Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 1 (2023) 88 – 93 ISSN Media Electronic: 2580-0760

Credit Scoring Model for Farmers using Random Forest

Kharida Aulia Bahri[1], Yeni Herdiyeni[2], Suprehatin Suprehatin[3]
[1, 2]Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University
[3]Department of Agribusiness, Faculty of Economics and Management, IPB University
[1]kharidabahri@apps.ipb.ac.id, [2]yeni.herdiyeni@apps.ipb.ac.id, [3]suprehatin@apps.ipb.ac.id

Abstract

One of the problems faced by farmers in Indonesia is capital. Based on Indonesian Central Statistics Agency survey results, the number of farmers who borrow capital from formal institutions such as banks is still small. This is because the process of applying for loans at banks is lengthy, farmers are considered high-risk and unbankable, and the rating of the agricultural sector is unattractive to banks. This study aims to determine the attributes and design a model of agricultural credit assessment. This study uses secondary data related to bank credit ratings and land productivity from banks in the Telagasari sub-district in 2018–2020 and Cipayung sub-district in 2020. Data were analyzed using random forests. The research process includes four stages: data collection, data pre-processing, model building, and model analysis and evaluation. This study produced five important variables that are relevant to farmers: planting costs, sales, land productivity, total production, and land area. The model built produces the most optimal accuracy of 83% with an AUC score of 81%. Based on the AUC performance classification, it can be concluded that the model that has been made is good at predicting the credit status of farmers because the AUC value is included in the good classification predicate.

Keywords: agriculture; credit scoring; farmer; land productivity; random forest

1. Introduction

The agricultural sector is one of the strategic sectors in Indonesia's economic development. One of Indonesia's agricultural problems is that there are obstacles for farmers in developing their farms such as access to capital, including small farmers who find it difficult to get capital.

Capital loans, also known as credit, are one of the main supporting factors for the development of the adoption of farm-business technology [1]. Based on the results of a survey of the cost structure of horticultural crops conducted by the Indonesian Central Statistics Agency in 2018, there were only 3.7% of horticultural farmers who borrowed capital, and only about 1.8% of farmers who borrowed from banking institutions [2].

The lack of farmers using banking services can be caused by internal and external factors. Internal factors affecting the low number of farmers borrowing capital to banks, such as the absence of collateral by farmers and farmers' lack of knowledge regarding lending procedures [2]. On the other hand, external factors that cause fewer farmers to get loans from banks are the long process of applying for loans to banks, farmers are considered high-risk and unbankable, and the assessment of agricultural sector ratings is not attractive by banks [3]. Furthermore, in some cases, creditors from banks do not have special expertise in agronomy, so the assessment of farmers' credit risk is quite complicated, and the possibility of credit results obtained from smallholders is also small [4].

Every creditor, especially banks, will evaluate potential debtors based on their ability to pay off loans. This ability will be assessed from various aspects, namely character, capacity, capital, collateral and condition of economic (5C) of the debtor customer [5]. In another sense, prospective debtors will be assessed on the basis of their background, their ability to run and develop their businesses, their financial statements, their collateral, and the current economic conditions. These five aspects will determine whether or not debtors are eligible to be given loans, including farmers. However, not all farmers are able to meet all these aspects, especially small farmers. Therefore, a way to increase lending in the agricultural sector is needed that is right on target and sustainable. This means that in the credit assessment for farmers, it can be seen from the aspect of data relevant to farmers, which can provide an overview of farmers and the socio-economic conditions of farmers.

Accepted: 11-10-2022 | Received in revised: 17-01-2023 | Published: 03-02-2023

According to [4], farmer transaction data related to buying and selling activities, agronomic data, farmer demographic data, satellite imagery data to see land productivity, credit history, and other data showing stability and a track record related to creditworthiness can help answer creditors' questions about farmers' ability to repay loans. One of the important data points that is influential in determining the creditworthiness of farmers and is the focus of this study is land productivity data. Land productivity is a measure of the land's ability to perform certain functions. Land productivity will depend on several variables, such as land area and production output [6]. On the other hand, according to [7], socio-economic variables such as gender, age, education, household size, farming experience, farmer income, membership in farmer organizations, and land area have a strong influence on agricultural land productivity. Meanwhile, according to [8], land labor, the use of pesticides and fertilizers, and household size have an influence on land productivity.

Based on this, several variables were obtained to describe land productivity. Agricultural land productivity is viewed based on the type of farming, agricultural production data, farmer sales data, farmer experience, land area, planting land conditions, planting period, and planting costs. In general, the productivity of this land has been used by banks to assess the creditworthiness of farmers who are business-to-business (B2B) or business-to-customer (B2C). However, these data are only analyzed manually to determine the Repayment Capacity (RPC) or the ability to pay back from farmers, as implemented by banks in Telagasari district, Karawang regency and Cipayung district, Bogor Regency. So, this study focuses on using data on the type of farming business, agricultural production data, farmer sales data, land area, planting period, land ownership, and planting costs in assessing creditworthiness, otherwise known as credit scoring for farmers.

Credit scoring is a statistical method for estimating the probability of a borrower's default using historical and statistical data to achieve a single indicator that can distinguish good borrowers from bad borrowers. By studying the historical customer data and its credit risk, banks can find out the factors that affect the smooth running of customer credit, which will later determine whether the customer is worthy of credit or not [9]. In credit scoring, machine learning can be used to create models. Credit scoring with machine learning can help financial institutions discover important features that affect an applicant's credit status and then assess applicants based on those important features, thereby reducing the rate of bad debt [10].

Research related to *credit scoring* has been widely carried out in various different credit contexts for both agricultural and non-agricultural credit. *Credit scoring* research in the agricultural sector was conducted by [11], who in this study conducted research related to the risks of controlling agricultural MSME credit using the *logistic regression* method. Further related research was conducted by [12]. In this study, modeling online credit risk assessment in agriculture using the Syncretic Cost-sensitive Random Forest (SCSRF) method. This research produced a model that can be used as a reference for decision-making for online loan platforms.

Research related to *credit scoring* in the non-agricultural sector such as technology company credit [13], banking ordinary customer credit [10], [14]–[16], corporate credit in ten different sectors (industry, consumer services, technology, utilities, telecommunications services, health, finance, energy, basic materials, consumer goods) [17], and retail credit [18].

In addition to various contexts, research related to credit scoring uses various methods. Commonly used methods are logistic regression [11], fuzzy logistic regression [13], classification and regression tree [14], ensemble classification based on supervised clustering [16], random forest by developing feature selection algorithms [15], and deep belief networks with restricted Boltzmann machines [17]. The application of random forest methods for credit scoring studies also varies, such as building a credit scoring model based on feature selection and grid search to optimize random forest algorithms [10] and using online credit data in agriculture with syncretic cost-sensitive random forest (SCSRF) [12].

Therefore, this study aims to create credit scoring models for farmers using farmer data. Credit scoring modeling is done using the random forest method. This model is expected to assist stakeholders in assessing creditworthiness for farmers by utilizing data relevant to farmers

2. Research Methods

This research process includes four stages, namely data collection, data preprocessing, model making, and model analysis and evaluation (Figure 1).

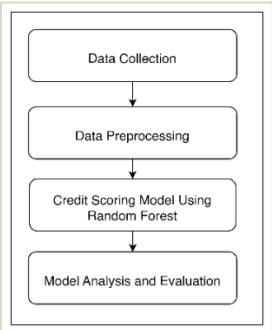

Figure 1. Research Methods

DOI: https://doi.org/10.29207/resti.v7i1.4583

2.1 Data Collection

The first stage of the research is data collection. The data used in this research are secondary data related to banking credit assessment and land productivity from banks in Telagasari district, Karawang regency in 2018–2020, and Cipayung district, Bogor regency in 2020. The attributes or variables used in this study are shown in Table 1.

Table 1. Description of Attribute Data (Variables)

No	Nama Atribut	Keterangan
1	Gender	Gender of the farmer (male, female)
2	Age	Farmer's age (years)
3	Marital status	Marital status of peasants (married, widower, widow)
4	Credit status	Credit status (good, bad)
5	Education	Farmer's last education (Elementary School, Junior High School, Senior High School, Bachelor's Degree)
6	Down payment	The amount of fees paid at the beginning of the credit process (percent)
7	Period	Approved credit term (month)
8	Total production	Number of agricultural production per season (tons)
9	Agricultural Productivity	Productivity of agricultural products (tons per ha)
10	Sales	The value of sales of agricultural products per season (Rupiah)
11	Planting costs	Planting cost per season (Rupiah)
12	Land area	Area of agricultural land (ha)
13	Land ownership	Land ownership status (owned, leased)
14	Home Ownership	Homeownership status (owned, rented)
15	Plafond	Maximum limit of loans provided by banks (Rupiah)
16	Harvest month	Harvest time (months)
17	Location	Address of the farmer (district)
18	Types of agricultural businesses	Types of agricultural businesses (rice, vegetables, ornamental plants, grains, flowers)
19	Number of dependents	Number of dependents farmers (people)

2.2 Data Preprocessing

The data used in this study were explored by understanding their shape and characteristics. At this stage, a descriptive analysis is also carried out to determine what parameters are used for credit scoring.

2.3 Credit scoring model using random forest

In this study, random forests were used to form a credit scoring model. The processed data will be partitioned into test data and training data. The resulting model will generate a scorecard, which contains a score for each variable. This score will be used to determine whether the customer is given credit or not.

2.4 Model Analysis and Evaluation

At this stage, the analysis and evaluation of the model are carried out to see and measure the extent to which the resulting model is able to predict the actual conditions. Evaluation of the model is performed using a confusion matrix. In the confusion matrix, the evaluation of the model is seen from several parameters such as accuracy, precision, recall, and f1-score. A model is considered good if its precision and recall values are higher. The model evaluation process is also carried out using the Area Under Receiver Operating Characteristic (AUROC) curve. According to [19], AUC performance can be classified into five groups, which can be seen in Table 2.

Table 2. AUC classification

Value Range	Classification
0.90 – 1.00	*Excellent Classification*
0.80 – 0.90	*Good Classification*
0.70 – 0.80	*Fair Classification*
0.60 – 0.70	*Poor Classification*
0.50 – 0.60	*Failure Classification*

3. Results and Discussions

Based on the research flow chart (Figure 1), the systematics of the results and discussion of this research are arranged through the following stages:

3.1 Data Preprocessing

Data preprocessing is the process of preparing data that will be processed into cleaner data so that it is ready for use at the next stage. This study used farmer credit data from banks in Telagasari district, Karawang regency in 2018–2020, and Cipayung district, Bogor regency in 2020, as many as 316 data points with **"good"** and **"bad"** credit status. An overview of the credit status of farmers can be seen in Figure 2.

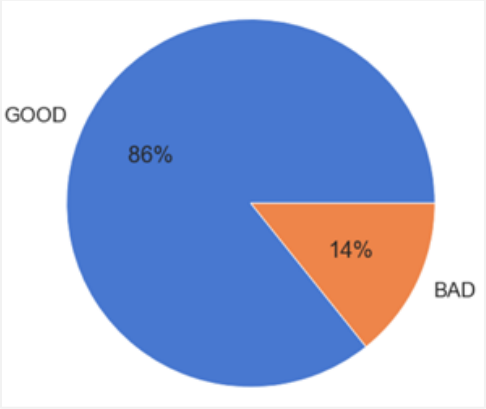

Figure 2. Farmer Credit Status

Based on Figure 2, it can be seen that the credit status of **"good"** farmers has a greater percentage compared to the credit status of **"bad"** farmers. In addition, farmers who have a **"good"** credit status are those who have a land area of more than 5 hectares with a credit period of under 10 months (Figure 3). This condition shows that farmers' loans to banks in Telagasari district, Karawang Regency and Cipayung district, Bogor Regency, are influenced by the large area of farmer land and the short period of credit taken.

DOI: https://doi.org/10.29207/resti.v7i1.4583

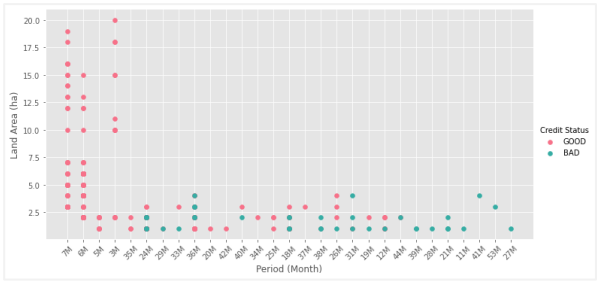

Figure 3. Farmer Credit Status Based on Land Area and Credit Term

Farmer credit data also has other common characteristics, as shown in Figure 4. Data on farmers' loans to banks in Telagasari district and Cipayung district is dominated by farmers with rice farming business types. This is in line with the location of farmers, namely in Tegalasari district, Karawang regency and Cipayung district, Bogor regency, where the majority of the area has a lot of rice fields. In the farmer credit data, the majority of the harvest months are in December and May, as shown in Figures 5.

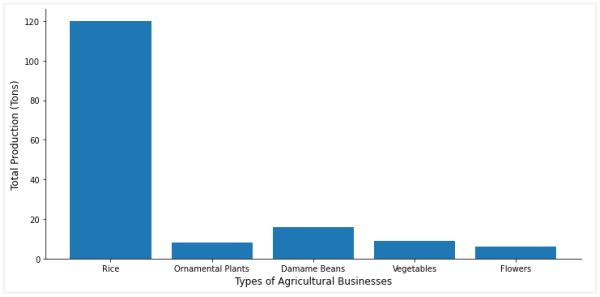

Figure 4. Total Production by Types of Agricultural Businesses

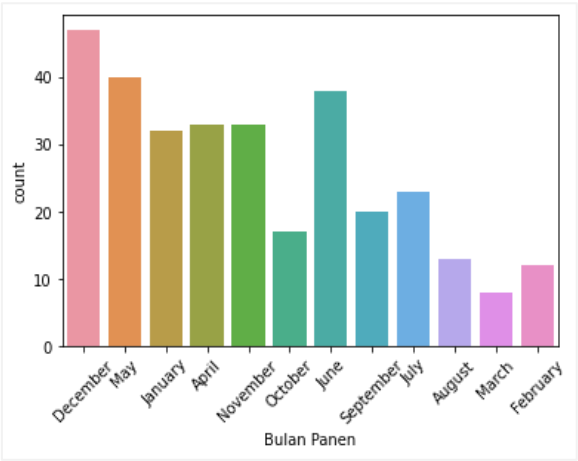

Figure 5. Farmer Credit Status Based on Land Area and Credit Term

Furthermore, the correlation between variables is examined to find out the extent of the influence of each variable on other variables. The results show that the sales data variable has a fairly strong correlation with the total production variable that can be seen in Figure 6.

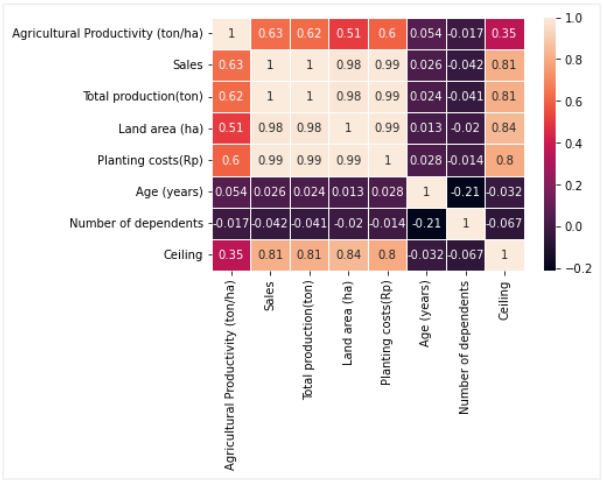

Figure 6. Correlation Between Variables

Based on Figure 6, it can be seen that the sales data variable has a fairly strong correlation with the total production variable. In this study, predictor variables and target variables were determined. The predictor variable consists of all variables in the data except the credit status variable. Variables – variables used for variables as predictor variables of type numeric and categorical data. Categorical variables are used as dummy variables with values of 0 and 1. The credit status variable is used as a target variable, consisting of 86% for **"good"** credit status and 14% for **"bad"** credit status. **"Good"** credit status data is greater than **"bad"** credit status data. This causes the data included in the data class to be unbalanced.

Unbalanced class data has several problems, such as overlapping classes and missing data patterns in minority class data that result in the model being less than optimal in predicting minority class data [20]. To overcome unbalanced class data in this study, the data balancing process was not carried out because the random forest method was able to process unbalanced data. Random forests perform data processing with an ensemble approach that uses a number of classifiers to work together to identify class labels for unlabeled instances. This approach has proven its high accuracy and superiority on unbalanced class data [21].

3.2 Credit Scoring Model using Random Forest

At this stage, the data is first partitioned into training data (80%) and test data (20%). The next process is credit scoring modeling with random forests using tuned training parameters to obtain the optimal model, which can be seen in Table 3.

Table 3. Tuning Parameters

Parameter	Values
n_estimators	100; 110; 120; 150
max_features	Sqrt; 0,1; 0,2; 0.3
min_sample_leaf	1; 3; 5; 7
max_depth	30, 40, 50

DOI: https://doi.org/10.29207/resti.v7i1.4583

In this study, the tuning parameters of the random forest were carried out with a grid search approach. Grid search works by considering all combinations of parameters to obtain optimal parameter values [22]. To calculate its accuracy, this study used cross-validation as many as five times. The results of the training process are shown in Table 4.

Table 4. Training Data Results

Best paramater	Best score
{'RndF__max_depth': 30, 'RndF__max_features': 0.3, 'RndF__min_samples_leaf': 3, 'RndF__n_estimators': 100}	0.83

Based on Table 4, it can be concluded that the most optimal random forest model has an accuracy of 0.83 or 83%. Modeling with random forests produces variable importance that show what variables are influential in determining the credit status of farmers.

Figure 7 shows the variable importance produced in this study. In this study, nine factors were identified that had an influence on producing accuracy in the random forest model, consisting of credit period, planting costs, sales, age, land productivity, total production, the maximum limit of loans provided by banks (plafond), the number of dependents, and land area. From the nine variables of importance, it can be concluded that some variables that have an influence on producing good accuracy in the model are variables of land productivity. Thus, land productivity is influential in determining the credit status of farmers.

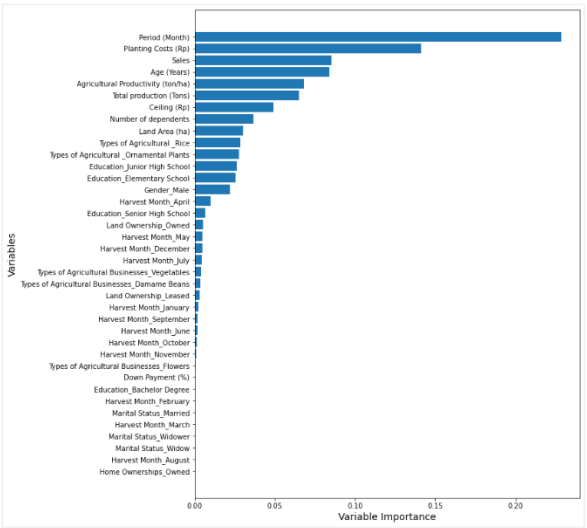

Figure 7. Variable Importance

3.3 Model Analysis and Evaluation

The model that has been used in the training process is evaluated. The evaluation process is carried out using test data. The results of the evaluation process are interpreted using the confusion matrix shown in Figure 8.

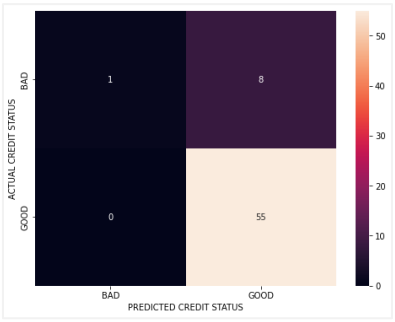

Figure 8. Confusion Matrix

Based on Figure 8, it can be seen that, out of 9 credit data points of farmers with **"bad"** credit status, 1 credit point was successfully classified correctly. While the credit data with a **"good"** credit status from 55 data, all data were successfully classified. To see the performance of the evaluation with the confusion matrix, see Table 5.

Table 5. Confusion Matrix Performance Results

Class	Precision	Recall	F1-Score	Support (Number of test data)
Bad	1,00	0,11	0,20	9
Good	0,87	1,00	0,93	55
Accuracy			0,88	64
Macro accuracy	0,94	0,56	0,57	64
Weighted avg	0,89	0,88	0,83	109

Based on Table 5, it can be concluded that the performance of the model after the evaluation process is quite good. In precision, recall, and f1-score, the result for the class is less than 0.90 or 90% with an accuracy of 0.88, or 88% which means that 100% of the farmer data evaluated the results as the same as the actual data for the good class, but for the bad class, it still cannot classify well. The model evaluation process is also carried out using the receiver operating characteristic (ROC) curve. ROC is used to determine the performance of the model as seen from the relationship between the sensitivity value (true positive) and 1-specificity (false positive). The results of the ROC can be seen in the area under the curve (AUC). AUC is the area under the ROC curve that shows the performance of the model. To see the ROC curve can be seen in Figure 9.

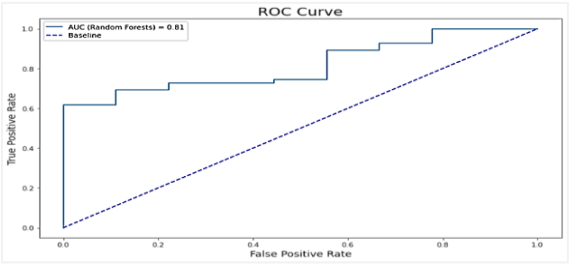

Figure 9. Receiver Operating Characteristic (ROC) Curve

Based on Figure 9, it can be seen that the ROC curve with a straight blue line, which results in an empty area

DOI: https://doi.org/10.29207/resti.v7i1.4583

above the curve, is getting smaller or closer to the value of 1.0 at the true positive rate (sensitivity). In addition, the area under the ROC Curve, called the AUC, produces a larger area with an AUC score of 0.81 or 81%. According to [19], AUC scores from 0.80 to 0.90 are included in the "good" classification. Then it can be concluded that the model created is already good at predicting the credit status of farmers.

4. Conclusion

Credit-scoring modeling using random forests produces variables of importance. The variables of importance that have an influence on producing accuracy in the random forest model consist of credit period, planting costs, sales, age, land productivity, total production, the maximum limit of loans provided by banks (plafond), the number of dependents, and land area. In terms of variable importance, there are variables relevant to farmers that are only manually analyzed by banks. The model is built using a random forest with tuning parameters, resulting in an accuracy of 83%. This model has been evaluated using test data, which results in an accuracy of 88%. The model also produced an AUC score of 0.81. Based on the AUC performance classification, it can be concluded that the model is good at predicting the credit status of farmers because the AUC value is included in the good classification predicate.

Reference

[1] Ashari, "Optimalisasi Kebijakan Kredit Program Sektor Pertanian Di Indonesia," *Anal. Kebijak. Pertan.*, vol. 7, no. 1, pp. 21–42, 2016, doi: http://dx.doi.org/10.21082/akp.v7n1.2009.21-42.

[2] BPS, *Hasil Survei Struktur Ongkos Usaha Tanaman Holtikultura (SOUH) 2018*. Jakarta: Badan Pusat Statistik Indonesia, 2018.

[3] D. Pratiwi, M. Ambayoen, and A. Hardana, "Studi Pembiayaan Mikro Petani Dalam Pengambilan Keputusan Untuk Kredit Formal dan Kredit Nonformal," *HABITAT*, vol. 30, no. 1, pp. 35–43, Apr. 2019, doi: 10.21776/ub.habitat.2019.030.1.5.

[4] Safira, Bappenas, Australian Goverment, and Grow Asia, "Penilaian Kredit Digital di Sektor Pertanian: Praktik Terbaik Penilaian Risiko Kredit dalam Rantai Nilai," 2018.

[5] N. Wahyuni, "Penerapan Prinsip 5C Dalam Pemberian Kredit Sebagai Perlindungan Bank," *Lex J. Kaji. Huk. Keadilan*, vol. 1, no. 1, 2017, doi: 10.25139/lex.v1i1.236.

[6] O. Dengiz and M. Sağlam, "Determination of land productivity index based on parametric approach using GIS technique," *Eurasian J. Soil Sci.*, vol. 1, no. 1, pp. 51–57, 2012, doi: 10.18393/ejss.12237.

[7] O. E. Emmanuel, N. C. Ehirim, E. . Eze, and M. N. Osuji, "Analysis Of Socio-Economic Variables On Agricultural Productivity Of Some Selected Arable Crops In Imo State. Nigeria," vol. 16, no. 1, pp. 1385–1391, 2013.

[8] T. Urgessa, "Review on the Determinants of Agricultural Productivity and Rural Household Income in Ethiopia," *Ethiop. J. Ecnomoucs*, vol. 24, no. 2, 2015, doi: 10.7176/jesd/11-18-01.

[9] Z. Khemais, D. Nesrine, and M. Mohamed, "Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression," *Int. J. Econ. Financ.*, vol. 8, no. 4, p. 39, 2016, doi: 10.5539/ijef.v8n4p39.

[10] X. Zhang, Y. Yang, and Z. Zhou, "A novel credit scoring model based on optimized random forest," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan. 2018, pp. 60–65, doi: 10.1109/CCWC.2018.8301707.

[11] Y. Fengge and W. Jing, "Agriculture microfinance risk control based on credit score model in China," in *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering*, Nov. 2013, pp. 316–320, doi: 10.1109/ICIII.2013.6703581.

[12] C. Rao, M. Liu, M. Goh, and J. Wen, "2-stage modified random forest model for credit risk assessment of P2P network lending to 'Three Rurals' borrowers," *Appl. Soft Comput. J.*, vol. 95, p. 106570, 2020, doi: 10.1016/j.asoc.2020.106570.

[13] S. Y. Sohn, D. H. Kim, and J. H. Yoon, "Technology credit scoring model with fuzzy logistic regression," *Appl. Soft Comput. J.*, vol. 43, pp. 150–158, 2016, doi: 10.1016/j.asoc.2016.02.025.

[14] S. I. Hermawan and R. F. Malik, "Credit Scoring Menggunakan Algoritma Classification And Regression Tree (CART)," in *Prosing Annual Research Seminar 2016*, 2016, vol. 2, no. 1, pp. 82–85, [Online]. Available: http://ars.ilkom.unsri.ac.id.

[15] H. Van Sang, N. H. Nam, and N. D. Nhan, "A novel credit scoring prediction model based on feature selection approach and parallel random forest," *Indian J. Sci. Technol.*, vol. 9, no. 20, 2016, doi: 10.17485/ijst/2016/v9i20/92299.

[16] H. Xiao, Z. Xiao, and Y. Wang, "Ensemble classification based on supervised clustering for credit scoring," *Appl. Soft Comput. J.*, vol. 43, pp. 73–86, 2016, doi: 10.1016/j.asoc.2016.02.022.

[17] C. Luo, D. Wu, and D. Wu, "A deep learning approach for credit scoring using credit default swaps," *Eng. Appl. Artif. Intell.*, vol. 65, no. September, pp. 465–470, Oct. 2016, doi: 10.1016/j.engappai.2016.12.002.

[18] A. Bequé and S. Lessmann, "Extreme learning machines for credit scoring: An empirical evaluation," *Expert Syst. Appl.*, vol. 86, pp. 42–53, 2017, doi: 10.1016/j.eswa.2017.05.050.

[19] F. Garunescu, *Data Mining: Concepts and Techniques*. Springers, 2011.

[20] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, and H. Yuanyue, "Learning from class-imbalanced data : Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[21] A. S. More and D. P. Rana, "Review of random forest classification techniques to resolve data imbalance," in *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, Oct. 2017, pp. 72–78, doi: 10.1109/ICISIM.2017.8122151.

[22] M. M. Ramadhan, I. S. Sitanggang, F. R. Nasution, and A. Ghifari, "Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency," in *DEStech Transactions on Computer Science and Engineering*, 2017, no. cece, doi: 10.12783/dtcse/cece2017/14611.

DOI: https://doi.org/10.29207/resti.v7i1.4583