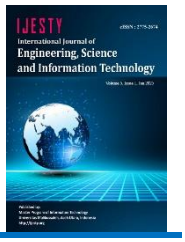
International Journal of Engineering, Science and Information Technology
Volume 5 No. 1 (2025) pp. 368-372
ISSN 2775-2674 (online)
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i1.752
Research Paper, Short Communication, Review, Technical Paper



Performance of K-Nearest Neighbor Algorithm and C4.5 Algorithm in Classifying Citizens Eligible to Receive Direct Cash Assistance in Bandar Mahligai Village

Wan Amalia Chaliza Nur *, Dahlan Abdullah, Rini Meiyanti

Department of Informatics, Faculty of Engineering, Universitas Malikussaleh, Aceh, Indonesia

**Corresponding author E-mail: wan.190170137@mhs.unimal.ac.id*

The manuscript was received on 18 May 2024, revised on 28 August 2024, and accepted on 12 January 2025, date of publication 21 January 2025

Abstract

Direct Cash Assistance, commonly called BLT, is one of the many programs the Indonesian government held to reduce the poverty rate of the Indonesian population. This study compares the KNN and C4.5 methods to determine the eligibility of residents eligible to receive Direct Cash Assistance in Bandar Mahligai Village. This study began with collecting resident data from the Bandar Mahligai village office. Then, the data obtained was taken into several attributes to be used in the classification process, namely the name of the head of the family, KK number, NIK, number of dependents, occupation, income, and monthly expenses. After the data is collected, the data will be classified using the KNN and C4.5 algorithms. There is a significant difference between the two algorithms in the classification process; the KNN algorithm by looking for the nearest neighbor data value, in this study, the K value = 9, while the C4.5 algorithm by building a decision tree from the attribute values taken based on resident data used as training data. The classification results of the two methods will be compared using a confusion matrix to obtain a higher accuracy technique. The results of testing using a confusion matrix for both algorithms are the accuracy produced by the KNN and C4.5 algorithms in classifying residents eligible for Direct Cash Assistance (BLT) of 90% in the system that has been built. The results of comparing the KNN and C4.5 algorithms for this study show that the KNN algorithm is better because the accuracy level reaches 90% in manual and system calculations. While the C4.5 method only gets 85% for the accuracy of its manual calculations, it receives an accuracy level of 90% in the system that has been built.

***Keywords**: KNN, Data Mining, C4.5, Algorithm, Classification.*

1. Introduction

Direct Cash Assistance, commonly referred to as BLT is one of the Indonesian government programs that help reduce the poverty rate of the Indonesian population. Aceh province follows government regulations in distributing the Direct Cash Assistance program to the community. Bandar Mahligai Village, Sekerak District, Aceh Tamiang Regency is a village that implements this Direct Cash Assistance. Every month, people who are categorized as underprivileged receive IDR 300,000. Based on the results of observations by researchers in Bandar Mahligai Village, the poverty rate is relatively high. Unfortunately, selecting recipients of Direct Cash Assistance in Bandar Mahligai Village is still manual, and staff must collect data files to choose prospective residents who receive Direct Cash Assistance, so it takes more time and drains more energy to make decisions. To Pramadhani and Setiadi (2014), data mining is a process of selection, exploration, and modeling from a large amount of data to find patterns or tendencies that are usually not realized [1].

K-Nearest Neighbor has a weakness in choosing the perfect K value to obtain the best system accuracy but has advantages in training data sets that have a lot of noise and is effective on high amounts of training data [2]; this method aims to classify new objects according to attributes and training samples. Given a query point, Next, several K objects or training points that are closest to the query point [3]. Like the K-Nearest Neighbor (KNN) algorithm, which has shortcomings, The C4.5 algorithm also has several weaknesses, one of which is that the quality of the decisions obtained highly depends on how the decision tree is made. So, if the decision tree that is designed is less than optimal, the quality of the decisions obtained will be affected [4].

Therefore, the researcher chose the K-Nearest Neighbor algorithm and the C4.5 algorithm for comparison between the two algorithms in shortening the time and level of accuracy of the results of determining prospective recipients of Direct Cash Assistance, as well as

creating a system to assist staff tasks in Bandar Mahligai Village, then implementing one of the more effective methods for use by village office staff, where several criteria will be processed to get a value. The value obtained will be compared with the training information, resulting in a classification of citizen information eligible to receive Direct Cash Assistance and a comparison of the level of accuracy of the two methods studied by the author. The results of this study are a web-based application that functions to classify the determination of citizens who are eligible to receive Direct Cash Assistance. This application can be beneficial for staff in Bandar Mahligai Village to shorten their performance time in the classification of the selection of Direct Cash Assistance recipients so that they can avoid errors that occur and more accurate results because of the comparison between the two algorithms applied by the author in this study [5] [6] [7].

2. Literature Review

2.1. Data Mining

According to Pramadhani and Setiadi (2014), data mining is selecting, exploring, and modeling large amounts of data to find patterns or similarities that are generally not realized. Data mining is digging for information in data until knowledge is obtained. It will be very valuable because it can be used as a basis for decision-making. Nowadays, companies all over the world use data mining techniques to support their business decisions so that they gain large business profits [8] [9] [10].

2.2. Classification

Description or classification means a model in data mining where a classifier is constructed to predict label categories, for example, "safe" or "risky" in money lending application data, "yes" or "no" in marketing data, or "treatment A," "treatment B," and "treatment C" in medical data. The group can be presented using values that suit their needs; the setting of these values does not have a special meaning.

Classification is also a process of finding a model that explains or differentiates concepts or data classes, aiming to estimate the class that comes from an object whose class is unknown. In classification, several records are received, called training sets, which consist of several attributes. The attributes can be continuous or categorical; one attribute gives a class or record.

The benefit of classification is to identify characteristics that indicate the group to which each case belongs. This pattern can be used to understand existing data and predict new cases. Classification models are created by data mining by examining data that has been classified (cases) and inductively finding predictive patterns [11] [12].

2.3. K-Nearest Neighbor Algorithm

K-Nearest Neighbor is a new object classification algorithm based on attributes and training data. Where the latest test data obtained is classified according to the majority in the K-Nearest Neighbor category, this algorithm uses neighborhood classification to estimate new test data. The K-Nearest Neighbor algorithm can be defined (Rahmawati, 2020) with the following meaning:

1. Calculate the distance between the test data and the existing training data. One similarity in calculating proximity distance can be using cosine similarity.
2. Set the parameter value k = number of nearest neighbors.
3. Sorting the smallest distance on test data.
4. Match types based on suitability.
5. Find the most significant number of nearest neighbors. Then, determine the type.
6. The distance used in this research is cosine similarity [13] [14].

Case model, if you want to find efforts to solve new patient problems using old patients' efforts. In finding efforts from new patients, old patient cases were used. Which of the old cases has an approach to the new case? Here is the formula for calculating the closeness between two cases:

$$Similarity\ (T,S) = \frac{\sum_{i}^{n}= 1\ \int(T_{i,}\ s_{i}) * W_{i}}{W_{i}} \quad \text{.....(1)}$$

Description:

T = New case
S = Case in storage
n = number of attributes in each case
i = Individual attribute between 1 and n
= similarity function of attribute I between case T and case S
W = weight given to the i attribute

The general approach is to have a value between 0 and 1. 0 means the two cases are different, whereas the value of 1 case is not different.

2.4. C4.5 Algorithm

The C4.5 algorithm was introduced by J. Ross Quinlan, a developer of the ID3 algorithm; this algorithm is used to create decision trees. Decision trees are considered to be one of the most well-known approaches. Decision tree classification is formed from a node in the form of a root; the root node has no input [15].

Initially, decision trees came from many facts and represented rules that could be obtained through the use of an algorithm called the C4.5 algorithm. The purpose of creating this decision tree is to make problems easy to resolve.

In its steps, the C4.5 algorithm has two working rules: building a decision tree and establishing rules (rule model). The rules built on a decision tree will form a situation as if then [16].

There are four stages in the process of creating a decision tree in the C4.5 algorithm, as follows:

1. Determine the attribute as the root and adjust it to the highest gain value for the attributes.
2. Building a branch for each value means building a branch based on the highest number of gain variable values.
3. Divide each case into branches according to the calculation of the highest gain value. The calculation is carried out after calculating the first highest gain value and then calculating the highest gain again without including the value of the initial gain variable.

4. Repeat the stages in each branch so that all cases in the branch have the same class, and repeat all stages of calculating the highest gain for each branch until no more calculation stages can be carried out [17] [18].

To calculate the gain, the formula stated in the following equation is used:

$$Gain(\mathrm{S},\mathrm{A}) = Entropy(\mathrm{S}) - \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{|\mathrm{Si}|}{|\mathrm{S}|} \times Entropy\ (\mathrm{Si}) \quad \text{..... (2)}$$

Description:
S = set of cases
A = Atribut/feature
N = number of attribute A partitions
|Si| = number of instances in the I partition
|S| = number of cases in S

The calculation of the entropy value can be understood in the equation below, as follows[19]:

$$Entropy\ (S) = \sum_{i=1}^{n} - pi\ \times \log_2 pi \quad \text{..... (3)}$$

Description:
S = set of cases
A = feature
N = number of S partition
pi = proportion from Si to S

3. Research Method

Data collection in the study began with collecting data, then the data obtained will be separated into 2, namely training data and testing data. Data was taken directly from the Bandar Mahligai village office, Sekerak District, Aceh Tamiang Regency, by asking for help from one of the village office staff on duty. Then, the data was arranged according to the author's requirements. The data used to train the system in the study is called training data. The purpose of this training data itself is to train the K-Nearest Neighbor algorithm and the C4.5 algorithm so that they can classify residents who are eligible for Direct Cash Assistance (BLT) in Bandar Mahligai Village, Sekerak District, Aceh Tamiang Regency. After testing the K-Nearest Neighbor algorithm and the C4.5 algorithm trained using training data, the next stage is to assess the performance of the two algorithms. The assessment will be carried out using test data. The function of the test data is to test the K-Nearest Neighbor algorithm and the C4.5 algorithm by entering new data. The K-Nearest Neighbor and C4.5 algorithms correctly classify the newly entered data and determine whether or not the citizen data entered is eligible for Direct Cash Assistance (BLT). After the test data classification results have come out from each algorithm, the K-Nearest Neighbor algorithm and the C4.5 algorithm will be tested regarding the level of accuracy produced by both algorithms using the matrix coefficient formula. This effort aims to see which of the two algorithms has a higher level of accuracy.

4. Results and Discussion

This section describes the research results and further explanations regarding the research. More details will be explained in the next sub-chapter.

4.1. C4.5 Algorithm Calculation

The discussion includes calculating the C4.5 algorithm for determining rules or implementing classification on direct cash assist recipient data. Group each variable to obtain attributes that will be used to calculate entropy based on cases to get roots and branches for the BLT recipient classification process.

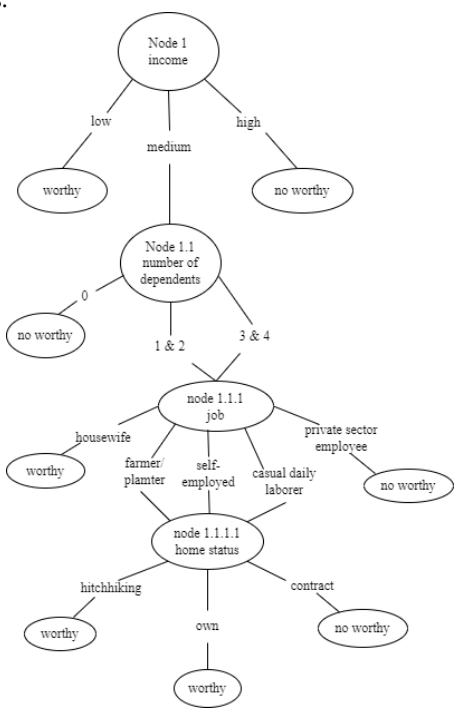


Fig 1. Final Result Decision Tree

Based on the final decision tree results, the following *general rules are obtained:*

1. If income = low, then feasible.
2. If income = high, then it is not feasible.
3. If income = moderate:
   a. If the number of dependents = 0, then it is not eligible.
   b. If the number of dependents = 1 or 2, or 3 or 4:
      - If job = housewife, then it is feasible.
      - If the job = is a private team member, it is not eligible.
      - If occupation = self-employed, farmer/planter, or casual laborer:
        - If the house status = freeloading, then it is feasible.
        - If the house status = owned, then it is eligible.
        - If the house status = contract, then it is not feasible.

After implementing the method in the information system, a more detailed decision tree is obtained for the classification process. The general rules received are also slightly different from those calculated manually. So now apply it to the test data.
After completing the test data classification process, the next step is to find the level of accuracy of the C4.5 method for this study using a confusion matrix. The calculation process is described as follows [20] [21].

Table 1. Confusion Matrix for C4.5 Method

	Prediction: Decent	Prediction: Not Worthy
Current: Eligible	11 (TP)	1 (FN)
Current: Not Eligible	1(FP)	7 (TN)

1. Precision $= \frac{TP}{TP+FP} = \frac{11}{11+1} = \frac{11}{12} = 0{,}9167 \times 100 = 91{,}67\%$
2. Recall$= \frac{TP}{TP+FN} = \frac{11}{11+1} = \frac{11}{12} = 0{,}9167 \times 100 = 91{,}67\%$
3. F1-Score$= 2 \times \frac{0{,}9167 \times 0{,}9167}{0{,}9167+0{,}9167} = 2 \times 0{,}4585 = 0{,}9167 \times 100 = 91{,}67\%$
4. Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{11+7}{11+7+1+1} = \frac{18}{20} = 0{,}9 \times 100 = 90\%$

From the *confusion matrix calculation* above, it is known that the accuracy level of the *C.45 method* for the tested data is 90%.

4.2. KNN Algorithm Calculation

The initial stage of the K-Nearest Neighbor method is to calculate the data distance using the Euclidean distance formula. Here is the Euclidean distance formula.

$$ed = \sqrt{(x_i - y_i)^2 + (x_j - y_j)^2} \quad \text{..... (4)}$$

The calculation starts from the first data in the test table. After completing the test data classification process, the next step is to find the level of accuracy of the K-Nearest Neighbor method for this study using *a confusion matrix.* The calculation process is described as follows.

Table 2. Confusion Matrix for KNN Method

	Prediction: Decent	Prediction: Not Worthy
Current: Eligible	11 (TP)	1 (FN)
Current: Not Eligible	1 (FP)	7 (TN)

1. Precision $= \frac{TP}{TP+FP} = \frac{11}{11+1} = \frac{11}{12} = 0{,}9167 \times 100 = 91{,}67\%$
2. Recall$= \frac{TP}{TP+FN} = \frac{11}{11+1} = \frac{11}{12} = 0{,}9167 \times 100 = 91{,}67\%$
3. F1-Score$= 2 \times \frac{0{,}9167 \times 0{,}9167}{0{,}9167+0{,}9167} = 2 \times 0{,}4585 = 0{,}9167 \times 100 = 91{,}67\%$
4. Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{11+7}{11+7+1+1} = \frac{18}{20} = 0{,}9 \times 100 = 90\%$

From the *confusion matrix calculation* above, it is known that the accuracy level of the *KNN method* for the tested data is 90%.

4.3. Results of Comparison of the Two Methods

After calculating the accuracy level of both methods using the confusion matrix, it can be concluded that the KNN method is superior to the C4.5 method. The KNN method gets a percentage of 90%, while the C4.5 method gets a percentage of 85% in its manual calculation classification. However, when implementing the program, both methods reach the same level of accuracy, which is 90%. So, it can be said that both programs have the same accuracy advantage in the system that has been built, but if only aimed at comparing manual calculations, the KNN method is more accurate.

5. Conclusion

From the results of the research and discussion that has been described, researchers can conclude that the implementation of the algorithm in determining recipients of Direct Cash Assistance (BLT) can be carried out by creating classification/eligibility rules through old recipient data, then classified according to the variables of work, number of dependents, housing status, income, and expenses, the results of the classification or rules can be used as a basis if there is a determination of the next BLT recipient. The level of accuracy obtained from the KNN and C4.5 algorithms in classifying citizens eligible for Direct Cash Assistance (BLT) is 90% in the system that has been built. The results of the comparison of the KNN and C4.5 algorithms for this study show that the KNN algorithm is better because, in manual and system calculations, the level of accuracy reaches 90%. While the C4.5 method only gets 85% for the accuracy of its manual calculations, the system that has been built gets an accuracy level of 90%.

References

[1] Ryanwar, “PENERAPAN METODE ALGORITMA C4 . 5 UNTUK MEMPREDIKSI LOYALITAS KARYAWAN PADA PT . XYZ BERBASIS WEB Laporan Skripsi Disusun oleh :,” 2020.

[2] S. S. Prihatin, P. D. Atika, and Herlawati, “Sistem Informasi Pemilihan Peserta Program Indonesia Pintar (PIP) Dengan Metode K-Nearest Neighbor pada SD Negeri Pejuang V Kota Bekasi,” vol. 2, no. 2, pp. 165–176, 2021.

[3] Rizal, H. A.-K. Aidilof, and W. Kuriniawan, “KLASIFIKASI BERITA OLAHRAGA PADA PORTAL BERITA ONLINE DENGAN METODE K-NEAREST NEIGHBOUR (KNN) DAN LEVENSHTEIN DISTANCE,” pp. 366–384.

[4] A. M. Hasibuan, “BANTUAN PROGRAM KELUARGA HARAPAN,” 2021.

[5] A. Widhianty *et al.*, “Perancangan Sistem Informasi Keuangan Dalam Penyaluran Bantuan Langsung Tunai Dana Desa (BLT-DD) Desa Cikuya Tahun Anggaran 2021 Berbasis Visual Basic,” vol. 10, no. 1, 2023.

[6] J. S. Pasaribu, "Development of a Web Based Inventory Information System," *Int. J. Eng. Sci. InformationTechnology*, vol. 1, no. 2, pp. 24–31, 2021, doi: 10.52088/ijesty.v1i2.51.

[7] R. Mirsa, M. Muhammad, E. Saputra, and I. Farhana, “Space Pattern of Samudera Pasai Sultanate,” *Int. J. Eng. Sci. Inf. Technol.*, vol. 1, no. 2, 2021, doi: 10.52088/ijesty.v1i2.120.

[8] A. Razi, “KLASIFIKASI PENERIMA BEASISWA ACEH CARONG (ACEH PINTAR) DI UNIVERSITAS MALIKUSSALEH MENGGUNAKAN ALGORITMA KNN (K-NEAREST NEIGHBORS),” vol. 7, no. 1, pp. 79–84, 2022.

[9] A. M. H. Pardede *et al.*, "Digital Image Security Application With Arnold Cat Map (ACM)," *J. Phys. Conf. Ser.*, 2018, doi: 10.1088/1742-6596/1114/1/012059.

[10] D. Riyan Rizaldi, E. Nurhayati, Z. Fatimah, and Z. Amni, "The Importance of Parental Assistance in Supervising the Use of Technology for Children During the Home Learning Program," *Int. J. Eng. Sci. Inf. Technol.*, vol. 1, no. 3, 2021, doi: 10.52088/ijesty.v1i3.78.

[11] F. Shidiq, E. W. Hidayat, N. I. Kurniati, and S. Artikel, “Innovation in Research of Informatics (INNOVATICS) Penerapan Metode K-Nearest Neighbor (KNN) Untuk Menentukan Ikan Cupang Dengan Ekstraksi Fitur Ciri Bentuk Dan Canny,” vol. 2, pp. 39–46, 2021.

[12] D. Abdullah and C. I. Erliana, “Model of Ict Goods Inventory Clustering Application Using K-Means Method,” *İlköğretim Online*, vol. 20, no. 1, pp. 1128–1132, 2021, doi: 10.17051/ilkonline.2021.01.116.

[13] A. D. Malik, “Jurnal teknoinfo,” vol. 17, pp. 236–243, 2023.

[14] R. Aryanto, M. A. Rosid, and S. Busono, “Penerapan Deep Learning untuk Pengenalan Tulisan Tangan Bahasa Akasara Lota,” *J. Inf. dan Teknol.*, vol. 5, no. 1, pp. 258–264, 2023, doi: 10.37034/jidt.v5i1.313.

[15] A. Muis, Sulistyawati, and A. Z. Arifin, “Pengaruh Pemberian Kombinasi Pupuk NPK dan Pupuk Kandang Sapi Terhadap Pertumbuhan dan Hasil Tanaman Sorgum (Sorghum bicolor L.),” *Agroteknologi Merdeka Pasuruan*, vol. 2, no. 2, 2018.

[16] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, 2008, doi: 10.1007/s10115-007-0114-2.

[17] Y. R. Amelia, “Penerapan data mining untuk prediksi penjualan produk elektronik terlaris menggunakan metode k-nearest neighbor,” 2018.

[18] C. I. E. Dahlan Abdullah, “Sistem informasi pendataan kendaraan hilang berbasis web pada polres binjai 1,” *Sist. Inf. pendataan kendaraan hilang Berbas. web pada polres binjai 1*, 2016.

[19] A. Deviyanto, “PENERAPAN ANALISIS SENTIMEN PADA PENGGUNA TWITTER,” vol. 3, no. 1, pp. 1–13, 2018.

[20] M. Iqbal *et al.*, "Design of Decision Support System Determination of Inventory Inventory Using Single Exponential Smoothing Forecasting Method," *J. Phys. Conf. Ser.*, 2018, doi: 10.1088/1742-6596/1114/1/012082.

[21] D. Hartama *et al.*, "A research framework of disaster traffic management to Smart City," 2018, doi: 10.1109/IAC.2017.8280607.