South German Credit Data Classification Using Random Forest Algorithm to Predict Bank Credit Receipts

Yoga Religia[1*)], Gatot Tri Pranoto[2], Egar Dika Santosa[3]
[1]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[2]Program Studi Teknik Informatika, Fakultas Industri Kreatif dan Telematika, Universitas Trilogi
[2]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
email: [1]yoga.religia@pelitabangsa.ac.id, [2]gatot.pranoto@trilogi.ac.id, [3]dikasantosaegar@gmail.com

Abstract *–* **Normally, most of the bank's wealth is obtained from providing credit loans so that a marketing bank must be able to reduce the risk of non-performing credit loans. The risk of providing loans can be minimized by studying patterns from existing lending data. One technique that can be used to solve this problem is to use data mining techniques. Data mining makes it possible to find hidden information from large data sets by way of classification. The Random Forest (RF) algorithm is a classification algorithm that can be used to deal with data imbalancing problems. The purpose of this study is to discuss the use of the RF algorithm for classification of South German Credit data. This research is needed because currently there is no previous research that applies the RF algorithm to classify South German Credit data specifically. Based on the tests that have been done, the optimal performance of the classification algorithm RF on South German Credit data is the comparison of training data of 85% and testing data of 15% with an accuracy of 78.33%.**

Keywords – data mining, classification, random forest, bank credit receipts.

I. INTRODUCTION

Credit financing is a provider of funds based on a loan and loan agreement between the customer and the Bank, which requires the borrower to pay off the loan at a certain time [1]. Every division of marketing banking needs to select prospective customers to be given credit by considering several things such as trust, risk level, grace period and credit object. This selection is necessary because a marketer banking must be able to protect his customers so that non-performing loans do not occur, where non-performing loans are often the main risk in every loan distribution [2]. One way that can be used to reduce the risk of applying for credit is by using data mining techniques so that it is possible to mine information from pre-existing credit application data sets [3].

Data mining (DM) is a powerful technique for finding meaningful and useful information from large data sets, so it is very useful for application in the real world [4]. Data mining techniques are generally divided into two categories, namely predictive and descriptive. The predictive method of data mining can be done with a classification model. Classification is the process of converting records data into a set of the same class [5]. The website www.kaggle.com has provided a dataset South German Credit consisting of 800 records credit application with 21 attributes and there is no missing value, so it can be used to build a classification model [6]. Based on the 21 existing attributes, the type of data from the data set is Bank Marketing included in the data category imbalance, so it requires the right algorithm to classify the data.

Several studies suggest that the algorithm Random Forest (RF)can be used for classification of data imbalance in large amounts of data by providing good performance results and relatively fast execution times [7] [8]. Research conducted by Wei Chen (2017) states that in terms of classification, the RF algorithm is able to provide greater accuracy when compared to algorithms such as the logistic model tree (LMT) and classification and regression tree (CART) [9].

Currently, there is no previous research that applies the RF algorithm to classify South German Credit data specifically. Based on this, this study will discuss the use of the RF algorithm for data classification of South German Credit. This research is expected to provide a contribution to the study regarding the making of prediction models for credit receipts in the banking world.

II. LITERATURE REVIEW

A. Acceptance of Bank Credit

Credit comes from the Italian language, namely Credere, which means trust. Trust in question is the trust of the creditor that the debtor will repay the loan and the interest in accordance with the agreement of both parties [10]. The implementation of credit extension usually goes through several stages, namely credit application, checking credit applications, credit analysis, credit approval, credit realization, and finally credit monitoring [11]. Normally, most of the bank's wealth is in the form of credit which is the source of bank income, therefore it is often referred to as productive assets. In channeling credit, management must use the principle of prudence so that loans are granted in the current category. However, there are often some customers whose interest and principal payments are not smooth and therefore fall into the non-performing loan (NPL) category. The higher the NPL, the greater the potential loss, so that banks must reduce their lending [12].

B. Data Mining

Data mining is one of the most important areas of research that aims to obtain information from data sets. Data mining is the process of extracting meaningful information and structures across complex data sets [13]. Data mining began in the 1990s as an effective way of extracting previously unknown patterns and information from data sets [14]. Data mining techniques are used to find the relationship between data to perform classifications that predict the values of several variables (classification), or to divide known data into groups that have similar characteristics (clustering). In its implementation, data mining can use various parameters to check data including association, classification and clustering. Data mining involves key steps which include problem definition, data exploration, data preparation, modeling, and evaluating and deployment [15]. Using data mining techniques makes it possible to search, analyze, and sort large data sets to find new patterns, trends, and relationships contained therein [16].

C. Classification by Random Forest

Random Forest (RF) is an algorithm that uses a recursive binary separation method to reach the end nodes in a tree structure based on a classification and regression tree [16]. Breiman in 2001 introduced the RF algorithm by showing several advantages including being able to produce relatively low errors, good performance in classification, being able to efficiently handle large amounts of training data, and an effective method for estimating missing data. RF generates multiple independent trees with subsets randomly selected via bootstrap from the training sample and from the input variables at each node. RF performs classification by adopting an ensemble approach of multiple trees through majority occurrences to reach a final decision [17].

The training dataset in the RF algorithm is formulated as $S = \{(x_i, y_j), i = 1, 2,\ldots, N; j = 1, 2,\ldots, M\}$, where x is the sample and y is the feature variable S. N is the number of training samples, and there is a feature variable M in each sample [18]. As for the development of the RF algorithm, it consists of 3 steps, namely: (1) Sampling of k training subsets, (2) Making each decision tree model, and (3) Collecting k trees into the RF model. The use of RF algorithms for classification can be applied to data imbalance large amounts of by providing good performance results and fast execution times [7] [8].

D. Algorithm Performance Testing

The classification performance appraisal indicator is very important for evaluating the performance of any machine learning algorithm. There are many assessment indicators in the field of classification including accuracy, precision, recall, area under the curve (AUC), Receiver Operating Characteristics (ROC), etc. However, classification testing by looking at the accuracy value is the most frequently used. Accuracy is the percentage of target and non-target samples correctly predicted and reflects the classifier ability to define the entire sample [19]. The accuracy can be measured by the following equation:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100$$

True Positive (TP) is the number of positive samples that are predicted to be correct; False Positive (FP) is the number of positive samples whose predictions are wrong; True Negative (TN) is the number of negative samples correctly predicted; False Negative (FN) is the number of negative samples that are predicted to be wrong.

III. RESEARCH METHODOLOGY

A. Data used

This study uses secondary data taken from the website www.kaggle.com in the form of a dataset of South German Credit [6]. The dataset South German Credit consists of three variables, namely bank client data, client last contact, and label. Each variable has its own attributes where the attributes of each variable can be seen in Table 1. The number of records data contained in data the Marketing Bank is 45,211 records consisting of 16 attributes and there is no missing value, so it does not require pre-processing data.

Table 1. Atribut Data *South German Credit*

Atribut	Information
status	status of the debtor's checking account with the bank (categorical)
duration	credit duration in months (quantitative)
credit history	history of compliance with previous or concurrent credit contracts (categorical)
purpose	purpose for which the credit is needed (categorical)
amount	credit amount in DM (quantitative; result of monotonic transformation; actual data and type of trans...
savings	debtor's savings (categorical)
employment duration	duration of debtor's employment with current employer (ordinal; discretized quantitative)
installment rate	credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative...
personal status sex	combined information on sex and marital status; categorical; sex cannot be recovered from the variab...
other debtors	Is there another debtor or a guarantor for the credit? (categorical)
present residence	length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitative...
property	the debtor's most valuable property, i.e. the highest possible code is used. Code 2 is used, if code...
age	age in years (quantitative)
other installment plans	installment plans from providers other than the credit-giving bank (categorical)

Atribut	Information
housing	type of housing the debtor lives in (categorical)
number credits	number of credits including the current one the debtor has (or had) at this bank (ordinal, discretiz...
job	quality of debtor's job (ordinal)
people liable	number of persons who financially depend on the debtor (i.e., are entitled to maintenance) (binary,d...
telephone	Is there a telephone landline registered on the debtor's name? (binary; remember that the data are f...
foreign worker	Is the debtor a foreign worker? (binary)
credit risk	Has the credit contract been complied with (good) or not (bad) ? (binary)

Data South German Credit is data that has been free of the missing value,so it can be directly used for the classification process because no longer need the data preprocessing.

B. Research Model

This study uses classification modeling to be used ondata South German Credit, where the resulting label is whether credit applications are accepted or rejected. The classification model is built using a process split validation to divide South German Credit into data training and data testing data. The validation process was carried out 5 times for the purposes of analyzing the classification results. Each validation process is carried out by comparing training data and testing data as seen in Table 2.

Table 2. Split Data Training dan Data Testing

No	Training : Testing	Training	Testing
1	75% : 25%	600	200
2	80% : 20%	640	160
3	85% : 15%	680	120
4	90% : 10%	720	80
5	95% : 5%	760	40

After the data is split, the classification will be carried out using the RF algorithm to measure its accuracy performance. The modeling stages in this study can be seen in Figure 1.

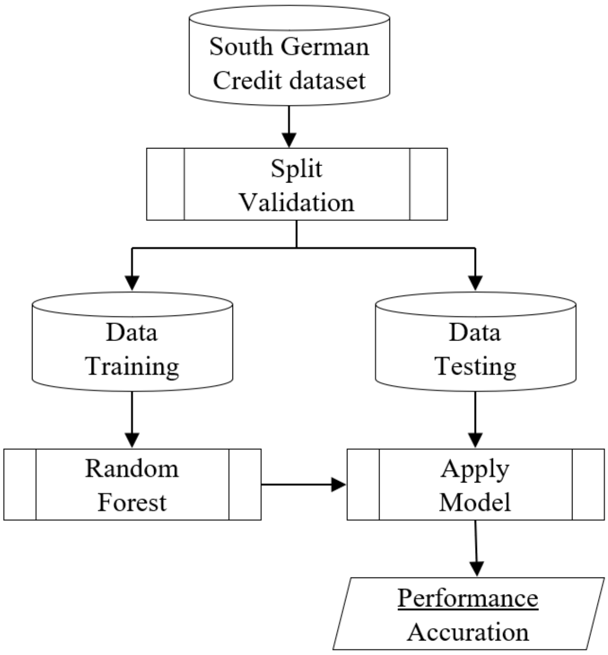

Figure 1. Research Model

After the stages of the research model shown in Figure 1 have been completed, this study will discuss the test results. discussion of the results of this test is needed to clarify the findings of this study.

IV. RESULTS AND DISCUSSION

A. Testing Steps

Testing in this study was carried out using the Rapid Miner version 5.0 tools. The choice of Rapid Miner tools was made because Rapid Miner was considered to be used for research, rapid prototyping, and supports all steps of the data mining process such as data preparation, result visualization, validation, and optimization [21], so it is suitable for use in this study. The first step in making this research model is calling the South German Credit, data after the data has been summoned then the data is distributed to the split validation process. As discussed in the research model section, the validation process is carried out 5 times for the purposes of analyzing the classification results. More details about the data calling and validation process can be seen in Figure 2.

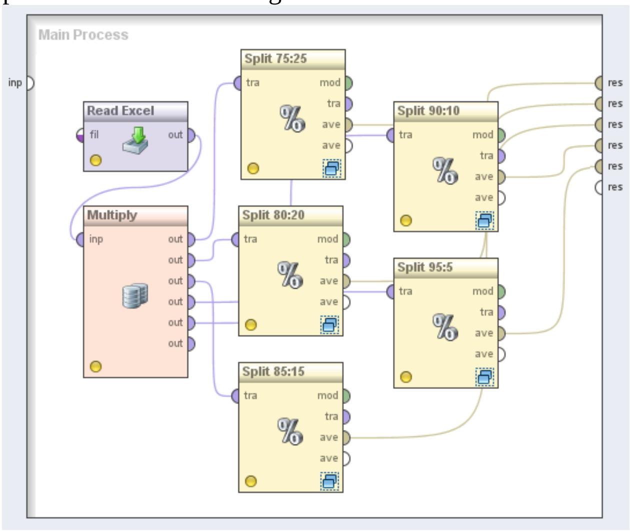

Figure 2. Data Calling and Validation Process

In each validation process in Figure 2, it contains a learning process with the RF algorithm which is then applied in the model to measure its accuracy performance. The learning process formed in this study can be seen in Figure 3.

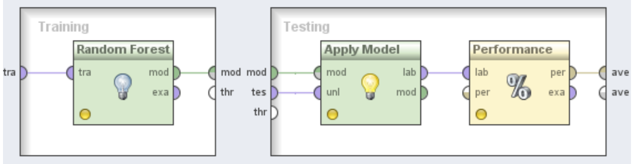

Figure 3. Learning Process and Apply Model

After all research models have been formed, the next step is to run the model that has been built in RapidMiner to see the results of the accuracy obtained.

B. Test result

Based on the tests that have been carried out, the results are as shown in Table 3.

Table 3. Test Result

No	Training : Testing	Akurasi
1	75% : 25%	72,50 %
2	80% : 20%	74,38 %
3	**85% : 15%**	**78,33** %
4	90% : 10%	77,50 %
5	95% : 5%	70,00 %

The test results in Table 3 show that the highest accuracy is obtained from training data of 85% and testing data of 15% with an accuracy of 78.33%. The lowest accuracy value is obtained from training data of 95% and testing data of 5% with an accuracy of 70.00%. Based on these results, it can be calculated that the average accuracy of the data classification South German Credit using the RF algorithm is 74.54%. When viewed from the results of the accuracy value from the lowest to the highest accuracy value from the distribution of training data and testing data is shown in Figure 4.

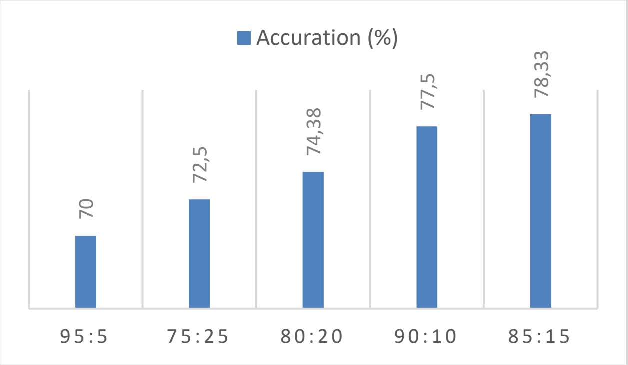

Figure 4. Comparison of Test Results

C. Discussion of Results

Based on the results of the RF algorithm classification test on the data South German Credit, it is known that the accuracy value obtained from 5 tests is still very small with an average accuracy of 74.54%. The small value is probably due to the data returns that are too high from the data used. The imbalance of this data results in the making tree in the RF algorithm process to be too biased. Although according to several studies, it is revealed that the RF algorithm is suitable for handling data imbalance, in fact this study proves that the RF algorithm has not been able to classify South German Credit data which has a fairly high data imbalance. In addition, this study shows that the amount of training data does not necessarily have an impact on the performance of the RF algorithm. This is evidenced by the increasing amount of non-linear training data with the increasing performance of the RF algorithm.

V. CONCLUSION

This study has tested the use of the RF algorithm to classify data South German Credit. The tests that have been carried out show the following results:

1. The highest accuracy of using the RF algorithm to classify data South German Creditis 78.33%, the lowest accuracy is 70%, while the average accuracy is 74.54%.
2. This study found that the higher the amount of training data is not linear with the increasing performance of the RF algorithm.
3. The RF algorithm has not been able to provide a good performance for classifying data South German Credit.

This study has not been able to provide a good enough accuracy value for the classification of data South German Credit, so further research is needed to obtain a better classification model. Based on the findings of this study, it is suggested that further research can apply optimization methods to further optimize the performance of the RF algorithm.

BIBLIOGRAPHY

[1] A. T. Rahmawati, M. Saifi dan R. R. Hidayat, "Analisis Keputusan Pemberian Kredit dalam Langkah Meminimalisir Kredit Bermasalah," *Jurnal Administrasi Bisnis,* vol. 35, no. 1, pp. 179-186, 2016.

[2] S. Somadiyono dan T. Tresya, "Tanggung Jawab Pidana Marketing Menurut Undang Undang Perbankan Terhadap Pembiayaan Bermasalah di Bank Muamalat Indonesia,Tbk," *Jurnal Lex Specialis,* vol. 21, pp. 22-38, 2015.

[3] S. Masripah, "Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit," *Bina Insani ICT Journal,* vol. 3, no. 1, pp. 187-193, 2016.

[4] W. Gan, J. C.-W. C. H.-C. Lin dan J. Zhan, "Data mining in Distributed Environment: A Survey," *Wiley Interdiscriplinary Reviews: Data Mining and Knowledge Discovery* , vol. 7, no. 6, pp. 1-19, 2017.

[5] S. Umadevi dan K. S. J. Marseline, "A Survey on Data Mining Classification Algorithms," dalam *International Conference on Signal Processing and Communication,* Coimbatore, India, 2017.

[6] "Kaggle," kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/c/south-german-credit-prediction/overview/data-overview. [Diakses 2 November 2020].

[7] A. S. More dan D. P. Rana, "Review of Random Forest Classification Techniques to Resolve Data Imbalance," dalam *International Conference on Intelligent Systems and Information Management*, Aurangabad, India, 2017.

[8] A. Parmar, R. Katariya dan V. Patel, "A Review on Random Forest: An Ensemble Classifier," dalam *International Conference on Intelligent Data Communication Technologies and Internet of Things*, Springer, Cham, 2018.

[9] W. Chen, X. Xie, B. Pradhan, H. Hong, D. T. Bui, Z. Duan dan J. Ma, "A Comparative Study of Logistic Model Tree, Random Forest, and Classification and Regression Tree Models for Spatial Prediction of Landslide Susceptibility," *Catena* , vol. 151 , pp. 147-160, 2017.

[10] M. S. Hasibuan, Dasar-Dasar Perbankan, Jakarta: PT Bumi Aksara, 2004.

[11] R. Widayati dan M. Efriani, "Aktivitas Pemberian Kredit Usaha Pada PT. Bank Perkreditan Rakyat Batang Kapas," dalam *OSF Preprints*, Batang, Indonesia, 2019.

[12] B. Panuntun dan Sutrisno, "Faktor Penentu Penyaluran Kredit Perbankan Studi Kasus Pada Bank Konvensional Di Indonesia," *Jurnal Riset Akuntansi & Keuangan Dewantara,* vol. 1, no. 2, pp. 57-66, 2018.

[13] M. S. Başarslan dan I. D. Argun, "Classification Of A Bank Data Set On Various Data Mining Platforms," dalam *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, Istanbul, Turkey, 2018.

[14] D. Tomar dan S. Agarwal, "A survey on Data Mining approaches for Healthcare," *International Journal of Bio-Science and Bio-Technology,* vol. 5, no. 5, pp. 241-266, 2013.

[15] V. Krishnaiah, G. Narsimha dan N. Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," *International Journal of Computer Science and Information Technologies*, vol. 4, no. 1, pp. 39-45, 2013.

[16] K. Sumiran, "An Overview of Data Mining Techniques and Their Application in Industrial Engineering," *Asian Journal of Applied Science and Technology,* vol. 2, no. 2, pp. 947-953, 2018.

[17] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[18] C. Yoo, D. Han, J. Ima dan B. Bechtel, "Comparison Between Convolutional Neural Networks and Random Forest for Local Climate Zone Classification in Mega Urban Areas Using Landsat Images," *Journal of Photogrammetry and Remote Sensing,* vol. 157, pp. 155-170, 2019.

[19] J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng dan K. Li, "A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment," *IEEE Transactions on Parallel and Distributed Systems,* vol. 28, no. 4, pp. 919-933, 2017.

[20] J. Lin, H. Chen, S. Li, Y. Liu, X. Li dan B. Yu, "Accurate Prediction of Potential Druggable Proteins Based on Genetic Algorithm and Bagging-SVM Ensemble Classifier," *Artificial Intelligence In Medicine,* vol. 98, pp. 35-47, 2019.

[21] A. Jeyaraj, R. S dan M. R. Raja, "A study of classification algorithms using Rapidminer," *International Journal of Pure and Applied Mathematics,* vol. 119, no. 12, pp. 15977-15988, 2018.