ComTech: Computer, Mathematics and Engineering Applications, 14(1), June 2023, 11-20 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v14i1.8243 E-ISSN: 2476-907X

Data-Driven Approach for Credit Risk Analysis Using C4.5 Algorithm

Muhammad Iqbal[1]* and Syahril Efendi[2]

[1]Program Studi Sistem Komputer, Fakultas Sains dan Teknologi, Universitas Pembangunan Panca Budi
Jln. Jend. Gatot Subroto KM 4.5, Sumatera Utara 20122, Indonesia
[2]Ilmu Komputer, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara
Jln. Dr. T. Mansur No. 9, Sumatera Utara 20222, Indonesia
[1]muhammadiqbal@dosen.pancabudi.ac.id; [2]syahril3fendiusu@gmail.com

Received: 19th February 2022/ **Revised:** 21st August 2022/ **Accepted:** 22nd August 2022

How to Cite: Iqbal, M., & Efendi, S. (2023). Data-Driven Approach for Credit Risk Analysis Using C4.5 Algorithm. *ComTech: Computer, Mathematics and Engineering Applications, 14*(1), 11−20.
https://doi.org/10.21512/comtech.v14i1.8243

Abstract **-** Credit risk is bad credit, resulting in bank losses due to non-receipt of disbursed funds and unacceptable interest income. However, credit services still have to be done to achieve profit. The absence of an approach that can assist in making policies to reduce credit risk makes the risk opportunities even more significant. So, data processing techniques are needed that produce information to be used as the basis for policies in triggering credit risk with data mining. The research presented an application of data mining as a credit risk approach considering the ability of data mining techniques to extract data into useful information with the C4.5 algorithm. The research used a sample of 30 data banks with 6 factors (credit growth, net interest margin, type of bank, capital ratio, company size, and bank compliance level). Credit risk was evaluated by making a decision tree and a RapidMiner test application. The results show that credit growth is the main factor causing credit risk, followed by bank compliance level, net interest margin, and capital ratio. Based on the results obtained, the C4.5 algorithm can be used in analyzing credit risk with results that are easy to understand and can be used as useful information for banks.

Keywords: data-driven approach, credit risk analysis, C4.5 algorithm

I. INTRODUCTION

Banks are one of the sectors with a very important role in moving the national economy (Lihani, Ngadiman, & Hamidi, 2013). However, in carrying out their business, banks often experience upheavals. Those are not only caused by the world economy but also by the bank itself related to borrowing or credit (Hakim & Oktaria, 2018; Saputro, Sarumpaet, & Prasetyo, 2019).

Moreover, many services are provided by the banking world, one of which is lending services. The existence of credit in the banking world is a form of service that is usually carried out in banking activities (Hanif, 2015). Moreover, many services are provided by the banking world, one of which is credit services. The existence of credit in the world of banking is a form of service that is usually carried out in banking activities. In this case, the research usually only analyzes one bank or the same type of bank that provides credit services based on the pattern of debtors who have been given credit before (Aji & Manda, 2021; Cristina & Artini, 2018; Wijaya & Tiyas, 2019).

Many factors can cause credit risk. So, it is essential to pay attention to the credit risk that may be experienced by banks, such as credit growth, net interest margin, bank type, capital ratio, company size, and level of compliance (Hakim & Oktaria, 2018). Furthermore, it is necessary to take advantage of technological developments in producing information quickly accompanied by data processing techniques to analyze data related to these factors correctly to produce information that can be used with data mining techniques. Hence, it can avoid upheaval in the banking world related to credit and assist in making policies to avoid credit risks that may arise.

Data mining refers to the process of extracting knowledge from big data (Wang, Zhou, & Xu, 2019). Data mining is also mentioned as a series of processes to explore added value in the form of knowledge that has not been known manually from a data set (Susanto & Indriyani, 2019). Data mining is also the mining or discovery of new information by looking for certain patterns or rules from a very large amount of data (Fauzi, Marpaung, & Pardede, 2018). Data mining is also the process of extracting data into information that has not been conveyed before. With the proper

*Corresponding Author

techniques, the data mining process will provide optimal results (Riandari & Sihotang, 2020). Data mining is also defined as a process that employs one or more computer learning techniques (machine learning) to analyze and extract knowledge automatically (Zulfami, 2017). According to Le (2022), data mining is a technology that is an organic combination of data warehouse technology and comprehensive database integration technology. It supports various algorithms to meet different mining needs. Data that have gone through the Knowledge Discovery in Database (KDD) process is also known as control data (data-driven). With the rapid development of data mining, many studies are used to predict a case, especially banking (Subarkah, Pambudi, & Hidayah, 2020).

Various techniques are available in data mining for knowledge extraction, including predicting, estimating, associating, clustering, and classifying (Ariawan, 2019). Similarly, according to Harlina (2018), data mining is divided into several groups based on the tasks that can be done: description, estimation, prediction, classification, clustering, and association. First, descriptions of patterns and trends often provide possible explanations for a pattern or trend. Second, estimation is almost the same as classification, except that the target variable is more numerical than categorical. The model is built using a complete record that provides the value of the target variable as the predicted value.

Furthermore, the estimated value of the target variable is made based on the value of the predictive variable. Third, prediction is almost the same as classification and estimation, except that the value of the results will be in the future. Fourth, in the classification, there is a categorical variable target. Fifth, clustering is grouping records and observing and forming classes of objects with similarities with one another and dissimilarities with records in other clusters. Last, the task of association in data mining is to find attributes that appear at one time.

Classification is also one of the main tasks of data mining. Classification means analyzing data patterns in the training set to find an accurate description model of each category and generalizing the known structures to apply them to new data. The classification procedure includes data acquisition, feature selection, model selection, training, and evaluation. Data to be used for training and testing must be collected beforehand. Then, feature selection is influenced by previous feature descriptions from the data set. Classifiers must be trained to define system parameters. Usually, there are several repetitions of the previous procedure based on the results of the previous evaluation to create better results (Liu, Jin, & Liu, 2011). Classification is the process of finding a model or function that distinguishes concepts or data classes. It predicts the object class whose class label is unknown based on the analysis of training data (data objects whose class is known) (Afrianto, Suseno, & Warsito, 2020). It is the most commonly applied data extraction technique to predict categorical attribute values (discrete or nominal) (Bedregal-Alpaca, Cornejo-Aparicio, Zarate-Valderrama, & Yanque-Churo, 2020).

One of the extraction techniques that is often used in prediction and classification is the C4.5 algorithm. The C4.5 algorithm is a classification algorithm with a decision tree technique that is well-known and preferred because of its advantages. For example, it can process numeric (continuous) and discrete data, handle missing attribute values, and generate rules that are easy to understand and the fastest among other algorithms.

In addition, the algorithm is a collection of commands written systematically to solve mathematical logic problems. Understanding the C4.5 algorithm can be used to control a device. Meanwhile, the decision tree can be interpreted as a powerful way of predicting or clarifying. Decision trees can divide large data sets into sets. There are many nodes in the decision tree and a number of nodes representing tests on a particular attribute, which have spread to the sample size in the lowest category of leaf nodes. There are many types of decision trees. The most famous algorithm in the industry is developed by the Rosquin Institute, which is mainly used to generate decision trees (An & Zhou, 2022).

The C4.5 algorithm is also one of the algorithms for converting big facts into a decision tree that represents the rules. The purpose of forming a decision tree in the C4.5 algorithm is to make it easier to solve existing problems. There are stages in turning the C4.5 algorithm into a rule (Kurniawan, Anggrawan, & Hairani, 2020). Thus, the decision tree is a classification method that uses a tree structure representation where each node represents an attribute, the branch represents the value of the attribute, and the leaf represents the class. The top node of the decision tree is called the root (Hozeng & Aisa, 2016).

The decision tree approach can potentially improve prediction accuracy as it plays a promising role in decision-making (Ramos, Faria, Morais, & Vale, 2022). Another previous research predicts lung cancer risk using the support vector machines classification technique, C4.5, and Naive Bayes algorithms for health services analysis. It concludes that the C4.5 algorithm predicts better (Pradeep & Naveen, 2018). Another previous research applies data mining using the decision tree method for predicting credit risk determination and focusing on building a BRI credit scoring model with a decision tree technique. It concludes that the decision tree technique can build a model with an objective and produce an easy-to-understand model and a high level of accuracy (Hozeng & Aisa, 2016).

Based on the previous explanation, a credit risk evaluation is carried out at the lending/credit service provider in the bank. The research is based on the factors that cause credit risk with a data mining approach with the C4.5 algorithm in pre-processing data according to the needs of the approach used. It helps the research to run in accordance with the

expected goals. The research aims to see how the application of data mining with the C4.5 algorithm analyzes credit risk. The results obtained are in the form of a decision tree that describes the pattern of causes of bad loans. It is expected to provide information about the pattern of causes of bad loans based on the variables that are beneficial to banks. These results are also expected to be a tool for banks in providing credit services to take appropriate policies so that credit risk does not occur for non-bank service providers so that service providers cannot only survive in running their business but also increasing company productivity and achieving the company's main goals.

II. METHODS

In research, the data mining technique for performing classification is the C4.5 algorithm. The C4.5 algorithm constructs a decision tree from training data as cases or records (tuples) in the database (Riandari & Simangunsong, 2019). The C4.5 decision tree algorithm was proposed by JR Quinlan in 1993 (Wang & Gao, 2021). It is a decision tree making algorithm based on the ID3 algorithm. The C4.5 algorithm overcomes the shortcomings of the ID3 algorithm in the application. In the C4.5 algorithm, the information gain rate is used as the basis for selecting test attributes (Wang & Gao, 2021). The C4.5 algorithm is chosen in the research because it can make predictions by providing an ideal level of accuracy in predicting credit risk.

The research has several stages, as seen in Figure 1. The initial stage of research is to formulate the problem in accordance with the problems that occur and the goals to be achieved. Then, the research focuses on how to apply data mining with the C4.5 algorithm in analyzing credit risk. After the formulation of the problem, the research objectives are formed to answer the predetermined problem formulation. Finally, the research objective is to apply data mining with the C4.5 algorithm in credit risk analysis. Furthermore, data and information collection on data mining with the C4.5 algorithm are carried out through literature studies through books, previous research, and other media related to research.

The implementation of the C4.5 algorithm analysis has several stages. After the problem to be analyzed is found, it process data related to the research objectives. Implementing the C4.5 algorithm is carried out after the data to be processed is complete and in accordance with the needs. The data processing process before being processed with the C4.5 algorithm is carried out according to the KDD stages. First, in data selection, the research object is 30 banks providing loan/credit services as samples. The qualitative data contain information about every factor that can lead to credit risks, such as credit growth, net interest margin, bank type, capital ratio, company size, and bank compliance level. Second, it is pre-processing/ data cleaning. After 30 samples of bank data have been taken, which are equipped with information from each factor, the data from the selection are processed to the data cleaning stage to eliminate data that are not appropriate/noisy and have the same value. This stage discards bank data with no value and same value in all factors used, such as alternatives A and B which are different alternatives but have the same value in all the variables used. So, one variable is discarded. At this stage, different patterns are searched. If a similar pattern is found, only one representative pattern will be left, and the rest will be cleaned. Because none of the samples used are noisy and have the same value, the final result of this stage finds the same amount of data as the sample, as many as 30 data bank samples. Third, in transformation, the previously processed data (qualitative data) are grouped and transformed into an appropriate assessment form to be processed in data mining. In the research, the data were converted into quantitative form to make it easier to define during testing. Based on the 30 data samples, the classification stage is carried out using the C4.5 algorithm approach, which produces a decision tree. It identifies patterns that cause credit risk in banks that provide loan/credit services based on existing factors.

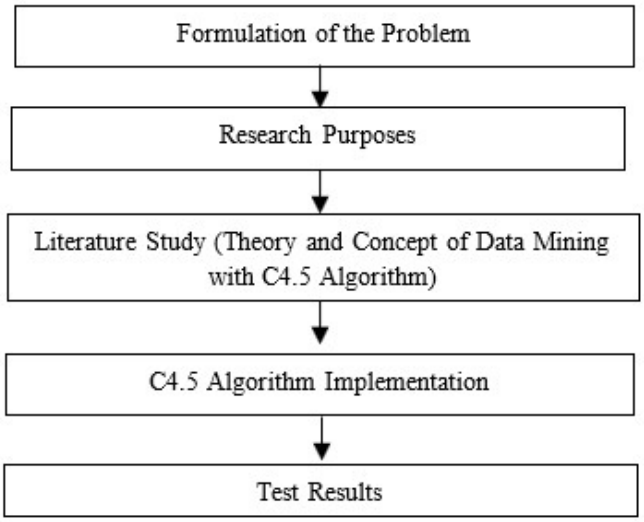

Figure 1 Research Stages

An attribute as root is selected based on the highest gain value of the existing attribute. Equation (1) is used to calculate the profit. It has *S* as the case set, *A* as the attribute, *n* as the number of partitions of *S*, and *Pi* as the proportion of *Si* to *S*. Meanwhile, Equation (2) calculates the entropy value. It has *S* as a set of cases, *A* as an attribute, *N* as the number of partitions in attribute *A*, |*Si*| as the number of cases on the *i*-th partition, *and* |*S*| as the number of cases in *S*.

$$Gain(S, A) = Entrophy(S) - \sum_{i=1}^{n} -\frac{|s|}{|si|} xEntrophy(Si) \quad (1)$$

$$Entropy\ (total) = -\sum_{i=1}^{n} - Pi\ x\ Log2\ Pi \quad (2)$$

To make a decision tree, the first step that needs to be done is to count the total number of cases: the number of cases with a low credit risk statement (S1) and the number of cases with a high credit risk statement (S2). Then, these cases are divided based on credit risk factors, such as credit growth factors, net interest margin, bank type, capital ratio, company size, and bank compliance level. It is calculated for each case obtained on each factor.

There are several steps involved in making a decision tree. It specifies the factor as the root and calculates the attribute value to get the information. It is done based on the highest gain value of all the existing factors to determine the factor as the root. The entropy value is needed to determine the highest gain.

According to Ginting, Kusrini, and Taufiq (2020), the purpose of the evaluation stage is to obtain the results of the analysis regarding banks providing loan/credit services. In addition, it is to see the potential to accurately assess credit risk based on credit growth factors, net interest margin, bank type, capital ratio, company size, and bank compliance level. Finally, the results are tested with one of the data mining testing applications, RapidMiner.

III. RESULTS AND DISCUSSIONS

Based on the KDD stages, it starts from the data selection, pre-processing, transformation, data mining, and evaluation stages. Then, the testing data obtained are used in the discussion, as seen in Table 1. The data presented in Table 1 consist of 30 bank data with an assessment of each attribute. The assessment of the attributes consists of five categories, namely high, currently, low, big, and small. The data have gone through the stages of selection, pre-processing, and transformation. The data also have different pattern and are ready to be analyzed using the C4.5 algorithm.

Table 1 Testing Data in the Research

List of Bank	Credit Growth	Net Interest Margin	Bank Type	Capital Ratio	Company Size	Bank Compliance Level	Credit Level
1	High	Low	Country	Big	Big	High	Low
2	High	Low	Public	Big	Big	High	Low
3	High	High	Country	Big	Big	High	Low
4	High	High	Public	Big	Big	High	Low
5	High	High	Public	Small	Big	High	Low
6	High	High	Country	Small	Big	High	Low
7	High	High	Country	Big	Small	High	Low
8	High	High	Country	Big	Small	Low	Low
9	Currently	Low	Country	Big	Big	High	Low
10	Currently	Low	Public	Big	Big	High	Low
11	Currently	Low	Public	Small	Big	High	Low
12	Currently	Low	Public	Small	Small	High	Low
13	Currently	Low	Public	Small	Small	High	High
14	Currently	High	Country	Big	Big	High	Low
15	Currently	High	Public	Big	Big	High	Low
16	Currently	High	Public	Small	Big	Low	Low
17	Currently	High	Public	Small	Small	Low	High
18	Currently	High	Public	Big	Small	Low	Low
19	Currently	High	Country	Small	Big	High	Low
20	Currently	High	Country	Big	Small	High	Low
21	Currently	High	Country	Big	Small	Low	Low
22	Low	Low	Country	Big	Big	High	High
23	Low	Low	Public	Big	Big	High	High
24	Low	Low	Public	Small	Big	High	High
25	Low	High	Country	Big	Big	High	High
26	Low	High	Public	Big	Big	High	High
27	Low	High	Public	Big	Small	Low	High
28	Low	High	Country	Small	Big	High	High
29	Low	High	Country	Small	Small	Low	High
30	Low	High	Public	Big	Small	Low	High

Based on Table 1, the initial step to be carried out is to determine the factor as the root and calculate the information value of the factor acquisition. Then, it is based on the highest gain value of each factor to determine the factor as the root. Next, it needs entropy value to determine the highest gain value. Meanwhile, the total value for each factor will be calculated to find the entropy of each case. The process of finding the value of each factor in the case is transformed into the following form in Table 2. It shows the transformation data from the five categories of attribute ratings to make it easier to remember the value of each attribute in the assessment process.

Table 2 Description of Factor Value

Description	
High	H
Currently	C
Low	L
Big	B
Small	S

Entropy (total) calculates the total value of the information on the low-risk level (S1), which has 8 cases. Meanwhile, the information on the high-risk level has 11 cases. Then, the total number of cases is 30 cases. Entropy (total) is calculated using Equation (2) as follows.

$$Entropy\ (\text{total}) = -\sum_{i=1}^{n} - Pi\ x\ Log2\ Pi$$

$$Entropy\,(total) = \left(-\frac{19}{30} * log_2\left(\frac{19}{30}\right)\right) + \left(-\frac{11}{30} * log_2\left(\frac{11}{30}\right)\right) = 0{,}94808$$

The next step is to calculate the entropy value for each factor used in analyzing credit risk. First, it is the credit growth factor. It is necessary to pay attention to Table 1 to calculate the entropy of credit growth. In Table 1, it can be seen that the credit growth with a high score is 8 cases. They are divided into a low credit risk level with 8 cases and a high credit risk level with 0 cases. For the number of cases with a moderate factor, it has 13 cases with a low credit risk level of 11 cases and a high credit risk level of 11 cases. Furthermore, for credit growth with low factor values, it has 9 cases with 0 cases of a low-risk level and 9 cases of high credit risk. Here is the entropy of credit growth.

$$Entropy(H) = \left(-\frac{8}{8} * log_2\left(\frac{8}{8}\right)\right) + \left(-\frac{0}{8} * log_2\left(\frac{0}{8}\right)\right) = 0$$

$$Entropy(C) = \left(-\frac{11}{13} * log_2\left(\frac{11}{13}\right)\right) + \left(-\frac{2}{13} * log_2\left(\frac{2}{13}\right)\right) = 0{,}61938$$

$$Entropy(L) = \left(-\frac{0}{9} * log_2\left(\frac{0}{9}\right)\right) + \left(-\frac{9}{9} * log_2\left(\frac{9}{9}\right)\right) = 0$$

From the results of the assessment, it can be seen that the entropy value of the credit growth attribute with two categories owned, namely the high and low categories. It gets an entropy value of 0 so that the search process is complete. It can be concluded that the credit growth attribute with a high level credit risk category is low and vice versa in the low category. Low level credit risk is in high category. Meanwhile, in the currently category, the entropy value obtained is 0,61938. Hence, further analysis is needed because the assessment process has not been completed, and the result is not 0.

The following calculation is for the net interest margin. The same action is taken for this factor. From the results of calculating the entropy value, it is known that the net interest margin has two categories, namely high and low. It produces a value that is not equal to 0 so that the calculation has not been completed. Then, further analysis will still be carried out in the entropy search process at the next root. Here is the calculation of entropy of the net interest margin.

$$Entropy\ (H) = \left(-\frac{13}{20} * log_2\left(\frac{13}{20}\right)\right) + \left(-\frac{7}{20} * log_2\left(\frac{7}{20}\right)\right) = 0{,}93407$$

$$Entropy(L) = \left(-\frac{6}{10} * log_2\left(\frac{6}{10}\right)\right) + \left(-\frac{4}{10} * log_2\left(\frac{4}{10}\right)\right) = 0{,}97095$$

The following calculation is bank type. From the results of calculating the entropy value, it is known that the bank type attribute with two categories (country and public) produces a value that is not equal to 0. So, the calculation has not been completed, and further analysis will still be carried out in the entropy search process at the next root. Here is the calculation of entropy of the bank type.

$$Entropy(C) = \left(-\frac{10}{14} * log_2\left(\frac{10}{14}\right)\right) + \left(-\frac{4}{14} * log_2\left(\frac{4}{14}\right)\right) = 0{,}86312$$

$$Entropy(P) = \left(-\frac{9}{16} * log_2\left(\frac{9}{16}\right)\right) + \left(-\frac{7}{16} * log_2\left(\frac{7}{16}\right)\right) = 0{,}9887$$

The next calculation is a capital ratio. From the results of calculating the entropy value, the capital ratio attribute with two categories (big and small) produces a value that is not equal to 0. Hence, the calculation has not been completed, and further analysis will still be carried out in the entropy search process at the next root. Here is the calculation of entropy of the capital ratio.

$$Entropy(B) = \left(-\frac{13}{19} * log_2\left(\frac{13}{19}\right)\right) + \left(-\frac{6}{19} * log_2\left(\frac{6}{19}\right)\right) = 0{,}89974$$

$$Entropy(S) = \left(-\frac{6}{11} * log_2\left(\frac{6}{11}\right)\right) + \left(-\frac{5}{11} * log_2\left(\frac{5}{11}\right)\right) = 0{,}99403$$

The next calculation is company size. From the results of calculating the entropy value, the company size attribute has two categories, namely big and small. It produces a value that is not equal to 0 so that the calculation has not been completed. Then, further analysis will still be carried out in the entropy search process at the next root. Here is the calculation of entropy of company size.

$$Entropy(B) = \left(-\frac{13}{19} * log_2\left(\frac{13}{19}\right)\right) + \left(-\frac{6}{19} * log_2\left(\frac{6}{19}\right)\right) = 0{,}89974$$

$$Entropy(S) = \left(-\frac{6}{11} * log_2\left(\frac{6}{11}\right)\right) + \left(-\frac{5}{11} * log_2\left(\frac{5}{11}\right)\right) = 0{,}99403$$

The next calculation is the bank compliance level. From the results of calculating the entropy value, the compliance level attribute with two categories (high and low) produces a value that is not equal to 0. So, the calculation has not been completed, and further analysis will still be carried out in the entropy search process at the next root. Here is the calculation of entropy of the bank compliance level.

$$Entropy(H) = \left(-\frac{15}{21} * log_2\left(\frac{15}{21}\right)\right) + \left(-\frac{6}{21} * log_2\left(\frac{6}{21}\right)\right) = 0{,}89974$$

$$Entropy(L) = \left(-\frac{4}{9} * log_2\left(\frac{4}{9}\right)\right) + \left(-\frac{5}{9} * log_2\left(\frac{5}{9}\right)\right) = 0{,}99108$$

Then, the researchers look for the gain value for each attribute. The first calculated gain value is the gain (total, credit growth). The calculation is done by adding up the entropy value for each category in credit growth. The gain value search determines the root. The result can be seen as follows.

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Credit Growth}_i|}{|Total|} * Entropy(\text{Credit Growth}_i)$$

$$= 1 - \left(\left(\frac{8}{30} * 0\right) + \left(\frac{13}{30} * 0{,}61938\right) + \left(\frac{9}{30} * 0\right)\right) = 0{,}67968$$

Next, the calculation of the gain value on the net interest margin (total, net interest margin) is carried out. The same steps are taken to calculate the gain value for this factor. It adds up the entropy values for each category on the net interest margin that has been analyzed in the previous stage. The result can be seen as follows.

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Net Interest Margin}_i|}{|Total|} * Entropy\ (Net\ Interest\ Margin_i)$$

$$= 1 - \left(\left(\frac{20}{30} * 0{,}93407\right) + \left(\frac{10}{30} * 0{,}97095\right)\right) = 0{,}00172$$

Then, the calculation of the gain value on the bank type (total, bank type) is also carried out. The same steps are taken to calculate the gain value for this factor by adding up the entropy values for each category on the bank type analyzed in the previous stage. The result can be seen as follows.

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Bank\ Type_i|}{|Total|} * Entropy(Bank\ Type_i)$$

$$= 1 - \left(\left(\frac{14}{30} * 0{,}86312\right) + \left(\frac{16}{30} * 0{,}9887\right)\right) = 0{,}01798$$

Next is the calculation of the gain value on the capital ratio (total, capital ratio). It also adds up the entropy values for each category on the capital ratio that has been analyzed in the previous stage. The result can be seen as follows.

$$= Entropy(S) - \sum_{1=1}^{n} \frac{(Capital\ Ratio_i)}{|Total|} * Entropy\ (Capital\ Ratio_i)$$

$$= 1 - \left(\left(\frac{19}{30} * 0{,}89974\right) + \left(\frac{11}{30} * 0{,}99403\right)\right) = 0{,}01376$$

Then, the calculation of the gain value on the company size (total, company size) will be carried out. The same steps are taken to calculate the gain value for this factor, namely by adding up the entropy values for each category on the company size that has been analyzed in the previous stage.

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Company\ Size_i|}{|Total|} * Entropy\ (Company\ Size_i)$$

$$= 1 - \left(\left(\frac{19}{30} * 0{,}89974\right) + \left(\frac{11}{30} * 0{,}99403\right)\right) = 0{,}01376$$

Then, the calculation of the gain value on the bank complience level (total, bank complience level) is carried out. It also adds up the entropy values for each category on the bank complience level that has been analyzed in the previous stage. The result can be seen as follows.

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Bank\ Complience\ Level_i|}{|Total|} * Entropy\ (Bank\ Complience\ Level_i)$$

$$= 1 - \left(\left(\frac{21}{30} * 0{,}86312\right) + \left(\frac{9}{30} * 0{,}99108\right)\right) = 0{,}04657$$

After the entropy value search process has been completed and all entropy values and gain values are known for each factor, the results of each entropy value for each factor category and the gain value for each factor are shown. It is easier to find out the comparison of the gain values obtained. The factor with the highest gain value will be the first root. The data are presented in Table 3.

In Table 3, all the factors in credit risk have obtained the calculation results of the entropy value and gain value. The factor with the highest gain value is credit growth, with a gain value of 0,67968. Hence, credit growth becomes the root. Based on the entropy value of each factor value, it is known that when credit growth has a high value, the credit risk has a low value. Meanwhile, when the credit growth has a low value, the credit ratio has a high value. The first node formed can be seen in Figure 2.

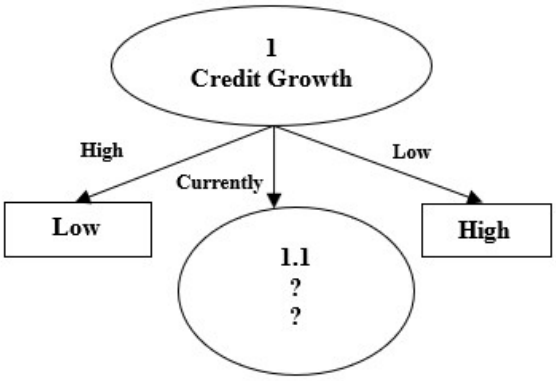

Figure 2 Decision Tree of Node 1

For the next node, it is done in the same way as in the first node search. The difference is that the number of credit growth factor cases with high and low values in the next node search is no longer counted in cases or eliminated. It is because they have been completed or have a value of 0. It can be said that information has

Table 3 Node 1

Node	Total Cases (S)		Low (S1)	High (S2)	Entropy	Info Gain
1	Total	30	19	11	0,94808	
	Credit Growth					0,67968
	High	8	8	0	0	
	Currently	13	11	2	0,61938	
	Low	9	0	9	0	
	Net Interest Margin					0,00172
	High	20	13	7	0,93407	
	Low	10	6	4	0,97095	
	Bank Type					0,01798
	Country	14	10	4	0,86312	
	Public	16	9	7	0,9887	
	Capital Ratio					0,01376
	Big	19	13	6	0,89974	
	Small	11	6	5	0,99403	
	Company Size					0,01376
	Big	19	13	6	0,89974	
	Small	11	6	5	0,99403	
	Bank Compliance Level					0,04657
	High	21	15	6	0,86312	
	Low	9	4	5	0,99108	

been obtained. So, for the following calculation, only 13 cases are counted for credit growth with a moderate value. The same thing is repeated until all values for each factor have been completed or are worth 0. In the research, the credit risk rules are obtained: credit growth as node 1, bank compliance level as node 2, net interest margin as node 3, and capital ratio as node 4.

Next, the research tests the data mining application with the RapidMiner. The data from 30 samples are shown in Table 1. The data are tested in the RapidMiner application by establishing a relationship between the database and the operators, as shown in Figure 3.

Figure 3 describes the process before classification with RapidMiner. In this section, the bank data database, as shown in Table 1, is imported into the worksheet to connect the tested database with the classification operator, namely the decision tree. It is necessary to set two operator roles and pull the decision tree operator into the worksheet. The process of connecting the tested dataset with the decision tree operator is carried out by pulling the port or wire, which can be seen in the figure that connects the dataset to the operator set role to decision tree operator.

The following rules are formed based on the graph formed from the RapidMiner test results in Figure 4. First, if credit growth is high, the credit risk is low. Second, the credit risk is low if credit growth is medium and the bank compliance level is high. Third, if credit growth is medium, the bank compliance level is low, and the net interest margin is low, credit risk is high. Fourth, if the credit growth is medium, the bank compliance level is low, the net interest margin is high, and the capital ratio is large, the credit risk is low. Fifth, if credit growth is medium, bank compliance level is low, net interest margin is high, capital ratio is small, and company size is large, credit risk is low. Sixth, if credit growth is medium, bank compliance level is low, net interest margin is high, capital ratio is small, and company size is small, credit risk is high. Last, if credit growth is low, the credit risk is high.

The research provides different results from previous studies, although there are similar variables in analyzing credit risk. For example, previous research mentions that bank size, leverage, bank age, and competing banks affect credit risk (Syamlan & Jannah, 2019). Another previous research states that financial inclusion on bank credit risk where an increase in the financial inclusion index will increase credit risk (Ghasarma, Muthia, Umrie, Sulastri, & Arianto, 2019). In the research, the variables that affect the credit risk of a bank are credit growth, bank compliance level, net interest margin, and capital ratio.

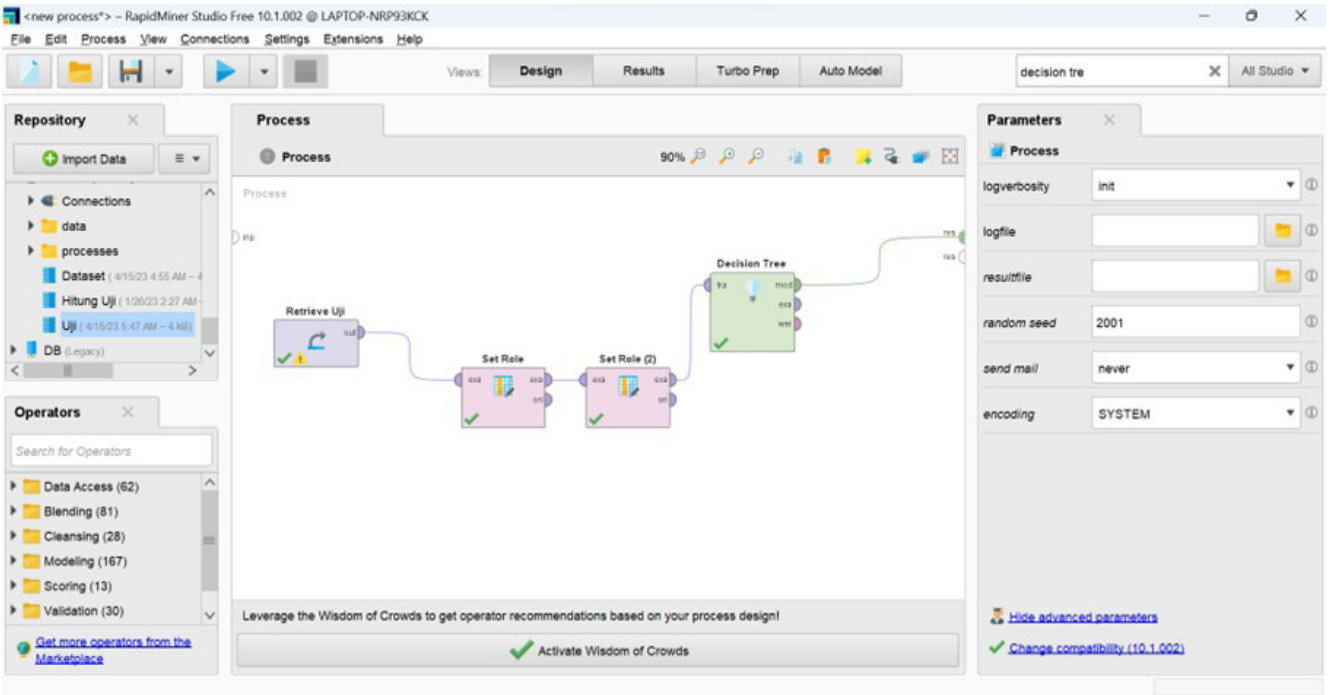

Figure 3 RapidMiner Design

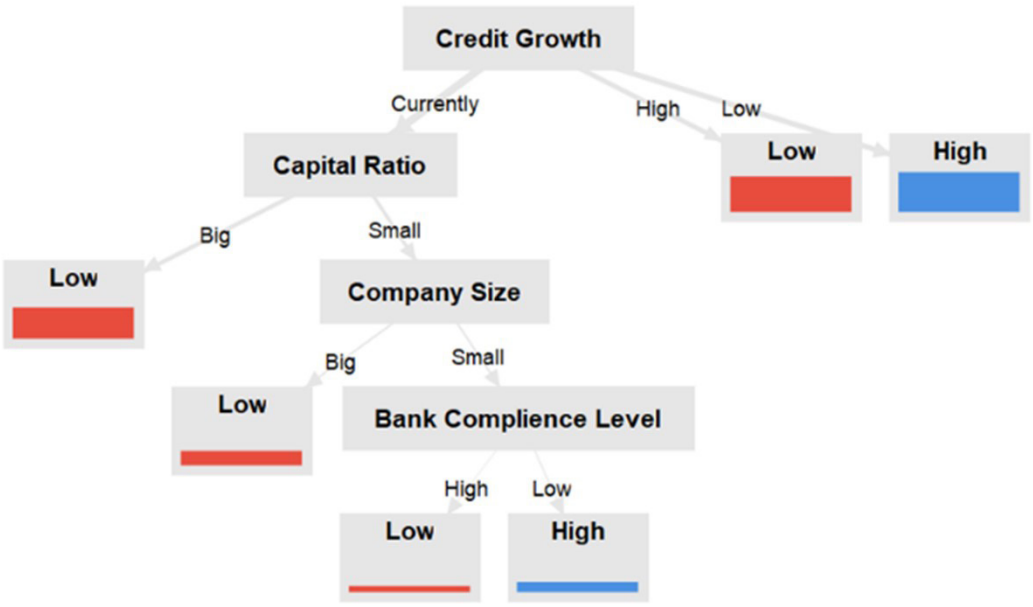

Figure 4 Graph of RapidMiner Test Results

IV. CONCLUSIONS

Based on the results, it can be concluded that the C4.5 algorithm can be applied to analyze credit risk with predetermined variables. Therefore, the research results can provide information for banks in minimizing the occurrence of credit risk and serve as a tool in policymaking in overcoming and reducing the number of occurred credit risks. Of the six variables used in analyzing credit risk, only four have an effect on the occurrence of credit risk, namely credit growth, bank compliance level, net interest margin, and capital ratio. Based on the results, it will be easier for banks to analyze the possibility of credit risk because the analysis only needs to be done on the empathy variable. Hence, banks can make policies to help to overcome credit risk problems.

The drawback of the research is that a wider variable that can affect credit risk is needed. It is hoped that further research can re-analyze the factors that may be credit risk factors other than those used in one of the world economic research factors. In addition, further research can use other approaches in classifying credit risk to gain knowledge about each algorithm used. Hence, credit service providers can choose an algorithm that suits their needs.

REFERENCES

Afrianto, E., Suseno, J. E., & Warsito, B. (2020). Decision tree method with C4.5 algorithm for students classification who is entitled to receive Indonesian Smart Card (KIP). In *IOP Conference Series: Materials Science and Engineering* (pp. 1–12). https://doi.org/10.1088/1757-899X/879/1/012072

Aji, I. K., & Manda, G. S. (2021). Pengaruh risiko kredit dan risiko likuiditas terhadap profitabilitas pada Bank BUMN. *JAD: Jurnal Riset Akuntansi & Keuangan Dewantara*, *4*(1), 36–45.

An, Y., & Zhou, H. (2022). Short term effect evaluation model of rural energy construction revitalization based on ID3 decision tree algorithm. *Energy Reports, 8*(July), 1004–1012. https://doi.org/10.1016/j.egyr.2022.01.239

Ariawan, P. A. (2019). Optimasi pengelompokan data pada metode k-means dengan analisis outlier. *Jurnal Nasional Teknologi & Sistem Informasi, 5*(2), 88–95. https://doi.org/10.25077/teknosi.v5i2.2019.88-95

Bedregal-Alpaca, N., Cornejo-Aparicio, V., Zarate-Valderrama, J., & Yanque-Churo, P. (2020). Classification models for determining types of academic risk and predicting dropout in university students. *International Journal of Advanced Computer Science and Applications, 11*(1), 266–272. https://doi.org/10.14569/ijacsa.2020.0110133

Cristina, K. M., & Artini, L. G. S. (2018). Pengaruh likuiditas, risiko kredit, dan dana pihak ketiga terhadap profitabilitas pada Bank Perkreditan Rakyat (BPR). *E-Jurnal Manajemen*, *7*(6), 3353–3383.

Fauzi, A., Marpaung, I. J. S., & Pardede, A. M. H. (2018). Sistem pendukung pemilihan pekerjaan menggunakan metode apriori berdasarkan korelasi jurusan dengan IPK untuk mengetahui pekerjaan yang tepat. *METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, *2*(2), 152–159.

Ghasarma, R., Muthia, F., Umrie, M. R., Sulastri, S., & Arianto, B. (2019). The influence of financial inclusion on credit risks in commercial banks in Indonesia. *Jurnal Akuntansi dan Bisnis*, *19*(2), 160–166.

Ginting, V. S., Kusrini, K., & Taufiq, E. (2020). Implementasi algoritma C4.5 untuk memprediksi keterlambatan pembayaran sumbangan pembangunan pendidikan sekolah menggunakan Python. *Inspiration: Jurnal Teknologi Informasi dan Komunikasi, 10*(1), 36–44. https://doi.org/10.35585/inspir.v10i1.2535

Harlina, S. (2018). Data mining pada penentuan kelayakan kredit menggunakan algoritma k-nn berbasis forward selection (Data mining on credit feasibility determination using K-NN algorithm based on forward selection). *CCIT Journal, 11*(2), 236–244. https://doi.org/10.33050/ccit.v11i2.591

Hakim, L., & Oktaria, E. T. (2018). Prinsip kehati-hatian pada lembaga perbankan dalam pemberian kredit. *Keadilan Progresif*, *9*(2), 164–176.

Hanif, R. A. (2015). Pengaruh pertumbuhan kredit, net interest margin, rasio modal dan ukuran perusahaan terhadap risiko kredit pada seluruh bank yang terdaftar di BEI periode 2010-2013. *PEKBIS*, *7*(3), 163–173.

Hozeng, S., & Aisa, S. (2016). Aplikasi data mining dengan menggunakan metode decision tree untuk prediksi penentuan resiko kredit. *Prosiding Seminar Ilmiah Sistem Informasi dan Teknologi Informasi, V*(2), 33–41.

Kurniawan, D., Anggrawan, A., & Hairani. (2020). Graduation prediction system on students using C4.5 algorithm. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 19*(2), 358–366. https://doi.org/10.30812/matrik.v19i2.685

Le, Y. (2022). Research on data resource management of biomass energy engineering based on data mining. *Energy Reports, 8*(July), 1482–1492. https://doi.org/10.1016/j.egyr.2022.02.048

Lihani, R., Ngadiman, & Hamidi, N. (2013). Analisis manajemen kredit guna meminimalkan risiko kredit (Studi pada PD BPR BKK Tasikmadu Karanganyar). *JuPE-Jurnal Pendidikan Ekonomi, 1*(3), 1–11.

Liu, Z., Jin, D., & Liu, Q. (2011). Prediction of water inrush through coal floors based on data mining classification technique. *Procedia Earth and Planetary Science, 3*, 166–174. https://doi.org/10.1016/j.proeps.2011.09.079

Pradeep, K. R., & Naveen, N. C. (2018). Lung cancer survivability prediction based on performance using classification techniques of Support Vector Machines, C4.5 and Naive Bayes algorithms for healthcare analytics. *Procedia Computer Science, 132,* 412–420. https://doi.org/10.1016/j.procs.2018.05.162

Ramos, D., Faria, P., Morais, A., & Vale, Z. (2022). Using decision tree to select forecasting algorithms in distinct electricity consumption context of an office building. *Energy Reports, 8*(June), 417–422. https://doi.org/10.1016/j.egyr.2022.01.046

Riandari, F., & Sihotang, H. T. (2020). Implementation of C4.5 algorithm to analyze library satisfaction visitors. *Jurnal Mantik*, *4*(2), 1076–1084.

Riandari, F., & Simangunsong, A. (2019). *Penerapan algoritma C4.5 untuk mengukur tingkat kepuasan mahasiswa*. CV. Rudang Mayang.

Saputro, A. R., Sarumpaet, S., & Prasetyo, T. J. (2019). Analisa pengaruh pertumbuhan kredit, jenis kredit, tingkat bunga pinjaman bank dan inflasi terhadap kredit bermasalah. *Ekspansi: Jurnal Ekonomi, Keuangan, Perbankan, dan Akuntansi*, *11*(1), 1–12.

Subarkah, P., Pambudi, E. P., & Hidayah, S. O. N. (2020). Perbandingan metode klasifikasi data mining untuk nasabah bank telemarketing. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 20*(1), 139–148. https://doi.org/10.30812/matrik.v20i1.826

Susanto, W., & Indriyani, L. (2019). Analisis penerapan Naïve Bayes untuk memprediksi resiko kredit anggota koperasi keluarga guru. *Jurnal Informatika, 6*(2), 262–270. https://doi.org/10.31311/ji.v6i2.5724

Syamlan, Y. T., & Jannah, W. (2019). The determinant of credit risk in Indonesian islamic commercial banks. *Share: Jurnal Ekonomi dan Keuangan Islam*, *8*(2), 181–206. https://doi.org/10.22373/share.v8i2.5051

Wang, H. B., & Gao, Y. J. (2021). Research on C4.5 algorithm improvement strategy based on MapReduce. *Procedia Computer Science, 183,* 160–165. https://doi.org/10.1016/j.procs.2021.02.045

Wang, X., Zhou, C., & Xu, X. (2019). Application of C4.5 decision tree for scholarship evaluations. *Procedia Computer Science, 151,* 179–184. https://doi.org/10.1016/j.procs.2019.04.027

Wijaya, E., & Tiyas, A. W. (2019). Analisis pengaruh kecukupan modal, likuditas, risiko kredit dan efisiensi biaya terhadap profitabilitas bank umum. *Jurnal Ekonomi, Manajemen dan Perbankan (Journal of Economics, Management and Banking)*, *2*(3), 99–109.

Zulfami, F. (2017). Analisa dan perancangan aplikasi data mining penentuan resiko kredit kepemilikan kendaraan bermotor menggunakan algoritma K-Nearest Neighbor. *Jurnal Inkofar, 1*(1), 32–39. https://doi.org/10.46846/jurnalinkofar.v1i1.1