems. Testing using Naive Bayes was done once because it had no parameters, CART was performed using the combination of "minimal number of observations at the terminal nodes" 1 to 6, SVM used the combination of "kernel type" (linear, polynomial, RBF and sigmoid) and "fixed coefficient" parameters (0, 0.5, 1 to 4), ANN used the combination of "learning rate" (0.1, 0.3, 0.6) and epochs parameters (400, 500, 600, 800 and 1000), KNN used the combination of "number of neighbors" parameters (1, 5 and 10). In this step, the KNN algorithm with the parameter "number of neighbors" 5 produced the highest ACC and AUC measures (80.53% and 92.40%), the list of two-class algorithms and the best parameters can be seen at Table 2.

TABLE 1.
COMPARISON OF ACC AND AUC BETWEEN DATASET WITHOUT FS AND USING FS

FS Algorithms	Avg. Correlation	Avg. ACC	Avg. AUC	Avg. TPR
Without FS	42.7%	69.0%	79.0%	68.1%
Relief (15 features, "nearest neighbors" 5)	33.8%	69.4%	79.3%	68.6%
Information Gain Ratio (22 features)	43.2%	69.3%	79.1%	68.4%
PCA ("variance covered" 0,99)	0.0%	79.4%	88.3%	79.5%

TABLE 2
THE RESULTS OF CLASSIFICATION ALGORITHMS CONSIDERING THE ACC AND AUC MEASURES

Approach	Algorithms	Parameters	ACC	AUC
Multiclass	Multiclass	Method Exhaustive correction code. Classifier Stacking, Meta classifier: ANN (learning rate 0.1, epochs 1000), base classifier: KNN (number of neighbors 5) and Random Forest (number of tress 150, maximal depth 0)	84.24%	95.30%
Ensemble	Stacking	Meta classifier: ANN (learning rate 0.1, epochs 1000), base classifier: KNN (number of neighbors 5) and Random Forest (number of tress 150, maximal depth 0)	84.09%	95.30%
	Random Forest	Number of tress 150, maximal depth 0	83.21%	94.90%
	Bagging	Sample rat.1 iter. 20 KNN, number of neighbors 5	80.75%	93.30%
	Boosting	Iterasi 20 CART, minimal number of observations at the terminal nodes 5	80.27%	93.10%
Two-class	KNN	Number of neighbors 5	80.53%	92.40%
	CART	Minimal number of observations at the terminal nodes 5	76.90%	89.60%
	ANN	Learning rate 0.1 epochs 1000	75.99%	90.30%
	SVM	Kernel types Polynomial, fixed coefficient 2	76.60%	82.30%
	Naive Bayes	-	39.88%	64.30%

The second step used the ensemble algorithm with the classifier parameters of each algorithm and the best parameters from the first step. Random Forest testing used the combination of "number of trees" (10, 25, 50, 100, 150 and 200) and "maximal depth" parameters (0 or infinite, 10 and 20). Bagging used the combination of "sample ratio" (0.7 and 1) and iterations parameters (5, 8, 10 and 20). Boosting with the combination of iterations parameters (5, 8, 10 and 20). The last one was Stacking with the combination "meta classifier" and two or three "base

classifier" parameters. Random Forest was specifically added as the classifier parameter of Stacking because the accuracy was good enough. In this step, the best algorithm was obtained using Stacking algorithm with meta classifier parameter: ANN ("learning rate" 0.1, epochs 1000), base classifier: KNN ("number of neighbors" 5) and Random Forest ("number of trees" 150, "maximal depth" 0), where the resulting accuracy was ACC 84.09%, AUC 95.3%.

The third step used the multiclass algorithm with method parameters (OVA, AVA, and ECOC) and classifier parameter of each algorithm and its best parameters from the first and second step. The best accuracy was achieved by the combination of method parameter of ECOC and the classifier parameter of Stacking (meta-classifier: ANN ("learning rate" 0.1, epochs 1000), base classifier: KNN ("number of neighbors" 5) and Random Forest ("number of trees" 150, "maximal depth" 0)), resulted the best multiclass algorithm with ACC 84.24% and AUC 95.30%, as well as the best classification model in this study.

V. CONCLUSION

In this research, a new data mining model is successfully built to produce the prediction model of seamless of accounts receivable with high accuracy, including oversampling, FS and multiclass approach. Using the sample dataset, this research proposes the data mining model that can produce the best classification model using the multiclass algorithm with ACC 84.24% and AUC 95.3%. In the early stage, the use of SMOTE and PCA algorithm proved effective for creating the balanced and non-multicollinearity dataset with principal features. As a result, the classification model can classify each class better, and this will impact on the level of accuracy.

The result of this study extends previous credit scoring studies that generally use two-class classification, and simultaneously confirm that the performance of the ensemble classification algorithm is better than single classification.

In future research, the authors suggest doing clustering at the beginning of the process to determine the number of class. Furthermore, detail research and significant methods to predict variable selections are needed, it along with setting algorithms parameters to the maximum result. Furthermore, there are many other algorithms types and parameters can be used further (FS algorithm: SA, PSO, ant colony, F-Score, LDA and classification algorithms: LR, C4.5, ID3, SMO).

REFERENCES

[1] C.-L. Huang, M.-C. Chen, and C.-J. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert Syst. Appl.*, vol. 33, no. 4, pp. 847–856, Nov. 2007.

[2] C.-H. Chen, R.-D. Chiang, T.-F. Wu, and H.-C. Chu, "A combined mining-based framework for predicting telecommunications customer payment behaviors," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6561–6569, Nov. 2013.

[3] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *J. Retail. Consum. Serv.*, vol. 27, pp. 11–23, Nov. 2015.

[4] L. Thomas, J. Crook, and D. Edelman, *Credit scoring and its applications.* SIAM, 2017.

[5] C. Tsai and J. Wu, "Using neural network ensembles for bankruptcy prediction and credit scoring," *Expert Syst. Appl.*, vol. 34, no. 4, pp. 2639–2649, May 2008.

[6] A. Hooman, M. Omidi, G. Marthandan, W. F. W. Yusoff, and S. Karamizadeh, "Statistical and data mining methods in credit scoring.," *J. Dev. Areas*, vol. 50, pp. 371–381, 2016.

[7] L. Yijing, G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, "Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data," *Knowledge-Based Syst.*, vol. 94, pp. 88–104, Feb. 2016.

[8] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, Feb. 2012.

[9] A. C. Liu, "The effect of oversampling and undersampling on classifying imbalanced text datasets," *Univ. Texas Austin*, 2004.

[10] A. Salappa, M. Doumpos, and C. Zopounidis, "Feature selection algorithms in classification problems: an experimental evaluation," *Optim. Methods Softw.*, vol. 22, no. 1, pp. 199–212, Feb. 2007.

[11] J. D. Kelleher, B. Mac Namee, and A. D. Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics*. London, England: The MIT Press, 2015.

[12] I. T. Jolliffe, *Principal Component Analysis*, Second. Springer, 2002.

[13] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, Oct. 2004.

[14] N. Mehra and S. Gupta, "Survey on Multiclass Classification Methods," *Int. J. Comput. Sci. Inf. Technol.*, vol. 4, 2013.

[15] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: from theory to algorithms*. New York, NY, USA: Cambridge University Press, 2014.

[16] R. Timofeev, "Classification and regression trees (cart) theory and applications," *Humboldt Univ. Berlin*, 2004.

[17] J. Han and M. Kamber, *Data mining: concepts and techniques*, 3rd ed. Burlington, MA: Elsevier, 2011.

[18] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 223–230, Jan. 2011.

[19] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[20] M. Aly, "Survey on multiclass classification methods," *Neural Netw.*, vol. 19, 2005.

[21] K. Kennedy, "Credit Scoring Using Machine Learning," 2013.

[22] I.-C. Yeh and C. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2473–2480, Mar. 2009.