IMPLEMENTATION OF DATA MINING ALGORITHM C4.5 TO PREDICT LOAN PAYMENTS IN THE HARUM MANIS WOMEN'S UNION IN SIRNOBOYO VILLAGE

[*]MIFTAHUL MUKTI ANAS

Department of Informatics, Faculty of Engineering, Universitas Muhammadiyah Gresik

Jl. Sumatra No.101 GKB Gresik

e-mail: mukty.anas@gmail.com

[*]Corresponding author

ABSTRACT

The women's union, "Harum Manis" is an active savings and loan cooperative that uses members' funds in savings and loans. Given the large number of prospective members who register each year, the union still needs to be more selective in accepting prospective members who only see from work and salary, thus causing lousy credit. To reduce the occurrence of bad loans, predicting prospective members' smooth payment status and finding prospective members, including bad credit or current loans, is necessary. This research applies classification data mining techniques using the Decision Tree C4.5 method to determine the smooth payment class, which is a jam class or a smooth class. The attributes used in this study consist of four variables, namely age, marital status, income, and home status. System testing is done three times testing. The data were taken from 102 data for the "Harum Manis" Women's Union Member Loan data. Based on the test results, it was found that the first test produced the highest accuracy, reaching 64%.

Keywords: *Data Mining, Classification, Decision Tree, Loan Payment*

1. INTRODUCTION

In general, the development of savings and loan businesses has become the primary basis in Indonesia. Savings and loan union have even become the leading destination for micro business actors in looking for friendly and easy-to-reach financing without complicated conditions based on family principles [1]. Savings and loan union are not only required to increase the profitability and welfare of members but also must maintain business continuity [2]. The "Harum Manis" Women's Union is a type of active savings and loan cooperative, utilizing funds from members in the form of savings and loans. The union still needs to be more selective in accepting prospective members who only look at job and salary aspects, thus causing lousy credit. This problem can affect the realization of loans or the circulation of money to other members.

Several studies have been carried out previously to resolve the problem of alleged slow payments, such as predicting bottlenecks in bank credit payments using Naïve Bayes and K-Nearest Neighbor with an accuracy of up to 80% [3], predicting the smoothness of bank credit payments based on forward selection up to 71% [4], Credit credit classification using decision trees with an accuracy of up to 96% [5]–[7], using a Support Vector Machine reaching up to 84% accuracy [8], using Naïve Bayes based on length of business and loan amount [9], and determining customer creditworthiness with an accuracy of up to 81% [10]. The solution can also use clustering, such as using K-Means [11] with data based on brand and place of residence [12].

From the previous research, it is necessary to design an application to predict the smooth payment status of union members, which will determine whether a credit is bad or good. It solves problems and reduces the occurrence of bad credit. The prediction results are hoped to allow cooperatives to provide early treatment for these problems.

2. RESEARCH METHODOLOGY

2.1 System Design

System Analysis is the decomposition of a complete system into its parts, which aims to identify and analyze problems, opportunities, obstacles that occur, and expected needs so that improvements can be proposed. Based on observations in the field, in providing loan funds. The "Harum Manis" Women's Cooperative only looks at the salary aspect to determine the smoothness of payments. Meanwhile, we only sometimes know someone's needs so that mistakes can occur. This problem can affect the realization of loans or the circulation of money to other members. The results of the predictions to determine whether loan payments are smooth or sour can later be used as a reference for providing loan funds to prospective cooperative members.

The prediction process is carried out by applying data mining techniques using the Decision Tree C4.5 method. This technique uses data on 102 union members obtained by the research of the "Harum Manis" women's union in Sirnoboyo village using the variables house status, marital status, age, and income. The built prediction process will produce informative output data in the form of predicted results in smooth or bad payments, which will be used as a consideration for providing loan funds. Using the Decision Tree C4.5 method, the system that will be created will be able to predict the cooperative in determining prospective members of the savings and loan cooperative.

Decision Trees have the advantage of being able to convert substantial facts into decision trees that represent rules. Rules can be easily understood with natural language. Decision Tree has several algorithms, one of which is C4.5. The C4.5 algorithm is the most popular compared to other algorithms in the Decision Tree group; besides that, the C4.5 algorithm has an acceptable level of accuracy. Apart from handling categorical type attributes, this algorithm can also handle numeric type attributes. Figure 1 will explain the flow of the decision support system for granting eligibility for credit applications using the Decision Tree C4.5 method. Algorithm C4.5 refers to Figure 1.

The C4.5 Decision Tree algorithm is explained as follows:

1. First, enter the training data stored in the database.
2. Are the attributes of the training data of continuous (numeric) type?
3. If the attributes of the training data are continuous, determine the V position; then, calculate the gain for each V position.
4. If the attributes of the training data are categorical / not continuous, then calculate the Gain for each attribute.
5. The results of continuous and categorical gain calculations determine the branch/attribute with the highest gain.
6. Division of cases into branches of selected attributes.
7. Does each branch have the same class?
8. If each branch has a different class, the calculation returns to point 2.
9. If each branch has the same class, it results in a decision tree's formation.
10. Next, enter the test data.
11. The system classifies test data using a decision tree that has been formed.
12. The system outputs the classification results.

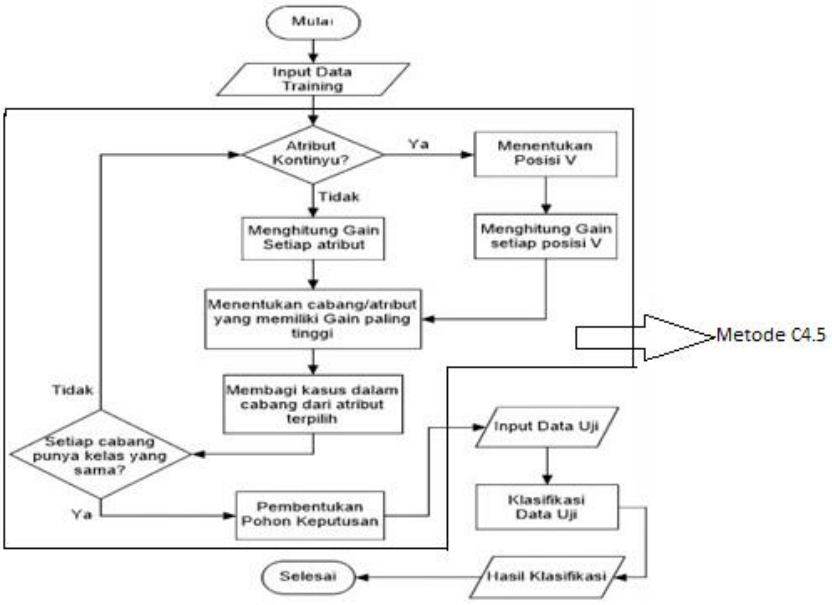

Figure 1. System Flowchart

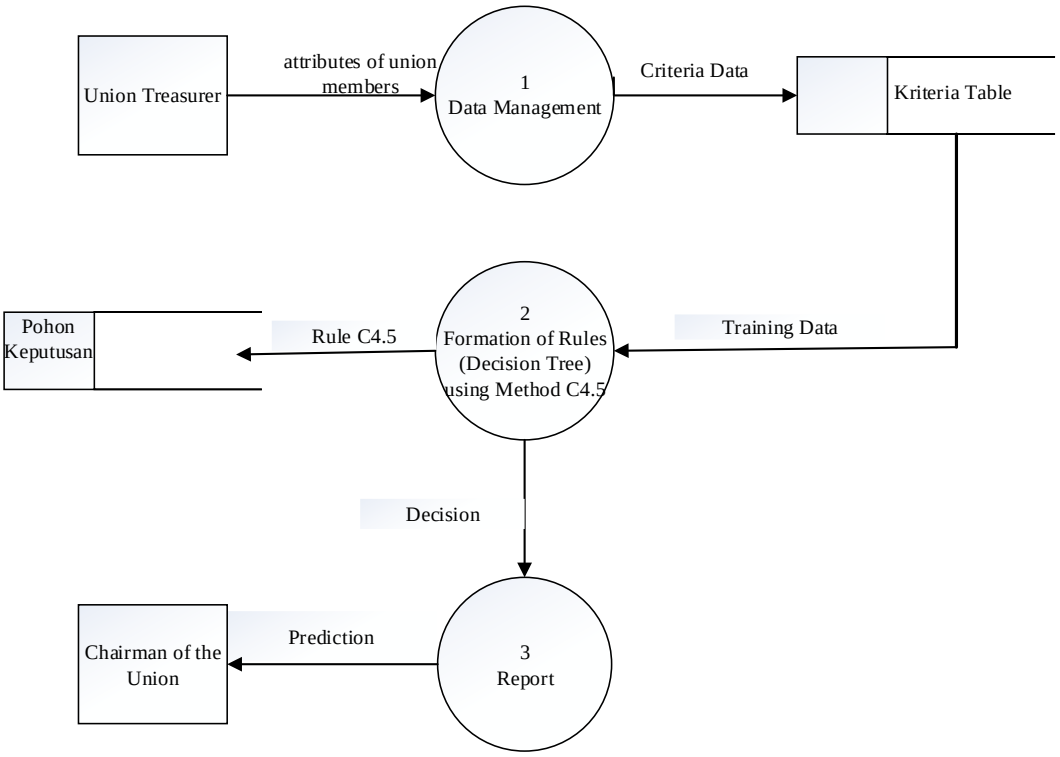

Figure 2. DFD Level 0

2.2 Context Diagram

Data Flow Diagram (DFD) level 0 in Figure 2 explains the data flow in the system. There are three processes in the system. The first process is managing member attribute data, which is input by the Cooperative treasurer. The data on cooperative members' skills is saved in the database, which will become training data for forming a decision tree. The second process is the formation of rules (decision trees) that will be used in the prediction process of test data, the results of which are stored in the decision tree database. In the third process, the report will be given to the cooperative's chairman.

2.3 Decision Tree with C4.5 Algorithm

Decision tree is a data mining classification method. A decision tree, in learning terms, is a tree structure where each tree node represents the attribute that has been tested. Each branch distributes test results, and the leaf nodes represent certain class groups. The top node level of a Decision Tree is the root node, which is usually the attribute with the greatest influence on a particular class. In general, Decision Trees carry out a top-down search strategy for solutions. In classifying unknown data, attribute values will be tested by tracing the path from the root node to the final node (leaf), and then the class belonging to a particular new data will be predicted. Quinlan introduced the C4.5 algorithm in 1996 as an improved version of ID3. In ID3, decision tree induction can only be done on categorical type features (nominal or ordinal), while numeric types (interval or ratio) cannot be used.

What is essential in decision tree induction is how to state the testing conditions for the nodes. There are three important groups in node testing requirements:

1. Binary features
   It is a feature that only has two different values. The test conditions when this feature becomes a node (root or interval) only have two branch options.
2. Categorical features
   Features whose values are of the categorical type (nominal or ordinal) can have several different values. In general, there are two solutions, namely binary splitting and multi-splitting.
3. Numerical features
   For numeric type features, testing conditions in nodes (root or internal) are expressed by comparison testing ($A \leq V$) or ($A > V$) with binary results.

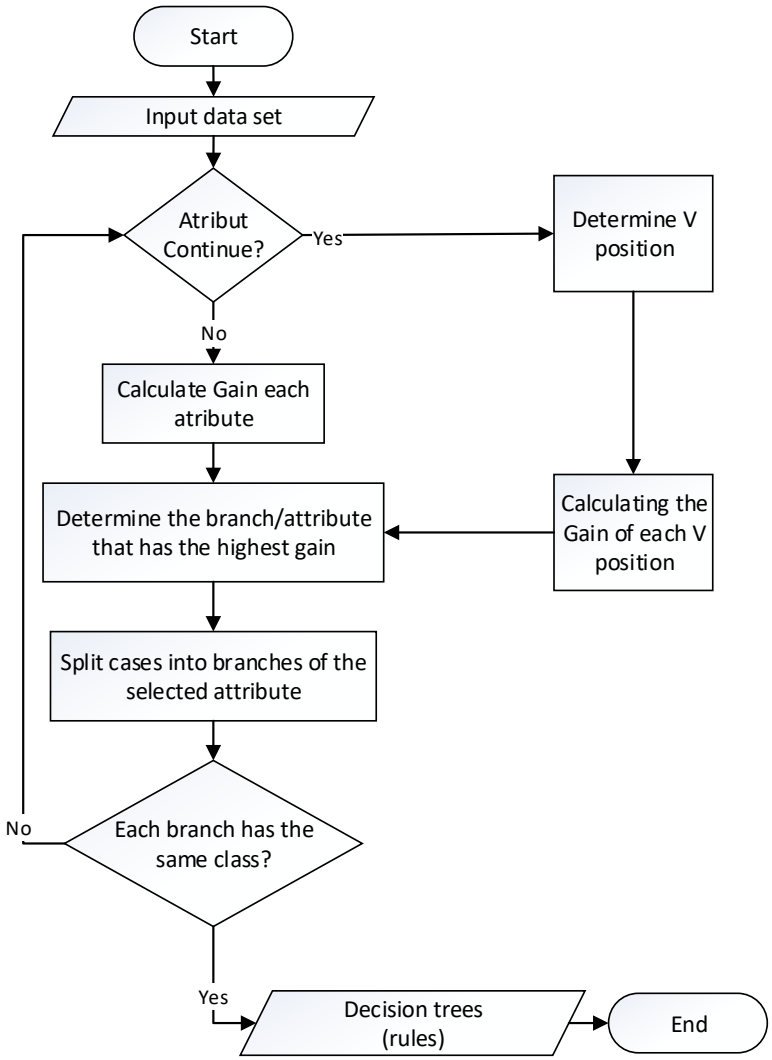

Figure 3. Flowchart algoritma Decision Tree C4.5

In general, the C4.5 algorithm for building a decision tree is as follows, as presented in Figure 3:

1. Select the attribute as root.
2. Create a branch for each value.
3. Divide cases within branches.
4. Repeat the process for each branch until all cases in the branch have the same class.

2.4 System Testing Scenario

Before creating a system for predicting the smooth payment of loans for cooperative members using data mining techniques with the C4.5 decision tree method, it is necessary to carry out several system testing scenarios first, such as validity in Excel so that the system can run according to the purpose for which it was created. Four criteria are used in the test: marital status, housing status, income, and age. The data used for system testing results from data classification from the Sweet Fragrance Women's Union in Sirnoboyo village, Benjeng sub-district, Gresik district. It is hoped that the system created can produce a prediction system that can provide helpful information for the Sirnoboyo village women's union in determining the predicted results of payments that will get a loan.

1. First testing
   The first test was carried out using 102 data, with 52 training data and 50 test data.
2. Second testing
   The second test was carried out using 102 data, with 59 training data and 43 test data.
3. Third testing
   The third test was carried out using 102 data, with 75 training data and 27 test data.

Table 1. Classification results

Test	Splitting data (training:test)	Accuracy	Error	Correct result	False result
1	52 :50	64%	36%	32	18
2	59:43	62.79%	37.21%	27	16
3	75:27	59.26%	40.74%	16	11

Each test will compare the original results with the results from the application using the accuracy formula as follows.

$$Accuracy = \frac{\sum true\ prediction}{\sum test\ data}\ X\ 100\% \quad (1)$$

3. RESULTS AND DISCUSSIONS

The accuracy, error rate, values from the results of trials that have been carried out three times are presented in Table 1. The prediction (classification) system can not work 100% correctly, so this section will evaluate the results of the prediction calculations for each experiment using the Confusion Matrix. This evaluation uses the Confusion Matrix, a table used to determine the performance of a classification model.

In 3 tests with different data splitting, it was found that the best accuracy was achieved when splitting 52 and 50 for training and testing data, respectively, namely 64%. This splitting of approximately half the data is sufficient to obtain the best classification performance. The higher the amount of training data, the better the performance accuracy. This problem can be seen in splitting data 75 and 27. These results are enough to recommend that determining the creditworthiness of members of the "Harum Manis" cooperative can be assisted using the Decision Tree classification.

4. CONCLUSION

The results obtained from this research show that this system can produce information to predict the smooth payment status of prospective cooperative members using the Decision Tree C4.5 method with an accuracy of up to 64%. From the results of the first, second, and third experiments, the first experiment, which got the best accuracy value with an accuracy value of 64%, was made into a decision tree. The training data's composition will affect the accuracy level obtained, where each experiment that has been carried out will obtain different accuracy and form the best decision tree.

Before carrying out the core process or forming a decision tree, it is best to explore (select) data to obtain quality data because good decisions must also be based on quality data. This research is only based on 102 data on the "Harum Manis" women's union members from 2010 to 2016. In future research, using more than 102 member data would be better. So this application can make more accurate predictions.

REFERENCES

[1] N. Asmita, (2020), "Peran Koperasi Simpan Pinjam dan Pembiayaan Syariah (KSPPS) dalam Pemberdayaan Ekonomi Masyarakat (Studi pada BMT Al-Ittihad Rumbai Pekanbaru," *Jurnal An-Nahl*, vol. 7, no. 2, pp. 171–176, doi:10.54576/annahl.v6i2.49.

[2] M. S. Soumokil, F. Edoway, and A. Numberi, (2022), "Analisa Sistem Pemberian Dan Pengawasan Kredit Serta Faktor Penyebab Terjadinya Kredit Macet Pada Bank Papua Cabang Timika," vol. 6, no. 1, pp. 21–34.

[3] F. Gultom and T. Simanjuntak, (2021), "Prediksi Tingkat Kelancaran Pembayaran Kredit Bank Dengan Menggunakan Algoritma Naïve Bayes Dan K-Nearest Neighbor," *METHOMIKA Jurnal Manajemen Informatika dan Komputerisasi Akuntansi*, vol. 4, no. 2, pp. 98–102, doi:10.46880/jmika.vol4no2.pp98-102.

[4] M. Hasan, (2017), "Prediksi Tingkat Kelancaran Pembayaran Kredit Bank Menggunakan Algoritma Naïve Bayes Berbasis Forward Selection," *ILKOM Jurnal Ilmiah*, vol. 9, no. 3, pp. 317–324, doi:10.33096/ilkom.v9i3.163.317-324.

[5] L. Desyanita and A. Wibowo, (2020), "Pemodelan Sistem Prediksi Kelayakan Pengajuan Kredit," *Elkom Elektronika Dan Komputer*, vol. 13, no. 2, pp. 10–22.

[6] T. Hidayatulloh, A. Fajria, R. N. Lestari, and N. S. Z. Nufus, (2022), "Algoritma C4.5 Untuk Menentukan Kelayakan Pemberian Kredit (Studi kasus: Bank Mandiri Taspen Kantor Kas Sukabumi)," *Jurnal Larik: Ladang Artikel Ilmu Komputer*, vol. 2, no. 2, pp. 66–74, doi:10.31294/larik.v2i2.1836.

[7] E. Wijaya, F. A. Tarigan, and Michael, (2021), "Aplikasi Prediksi Penentuan Kelancaran Pembayaran Koperasi Dengan Algoritma C5.0," *Jurnal Times: Technology Informatics & Computer System*, vol. 10, no. 1, pp. 31–38.

[8] I. P. Casuarina, M. N. Hayati, and S. Prangga, (2022), "Klasifikasi Status Pembayaran Kredit Barang

Elektronik dan Furniture Menggunakan Support Vector Machine," *Eksponensial*, vol. 13, no. 1, p. 71, doi:10.30872/eksponensial.v13i1.887.

[9] Nurdina Rasjid, Nurhikmah Arifin, and Nilam Cahya, (2021), "Klasifikasi Nasabah Bank Layak Kredit Menggunakan Metode Naive Bayes," *Jurnal ilmiah Sistem Informasi dan Ilmu Komputer*, vol. 1, no. 1, pp. 01–10, doi:10.55606/juisik.v2i2.187.

[10] F. M. Akbar, F. A. Bachtiar, and W. Purnomo, (2020), "Klasifikasi Kredit Macet berdasarkan Profil Nasabah pada Koperasi Serba Usaha Surya Abadi menggunakan Algoritme C5. 0," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 9, pp. 3047–3056.

[11] M. N. R. Fitriani, B. Priyatna, B. Huda, A. L. Hananto, and T. Tukino, (2023), "Implementasi Metode K-Means Untuk Memprediksi Status Kredit Macet," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 3, p. 554, doi:10.30865/json.v4i3.5953.

[12] J. B. Sembiring, H. Manurung, and A. Sihombing, (2023), "Pengelompokan Data Tunggakan Pembayaran Kredit Mobil Menggunakan Metode Clustering (Studi Kasus: Cv Citra Kencana Mobil)," *Jurnal Manajemen Informatika Jayakarta*, vol. 3, no. July, pp. 275–291.