Journal Of Engineering And Technology Innovation ( JETI )
Februari 2023 | Vol. 1 | No. 1
E-ISSN : 2828-0288
DOI: https://doi.org/10.66084/jeti.v1i01.7

# Implementation of the C4.5 Algorithm in Determining Priority Locations for Family Planning (KB) Program Counseling

**David Saro**[1]
[1]Informatics Engineering Study Program, Universitas Ibnu Sina, Batam, Indonesia
e-mail: davitsaro@gmail.com

***Abstract***

*Indonesia is one of the most populous countries in the world due to its high population growth rate. To suppress this rate, the government runs the Family Planning Program (Keluarga Berencana/KB). However, KB counseling has not been fully on target because the determination of counseling locations has not been based on adequate data analysis. This study aims to apply the C4.5 algorithm to classify birth data so that priority locations for KB counseling can be determined. The C4.5 algorithm is a data mining technique that builds a decision tree from training data. The data were obtained from the birth database of the Population and Civil Registry Office of Dumai City, comprising 934 records, divided into 70% training data and 30% testing data, and processed using the WEKA tool. The test results show that the C4.5 classification model achieved an accuracy of 99.5% with precision and recall values of 0.995. Based on the resulting decision tree, three urban villages (kelurahan) were identified as priority counseling locations, namely Jaya Mukti, Buluh Kasap, and Bukit Batrem. Therefore, the C4.5 algorithm is proven applicable for determining priority locations of Family Planning Program counseling.*

***Keywords:*** *Data Mining, C4.5 Algorithm, Decision Tree, Family Planning, WEKA*

## INTRODUCTION

One of the government's efforts in dealing with population problems is the Family Planning Program (Keluarga Berencana/KB). The national KB program is one of the programs aimed at suppressing the rate of population growth. The program is formulated as an effort to increase public awareness and participation through marriage age limits, birth control, family resilience development, and family welfare improvement, in order to realize the Small, Happy, and Prosperous Family Norm (NKKBS). The cultivation of this norm is intended to improve the quality of Indonesian human resources by controlling births while ensuring controlled population growth.

The KB program launched by the government has not been fully implemented properly because KB counseling is often held in several areas with low birth rates. If the counseling location is not on target, the program is feared not to achieve its objectives. Therefore, determining the appropriate priority locations for KB counseling is very important in order to reduce the birth rate. An algorithm-based system is needed to assist the National Population and Family Planning Agency (BKKBN) in determining priority counseling locations so that the program is on target.

Based on this description, the research problem is how the C4.5 algorithm can determine the priority locations of KB program counseling. The C4.5 algorithm is used for classification, so the result of processing the testing data is the grouping of data into classes, which are divided into two, namely “not priority” or “priority”. The purpose of this study is to analyze the results of implementing the C4.5 algorithm in classifying birth data so that it can determine the locations that become priorities for KB program counseling.

Several previous studies related to the use of the C4.5 algorithm are as follows. Swastina (2013) applied the C4.5 algorithm to determine student majors and concluded that the C4.5

decision tree can provide solutions that help institutions determine appropriate majors for students. Ridwan (2013) used the C4.5 algorithm to determine graduation predictions based on the attributes of gender, school origin, and Grade Point Average (GPA) from the first to sixth semesters. Kumara and Supriyanto (2013) applied data mining classification to select civil servant candidates using the C4.5 decision tree and obtained a fairly high accuracy level, so the algorithm is considered suitable for application in recruitment processes. Hartato (2014) applied data mining with the C4.5 algorithm to predict student graduation rates into four categories, with the most influential attribute being the sixth-semester GPA.

From these literature reviews, it can be seen that research on implementing the C4.5 algorithm to determine priority locations of KB program counseling has never been conducted. The use of the C4.5 algorithm is considered appropriate because the final result is a decision tree that describes the grouping of data by class.

## RESEARCH METHODS

### *Data Mining*

*Data mining* is the process of discovering meaningful relationships, patterns, and trends by examining large sets of stored data using pattern recognition techniques such as statistical and mathematical methods. *Data mining* is a combination of several disciplines that brings together techniques from *machine learning*, pattern recognition, statistics, databases, and visualization to deal with the problem of retrieving information from large databases (Larose, 2005).

A *decision tree* is one of the classification methods in *text mining*. Classification is the process of finding a collection of patterns or functions that describe and separate data classes from one another, to be used in predicting data that does not yet have a certain class (Han, 2006). Kusnawi (2007) states that *data mining* is a technology that combines traditional analytical methods with sophisticated algorithms to process large volumes of data. *Data mining* is one stage of *Knowledge Discovery in Database* (KDD). The stages of *data mining* are illustrated in Figure 1.

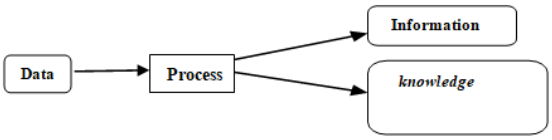


Figure 1. Data Mining Stages (Source: Hermawati, 2009)

### *Knowledge Discovery in Database (KDD)*

KDD is defined as the discovery or search for knowledge (added value) within a database (Hermawati, 2009). Because *data mining* is a series of processes, KDD can be divided into several stages, namely: data cleaning (removing inconsistent data and noise); data integration (merging data from several sources); data transformation (converting data into a form suitable for data mining); the application of data mining techniques; the evaluation of discovered patterns (to find interesting or valuable ones); and the presentation of knowledge using visualization techniques. These stages are illustrated in Figure 2.

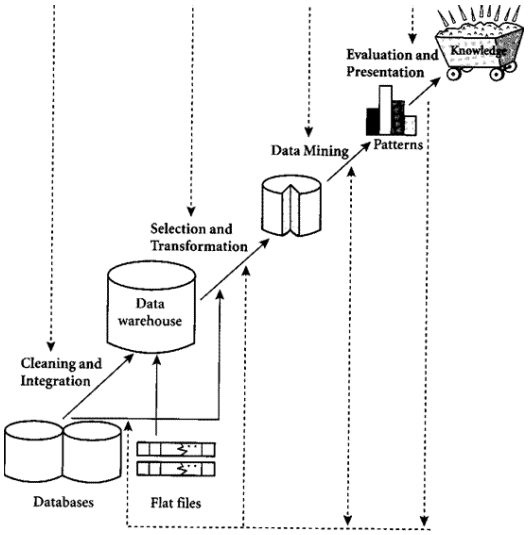


Figure 2. Stages of Knowledge Discovery in Database (Source: Hermawati, 2009)

### *C4.5 Algorithm*

The C4.5 algorithm is a very popular algorithm used by many researchers worldwide. This is explained by Wu and Kumar in their book *The Top Ten Algorithms in Data Mining* (Wu & Kumar, 2009). In addition, the C4.5 algorithm is one of the most effective *decision tree* algorithms for classification (Chauhan, 2013).

The C4.5 algorithm is one of the *machine learning* algorithms. The machine (computer) is given a group of data to study, called a *learning dataset* (Hamdan, 2010). The learning results are then used to process new data called a *test dataset*. The C4.5 algorithm is used to build a decision tree from the data and is a development of the ID3 algorithm. The C4.5 algorithm recursively visits each decision node, choosing the optimal branch, until no more branches are possible (Rahmayuni, 2014). In general, the steps of the C4.5 algorithm in building a decision tree are as follows: (a) select an attribute as the root; (b) create a branch for each value; (c) split cases into branches; and (d) repeat the process for each branch until all cases on a branch have the same class.

The selection of an attribute as the root is based on the highest Gain value among the existing attributes. The Gain value is calculated using Equation (1).

$$Gain(S, A) = Entropy(S) - \Sigma\ (|S_i| / |S|) \times Entropy(S_i) \quad (1)$$

where S is the set of cases, A is the attribute, $|S_i|$ is the number of cases in partition i, and $|S|$ is the number of cases in S. Meanwhile, the Entropy value is calculated using Equation (2).

$$Entropy(S) = \Sigma - p_i \times log_2\, p_i \quad (2)$$

where $p_i$ is the proportion of $S_i$ to S. The iteration process in the decision tree method stops when: (1) all data have been evenly divided; (2) there are no more attributes that can be divided; or (3) there is no data record in an empty branch.

### *WEKA*

WEKA (*Waikato Environment for Knowledge Analysis*) is a package of practical *machine learning* tools developed at the University of Waikato, New Zealand, for research, education, and various applications. WEKA is able to solve real-world data mining problems, especially classification that underlies the machine learning approach. The software is written in a Java class hierarchy using object-oriented methods and can run on almost all platforms (Bouckaert, 2008).

### *Research Design and Stages*

This study uses the decision tree method with the C4.5 algorithm, applied to the WEKA software to determine priority locations for KB program counseling. The research was conducted at the BKKBN of Dumai City to obtain valid attributes or criteria in accordance with the research object, over a period of ten months. The data were sourced from the birth database taken directly through the Population Administration Information System Application at the Population and Civil Registry Office of Dumai City. The sample consisted of infant birth data in each kelurahan (urban village) in East Dumai District in 2016, because the success of the KB program in an area is determined by the high or low birth rate in that area.

Data were collected through direct observation of the problems faced by the BKKBN of Dumai City, interviews with related parties (the BKKBN and the Population and Civil Registry Office), and literature studies of books, journals, and other relevant references. The collected quantitative data were then analyzed using the C4.5 algorithm to classify the data into classes, which were subsequently used to build a decision tree. The research stages are shown in Figure 3.

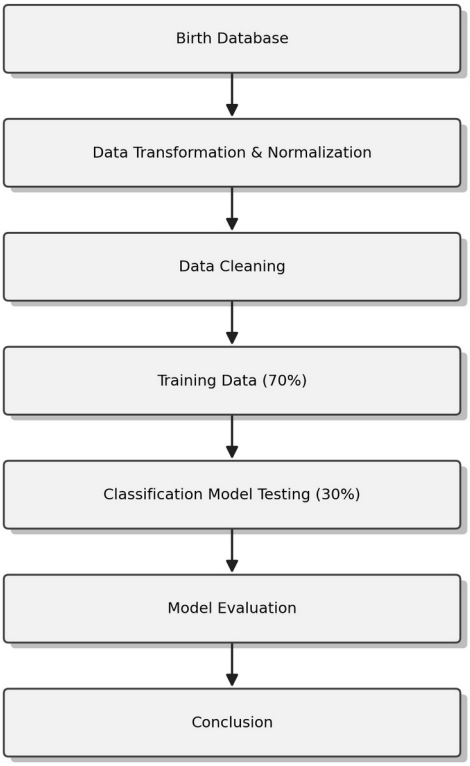


Figure 3. Research Stages

Based on Figure 3, each step can be described as follows. In data transformation, the data from the birth database still contain many unnecessary attributes, so the data are transformed by removing attributes unrelated to the research topic. In data normalization, the measurement scale type is changed from numerical to nominal. In data cleaning, irrelevant data, including missing values, are removed. In the training stage, 70% of the data are used to build the model, while the remaining 30% are used as testing data to evaluate the model. Model evaluation is carried out by examining the accuracy level through the confusion matrix as well as the accuracy and precision tables of the model used.

## RESULTS AND DISCUSSION

This section explains in detail the decision tree calculation process using the C4.5 algorithm to determine priority locations for KB program counseling. The criteria used include Mother's Age, Father's Occupation, and Sub-district (Kelurahan). Each criterion has attributes, with the target attribute being the Decision, valued “Yes” (priority) or “No” (not priority). The primary data from the database are shown in Table 1.

Table 1. Primary Data from the Database

| No | Attribute | | No. of Cases | Yes | No |
|---|---|---|---|---|---|
| | **Total** | | 934 | 849 | 85 |
| **1** | **Age** | | | | |
| | | Productive | 793 | 793 | 0 |
| | | Not Productive | 141 | 56 | 85 |
| **2** | **Occupation** | | | | |
| | | Not Working | 53 | 53 | 0 |
| | | Daily Laborer | 229 | 229 | 0 |
| | | Self-Employed | 413 | 359 | 54 |
| | | Employee | 239 | 208 | 31 |
| **3** | **Sub-district** | | | | |
| | | Teluk Binjai | 242 | 224 | 18 |
| | | Tanjung Palas | 147 | 137 | 10 |
| | | Jaya Mukti | 293 | 259 | 34 |
| | | Buluh Kasap | 104 | 90 | 14 |
| | | Bukit Batrem | 148 | 146 | 2 |

The next step is to calculate the Entropy and Gain values for each attribute. The Entropy of the total data is $-0.908 \cdot \log_2(0.908) - 0.090 \cdot \log_2(0.090) = 0.438$, while Entropy(Productive) = 0 and Entropy(Not Productive) = 0.969. The Gain values are then computed as Gain(Age) = 0.292, Gain(Occupation) = 0.050, and Gain(Sub-district) = 0.043. The recapitulation of the Entropy and Gain calculations for Node 1 is presented in Table 2.

Table 2. Entropy and Gain Calculation Results for Node 1

| Node 1 | | No. of Cases | Yes | No | Entropy / Gain |
|---|---|---|---|---|---|
| Total | | 934 | 849 | 85 | E = 0.438 |
| **Age** | | | | | **G = 0.292** |
| | Productive | 793 | 793 | 0 | 0 |
| | Not Productive | 141 | 56 | 85 | 0.969 |
| **Occupation** | | | | | **G = 0.050** |
| | Not Working | 53 | 53 | 0 | 0 |
| | Daily Laborer | 229 | 229 | 0 | 0 |
| | Self-Employed | 413 | 359 | 54 | 0.558 |
| | Employee | 239 | 208 | 31 | 0.555 |
| **Sub-district** | | | | | **G = 0.043** |

| | | | | |
|---|---|---|---|---|
| Teluk Binjai | 242 | 224 | 18 | 0.382 |
| Tanjung Palas | 147 | 137 | 10 | 0.359 |
| Jaya Mukti | 293 | 259 | 34 | 0.519 |
| Buluh Kasap | 104 | 90 | 14 | 0.569 |
| Bukit Batrem | 148 | 146 | 2 | 0.101 |

Based on Table 2, the attribute with the highest Gain value is Age, with a value of 0.292; thus, Age becomes the root node. Age has two values, namely Productive and Not Productive. The productive age has classified the cases into one class, namely the “Yes” decision, while the not-productive age attribute still needs to be recalculated because it still contains both “Yes” and “No” decisions. The decision tree of Node 1 is shown in Figure 4.

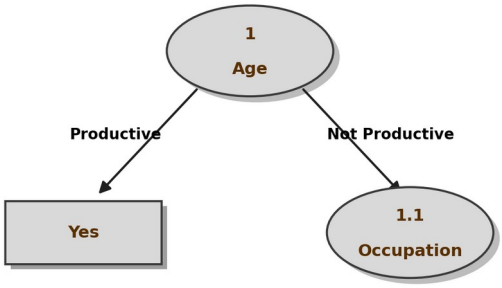


Figure 4. Decision Tree of Node 1

Subsequently, the Entropy and Gain calculation process continues for each branch until all cases on a branch have the same class. Figure 5 shows the final decision tree obtained from the calculation results.

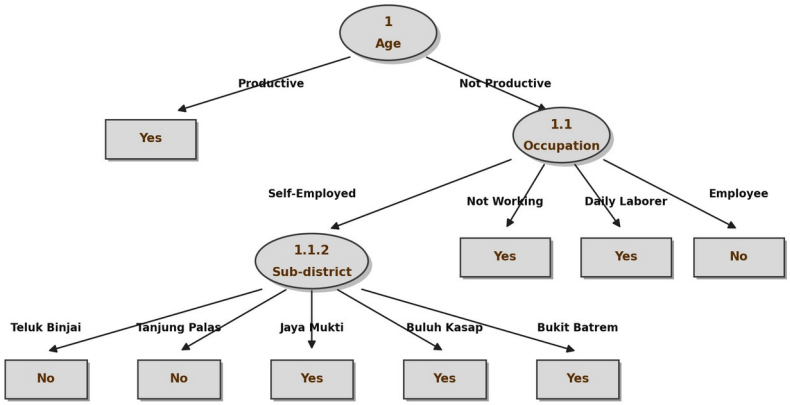


Figure 5. Final Decision Tree

Based on the decision tree in Figure 5, all cases have been grouped into their respective classes. From the resulting tree, nine rules were formed: (1) if Age = Productive, then Class = Yes; (2) if Age = Not Productive and Occupation = Not Working, then Class = Yes; (3) if Age = Not Productive and Occupation = Daily Laborer, then Class = Yes; (4) if Age = Not Productive

and Occupation = Employee, then Class = No; (5–7) if Age = Not Productive, Occupation = Self-Employed, and Sub-district = Bukit Batrem / Buluh Kasap / Jaya Mukti, then Class = Yes; and (8–9) if Age = Not Productive, Occupation = Self-Employed, and Sub-district = Tanjung Palas / Teluk Binjai, then Class = No. From the 934 records of sample data, nine rules were formed with Age as the root of the decision tree and the other attributes as child nodes.

***Implementation in WEKA***

The C4.5 algorithm was implemented using the WEKA *machine learning* tool. Before processing, the data were divided into two parts: 70% training data and 30% testing data. The model was built using the training data and then tested again using the testing data. The completion steps used WEKA version 3.7.4. The data to be tested were first prepared and stored in a *.csv* file using Microsoft Excel. The WEKA tool was then run; its main view is shown in Figure 6. Next, the Explorer was opened and the *.csv* file was imported through the Open file menu. After the file was imported, the Classify tab was selected, followed by Choose to select the classification method; the Trees method was used, then J48 was chosen. Finally, the Start button was clicked, producing the display shown in Figure 7.

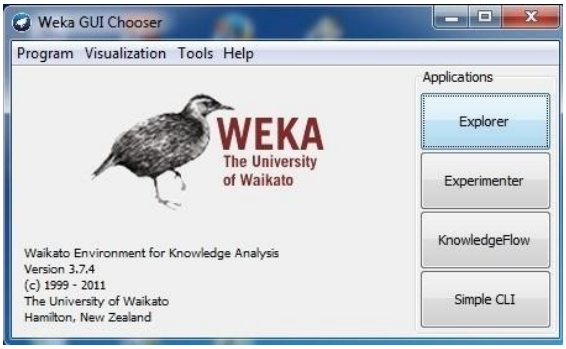


Figure 6. Main View of WEKA Version 3.7.4

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         929               99.4647 %
Incorrectly Classified Instances         5                0.5353 %
Kappa statistic                          0.8753
Mean absolute error                      0.0059
Root mean squared error                  0.074
Relative absolute error                 14.2961 %
Root relative squared error             52.3887 %
Coverage of cases (0.95 level)          99.4647 %
Mean rel. region size (0.95 level)      50.1071 %
Total Number of Instances              934

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.996     0.053      0.999     0.996     0.997      0.971    PRIORITAS
                 0.947     0.004      0.818     0.947     0.878      0.971    TIDAK PRIORITAS
Weighted Avg.    0.995     0.052      0.995     0.995     0.995      0.971

=== Confusion Matrix ===

   a   b   <-- classified as
 911   4 |   a = PRIORITAS
   1  18 |   b = TIDAK PRIORITAS
```

Figure 7. Classification of Training Data

The classification results are presented as a confusion matrix comparing the predicted class with the actual class. The 2×2 confusion matrix model is shown in Table 3.

Table 3. Confusion Matrix Model

| | ***Predict Class*** | |
|---|---|---|
| | **Class A** | **Class B** |

| | | | |
|---|---|---|---|
| ***Actual Class*** | **Class A** | AA | AB |
| | **Class B** | BA | BB |

The *True Positive (TP) Rate* is the proportion of samples classified as class x among all examples that actually belong to class x, equal to recall. The *False Positive (FP) Rate* is the proportion of samples classified as class x but actually belonging to a different class. For the "Yes" class, TP Rate = 911/(911+4) = 0.996 and FP Rate = 1/(1+18) = 0.053; for the "No" class, TP Rate = 18/(18+1) = 0.947 and FP Rate = 4/(4+911) = 0.004. The precision values are 0.999 for the "Yes" class and 0.818 for the "No" class. The confusion matrix produced by the C4.5 algorithm is shown in Figure 8.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 911   4 |   a = PRIORITAS
   1  18 |   b = TIDAK PRIORITAS
```

Figure 8. Confusion Matrix of the C4.5 Algorithm

In addition to accuracy and the *confusion matrix*, the performance of the classification model can be observed from the recall and precision values. Precision is the probability that a selected item is relevant, while recall is the ratio of relevant selected items to the total number of relevant items. The precision and recall values obtained from the model are both 0.995, ranging between 0 and 1; the higher the value, the better the model. The accuracy and error rate are calculated using Equations (3) and (4): Accuracy = (TP + TN)/(TP + TN + FP + FN) = (911 + 18)/934 = 0.995 = 99.5%, and Error Rate = (FP + FN)/(TP + TN + FP + FN) = (1 + 4)/934 = 0.005 = 0.5%. With an accuracy value above 90% (99.5%), the C4.5 algorithm can be applied to the infant birth database at the Population and Civil Registry Office.

## CONCLUSION

Based on the research results, the confusion matrix of the classification model produced a True Positive (TP) value of 911, a False Negative (FN) value of 4, a False Positive (FP) value of 1, and a True Negative (TN) value of 18. The precision and recall values close to 1 indicate that the selected attributes are relevant, supported by an accuracy value of 99.5%. The final decision tree shows that there are three sub-districts (kelurahan) that become priority counseling locations, namely Jaya Mukti, Buluh Kasap, and Bukit Batrem. Therefore, it can be concluded that the C4.5 algorithm can be implemented to determine priority locations for conducting Family Planning Program counseling.

## SUGGESTION

For future research, optimization can be carried out at the attribute selection stage so that the complexity of the attributes can be reduced, and the accuracy value is expected to increase.

## BIBLIOGRAPHY

[1] L. Swastina, "Penerapan Algoritma C4.5 untuk Penentuan Jurusan Mahasiswa," Jurnal GEMA Aktualita, vol. 2, no. 1, pp. 93–98, 2013.

[2] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," Jurnal EECCIS, vol. 7, no. 1, pp. 59–64, 2013.

[3] P. Kumara and C. Supriyanto, “Klasifikasi Data Mining untuk Seleksi Calon Pegawai Negeri Sipil Menggunakan Algoritma Decision Tree C4.5,” Jurnal Teknologi Informasi, 2013.

[4] E. Hartato, “Penerapan Data Mining untuk Prediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma C4.5,” Jurnal Sistem Informasi, 2014.

[5] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, NJ: John Wiley & Sons, 2005.

[6] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2006.

[7] Kusnawi, “Pengantar Solusi Data Mining,” in Seminar Nasional Teknologi 2007 (SNT 2007), Yogyakarta: STMIK AMIKOM, 2007.

[8] F. A. Hermawati, Data Mining. Yogyakarta: Andi Offset, 2009.

[9] X. Wu and V. Kumar, Eds., The Top Ten Algorithms in Data Mining. Boca Raton, FL: Chapman & Hall/CRC, 2009.

[10] H. Chauhan and A. Chauhan, “Implementation of Decision Tree Algorithm C4.5,” International Journal of Scientific and Research Publications, vol. 3, no. 10, pp. 1–3, 2013.

[11] A. R. Hamdan, “Penerapan Algoritma C4.5 untuk Klasifikasi Data,” Jurnal Ilmu Komputer, 2010.

[12] I. Rahmayuni, “Perbandingan Performansi Algoritma C4.5 dan CART dalam Klasifikasi Data,” Jurnal Teknologi Informasi dan Pendidikan, vol. 7, no. 1, 2014.

[13] R. R. Bouckaert et al., WEKA Manual for Version 3-7-4. Hamilton, New Zealand: University of Waikato, 2008.