Journal Of Engineering And Technology Innovation (JETI)
Februari 2022 | Vol. 1 | No. 1
E-ISSN : 2828-1209
P-ISSN : 2828-1462
DOI: -

APPLICATION OF THE C4.5 ALGORITHM TO FIND THE BEST LOCATION FOR FAMILY PLANNING PROGRAM EXPANSION

David Saro *[1]
[1]Universitas Ibnu Sina
e-mail: ***[1]Davitsaro@gmail.com**,

Abstract

Indonesia is one of the most populous countries in the world. This population explosion occurred because of the very high rate of population growth. This situation causes the burden of the state to be even greater. Because it is related to the high and low burden of the state to provide a decent life for every citizen, the government provides a series of efforts to suppress the rate of population growth so that a larger population explosion does not occur. One way that the government does is to promote the Family Planning Program (KB). The family planning program launched by the government to reduce the high birth rate has not been fully implemented because the location of the family planning counseling program is not right on target. Therefore, we need a system that can assist the National Population and Family Planning Agency (BKKBN) in determining the location of Priority Counseling for Family Planning Programs so that the counseling is right on target. Data Mining is the activity of extracting or mining knowledge from big data, the C4.5 algorithm is the algorithm used to build a Results Tree from the data. The C4.5 algorithm is usually used to classify which will then be used in determining the priority locations of family planning programs.

Keywords— *Data Mining, C4.5 Algorithm, Family Planning*

INTRODUCTION

One of the government's efforts in dealing with population problems is the family planning program. The National Family Planning Program (KB) is one of the programs in order to suppress the rate of population growth. The family planning program is formulated as an effort to increase awareness and community participation through marriage age limits, birth control, fostering family resilience, increasing family welfare, to realize the Small Happy and Prosperous Family Norm (NKKBS). Cultivating the norms of small happy and prosperous families in order to improve the quality of Indonesian human resources. The method used is to control births while at the same time ensuring the control of population growth.

The family planning program launched by the government to cultivate the norm of happy and prosperous small families as well as suppress the high birth rate has not been fully implemented properly because family planning program counseling is held in several areas with low birth rates, if the location of the extension is not always well targeted. it is feared that this program will not achieve its objectives. Therefore, determining the priority location of the right family planning program counseling is very important to be able to reduce the birth rate, so we need a system algorithm that can assist the National Population and Family Planning Agency (BKKBN) in determining the priority location of Family Planning Program Extension so that the program is right on target.

From the explanation above, the formulation of the research problem is how the C4.5 algorithm can determine the priority location of family planning program counseling, because

the C4.5 algorithm is used to classify, so the result of processing the test dataset is in the form of grouping data into classes, which class divided into two, namely no priority or yes priority.

The purpose of this study is to analyze the results of the implementation of the C4.5 Algorithm, in classifying birth data so that it can determine which locations are priorities for family planning program counseling.

There are several previous studies that have been made related to the use of the C4.5 Algorithm, including:

1. Research conducted by Liliana Swastina (2013). About the Application of the C4.5 Algorithm for Determining Student Majors With the application of Decision Tree C4.5 can provide solutions for students and help STMIK Indonesia in determining the appropriate majors that will be taken by students during their studies so that the chances of success in higher education are greater .

2. Research conducted by Mujib Ridwan (2013). In this study, the researcher used the C4.5 algorithm in determining graduation predictions based on the attributes of gender, high school origin and GPA in semesters one to six.

3. Next is the research conducted by Kumara and Supriyanto (2013), with the title Classification of Data Mining for the Selection of Candidates for Civil Servants 2014 Using the Decision Tree C4.5 Algorithm. The level of accuracy obtained using the C4.5 algorithm is quite high, therefore it can be concluded that this algorithm is suitable for implementation in research involving the recruitment process.

4. Research conducted by Hartato (2014). About the Application of Data mining with the C4.5 algorithm can be implemented to predict the graduation rate of students with four categories, namely fast passing, right passing, late passing and dropping out. The most influential attribute in the prediction results is the sixth semester GPA.

From several sources of Literature review, researchers can find out that research on the implementation of the C4.5 Algorithm to find out which locations are priorities in the implementation of planned program counseling has never been carried out and the use of the C4.5 algorithm can be used because the final result is a decision tree that describes the grouping of data. by class

Data Mining

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

In short, 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 data class (Han, 2006).

Meanwhile (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 starts from data which is then processed to produce information or generate

knowledge and is one of the steps of Knowledge Discovery in Database (KDD). The following are the stages in Data Mining which are illustrated in Figure 1.

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

Knowledge Discovery in Database (KDD)

The definition of KDD is the discovery or search for knowledge (added value) in a database (Hermawati, 2009), because data mining is a series of processes, data mining can be divided into several stages, namely:

1. Data cleaning (to remove inconsistent data and noise).
2. Data integration (merging data from several sources).
3. Data transformation (data is converted into a form suitable for Data Mining).
4. Application of Data Mining techniques.
5. Evaluate the patterns found (to find interesting/valuable ones).
6. Presentation of knowledge (with visualization techniques).
7. The stages can be illustrated in Figure 2.

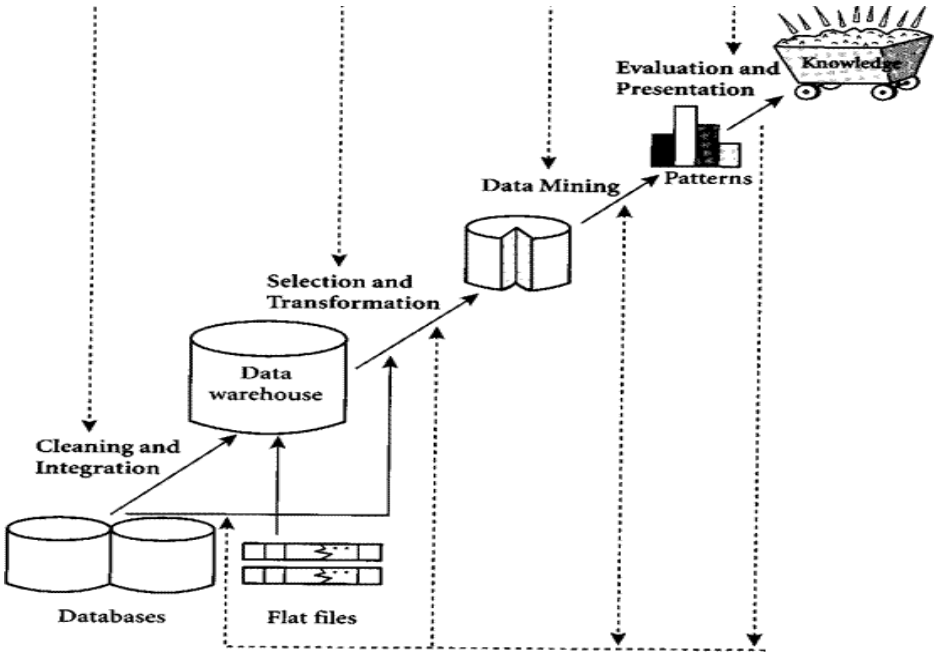


Figure 2. Stages of Knowledge Discovery in Database
Source: (Hermawati, 2009)

C4.5 . Algorithm

The C4.5 algorithm is a very popular algorithm used by many researchers in the world, this is explained by Xindong Wu and Vipin Kumar in their book entitled The Top Ten Algorithms in Data Mining (Wu, 2009), besides the C4.5 algorithm is a one of the most effective Decision Tree algorithms for classifying (Chauhan, 2013).
The C4.5 algorithm is one of the machine learning algorithms, with the C4.5 algorithm the machine (computer) will be given a group of data to study which is called a learning dataset (Hamdan, 2010). Then the results of learning

Then it will be used to process new data called test dataset. C4.5 algorithm is an algorithm used to build a decision tree (Decision Tree) from the data. The C4.5 algorithm is the development of the ID3 algorithm which is also an algorithm for building a decision tree. The C4.5 algorithm recursively visits each decision node, choosing the optimal branch, until no more branches are possible (Rahmayuni, 2014). In general, the C4.5 algorithm in building a decision tree, the steps are as follows.

a. Select attribute as root
b. Create a branch for each value
c. Split cases in branches
d. Repeat the process for each branch until all cases on the branch have the same class.

To select an attribute as the root, it is based on the highest Gain value of the existing attributes. To calculate the Gain, the formula as shown in equation (1) is used below.

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

S : case set
A : attribute
n : number of attribute partition A
$|S_i|$: number of cases on partition i
|S| : number of cases in S

Meanwhile, the calculation of the entropy value can be seen in equation (2) below.

$$Entropy(S) = \sum_{i=1}^{n} - pi * Log2\ pi \quad \text{.....(2)}$$

Description :
S : case set
A : features
n : number of partitions S
pi : the proportion of Si to S

The iteration process in the decision tree method will stop if:
1. All data has been evenly divided
2. No more shareable attributes
3. There is no data record in an empty branch

RESEARCH METHODS

Research methodology is a sequence that is carried out in a study. This research methodology aims to make research more conceptual and structured, so that each stage will be able to see its achievement in accordance with the expected goals of the research.

Research design

This study uses the Decision Tree Method, with the C4.5 Algorithm. which will be applied to the Weka software in determining the priority locations of family planning program counseling, so that research obtains maximum results, of course it must follow the rules (methods) that have been set.

Research Location and Time

A. Research Location
   This research was conducted at the Dumai City BKKBN. The consideration for choosing this location is to obtain valid attributes or criteria, because it is in accordance with the object of research
A. Research Time
   This research will be carried out within a period of 10 months.

Research Data Sources and Samples

A. Research Data Source
   The data of this research is sourced from the Birth database which is directly taken through the Population Administration Information System Application at the Dumai City Population and Civil Registry Office.

B. Research Sample
   The sample in this study is data on births of babies in every Kelurahan in Dumai Timur District in 2016, because the success of the family planning program in an area is determined by the high and low birth rates in that area.

   Data Collection and Analysis Techniques
   A. Data collection techniques that will be carried out by researchers are:

   a. Direct Observation, the researcher observed directly the complaints and problems faced by the Dumai City BKKBN in determining priority locations for holding family planning program counseling.
   b. Interviews, researchers conducted questions and answers with related parties, namely BKKBN as a party directly involved in the study in order to obtain data on the criteria needed to determine the location of family planning program counseling and the Dumai City Population and Civil Registration Office, as a party which will provide the necessary visitor data in the research.
   c. Library Studies, researchers look for supporting data such as books, journals and other literature from internet access.

   B. Data Analysis
   Quantitative data that has been collected is then analyzed using the C4.5 Algorithm to classify the data into classes and then used to build a decision tree.

Research Stages

The stages of this research are the steps that will be taken in solving the problems that will be discussed, while the method used in this study aims to show how a data mining classification model can provide a solution to determine the priority location of family planning program counseling based on existing attributes. The stages of this research can be seen in Figure 3.

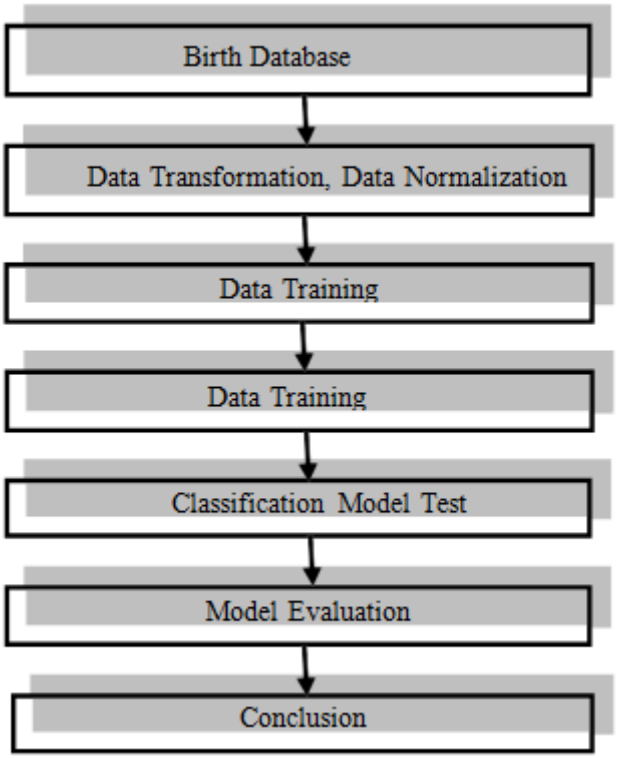


Figure 3. Research Stages

Based on the Research Stages in Figure 3, each step can be described as follows:

a. Data Transformation
   The data obtained from the birth database originating from the Dumai City Population Administration Information System Application, is still in the form of data that contains many attributes that are not needed so that it is necessary to transform the data by removing some attributes that are not related to the research topic.
b. Data Normalization
   The process of normalizing the data in question is to change the type of measurement scale which was originally numerical into nominal
c. Data Cleaning
   The process of cleaning irrelevant data including missing data in attributes
d. Training Data
   The data training process was taken from some of the data contained in the birth database. The proportion of data that was tested was 70% for training, while the rest was used for model testing.
e. Classification Model Test
   The process of model testing is carried out after the data training process is completed, the amount of data that is carried out for model testing is 30% of the baby birth database.

f. Model Evaluation
The evaluation of the model is done by looking at the level of accuracy of the method through the confusion matrix and the accuracy and precision table for the model used.

WEKA Machine Learning

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

RESULTS AND DISCUSSION

In this section, we will explain in detail the decision tree calculation process using the C4.5 algorithm, to determine the priority locations for family planning program extensions. The criteria used include Mother's Age, Father's Occupation and Kelurahan. Each criterion has attributes. One of the attributes is the solution data per data item called the target attribute, the target attribute is the Decision with a value of "Yes" or "No

Table 1. Primary Data from Database

No	Atribut		Amount Case	Yes	No
	Total		934	849	85
1	Age				
		Productive	793	793	0
		Not productive	141	56	85
2	Work				
		Not yet working	53	53	0
		Day Laborer	229	229	0
		Entrepreneur	413	359	54
		Employee	239	208	31
3	Ward				
		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 value and Gain value for each attribute using the following formula:

$Entropy(S) = \sum_{i=1}^{n}$ $- pi * Log_2 pi$ $Entropy(Total) = - (849/934)*\log_2(849/934) - (85/934)*\log_2(85/934)$

$= - (0{,}908* \log_2 0{,}908) – (0{,}090* \log_2 0{,}090)$

$= - (0{,}908* -0{,}139) – (0{,}090* -3{,}473)$

= 0,126 + 0,312

= 0,438

Entropy(Productive) = 0

Entropy(Not productive) = - (56/141)*$\log_2$(56/141)

– (85/141)*$\log_2$(85/141)

= - (0,397* $\log_2$0,397) – (0,602* $\log_2$0,602)

= - (0,397* -1,332) – (0,602* -0,732)

= 0,529 + 0,440

= 0,969

Then calculate the Gain value of each Attribute with the following formula:

Gain(S, A) = Entropy(S) $-\sum_{i=1}^{n} \frac{|Si|}{|S|}$ * *Entropy(Si) Gain(Usia)*

= 0,438 – (793/934*0 + 141/934*0,969)

= 0,438 – (0 + 0,146)

= 0,438 – 0,146

= 0,292

Gain(Work)

= 0,438 – (0 + 0 + 413/934*0,558 + 239/934*0,555)

= 0,438 – (0 + 0 + 0,246 + 0,142)

= 0,438 – 0,388

= 0,050

Gain(Ward)

= 0,438 – (242/934*0,382 + 147/934*0,359 + 293/934*0,519 + 104/934*0,569 + 148/934*0,101)

= 0,438 – (0,098 + 0,056 + 0,162 + 0,063 +0,016)

= 0,438 – 0,395

= 0,043

Table 2. Calculation Results of Entropy Value and Gain Node 1

Node 1			Amount Case	Yes	No	*Entropy*	*Gain*
1	**Total**		934	849	85	0,438	
	Age						**0,292**
		Productive	793	793	0	0	
		Not productive	141	56	85	0,969	
	Work						**0,050**
		Not yet working	53	53	0	0	
		Day Laborer	229	229	0	0	
		Entrepreneur	413	359	54	0,558	
		Employee	239	208	31	0,555	
	Ward						**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	

As seen in Table 2, the attribute with the highest gain value is Age, with a value of 0.292, Age becomes the root node. Age has two values, namely Production and Unproductive. The productive age has classified the cases into 1, namely the "Yes" decision, while the unproductive age attribute still needs to be calculated again because there are still Yes and No decisions.

From the results of these calculations, the decision tree of node 1 can be described as follows.

Next, return to the completion steps and the Entropy and Gain calculation process for node 1.1. Table 3.

Table 3. Calculation Results of Entropy Value and Gain Node 1.1

Node			Amount Case	Priority	No Priority	*Entropy*	*Gain*
1.1	**Age**	Not productive	141	56	85	0,969	
	Work						0,509
		Does not work	5	5	0	0	

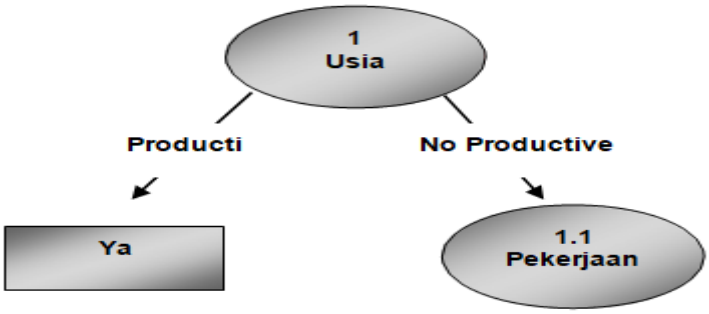
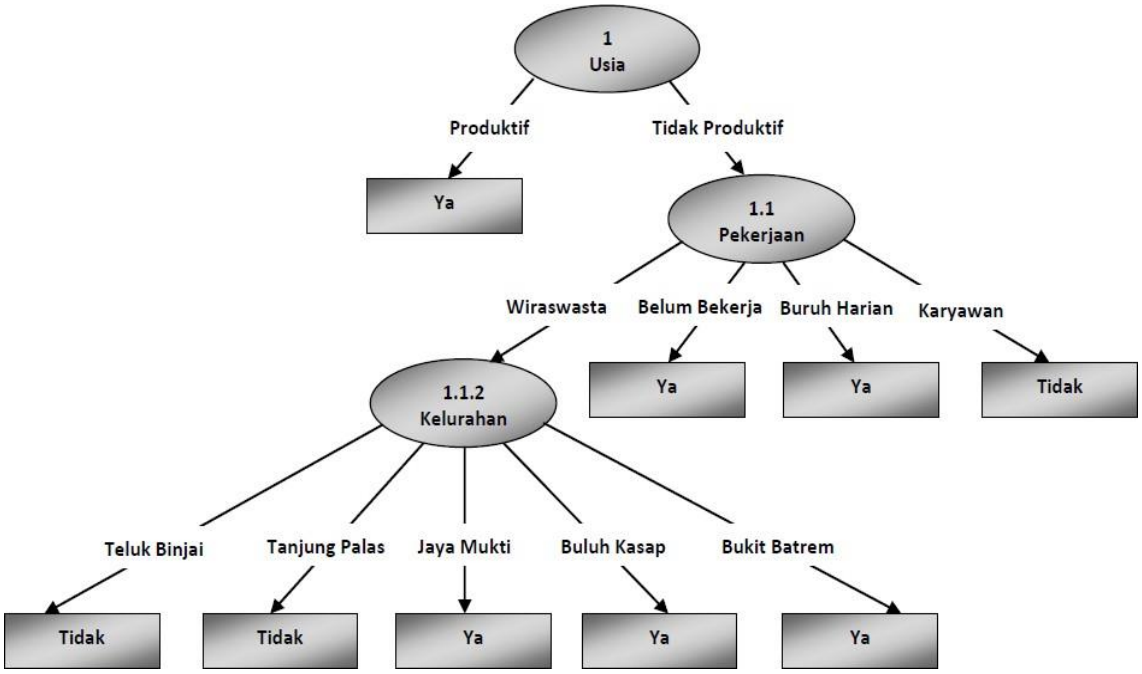


Figure 4. Node.1 Keputusan Decision Tree

Is a recapitulation of the results of the calculation of the value of Entropy and Node Gain 1.1

Next, go back to the completion steps and the process of calculating Entropy and Gain for each branch until all cases in the branch have the same classes. Figure 5. is the final decision tree obtained from the calculation results.



By paying attention to the decision tree in Figure 5, it is known that all cases have been included in their respective classes, from the tree

formed resulted in a number of rules. The rules formed are as follows:
”If Age = Productive Then Class = Yes

”If Age = Unproductive, and Work =
Haven't Worked Then Class = Yes
”If Age = Unproductive, and Work =
Daily Labor Then Class = Yes
“If Age = Unproductive, and Work =
Employee Then Class = No

“If Age = Not Productive, and Occupation = Entrepreneur, and Kelurahan = Bukit Batrem, then Class = Yes
"If Age = Not Productive, and Occupation = Entrepreneur, and Kelurahan = Buluh Kasap, then Class = Yes
"If Age = Not Productive, and Occupation = Entrepreneur, and Kelurahan = Jaya Mukti, then Class = Yes
”If Age = Not Productive, and Occupation = Self Employed, and Kelurahan = Tanjung Palas Then Class = No
"If Age = Not Productive, and Occupation = Self Employed, and Kelurahan = Teluk Binjai, then Class = No
The classification results on the sample data attribute Age as the root of the decision tree, while other attributes as child nodes, from the sample data with 934 records the number of rules formed is 9 rules.

WEKA machine learning implementation
implementing the C4.5 algorithm with the help of WEKA machine learning tools, before processing the data is divided into two parts, the first is training data at 70% and the second is testing data at 30%. This is done so that a model is formed using training data, then the data formed using training data will be tested again using data testing. The following are the completion steps using the WEKA Version 3.7.4 tools.

1. Before we implement the system on the data we want to process, we must first prepare the data to be tested. The data is stored in the form of a file with a .csv extension in Microsoft Excel.
2. Then run the WEKA tool. Figure 6 The following is the main view of the WEKA application.

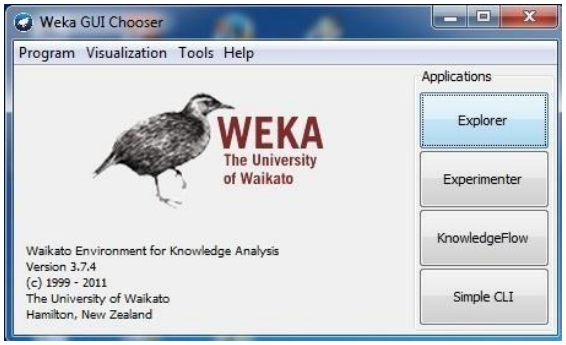


Figure 6. WEKA Version 3.7.4.

3. Then click Explorer, select the data to be processed, which has been saved in csv format, by clicking Open file.
4. The file to be processed has been successfully imported, then click Classify, then click Choose to select the classification method to be processed, in this study the researcher used the Trees method, then chose J48.
5. The next step is to click the start button, a display will appear as shown in Figure 7.

```
=== 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

Table 4. is the classification results that will be presented in the form of a Confusion Matrix. Predict Class and Actual Class. Confusion Matrix 2x2 models.

Table 4. Confusion Matrix Model

		Predict Class	
		Class A	Class B
Actual Class	Class A	AA	AB
	Class B	BA	BB

Figure 8. is the result of the level of accuracy in detail generated from the WEKA Application process.

The True Positive (TP) Rate is the proportion of samples classified as class x, among all examples that actually have class x which has the same value as recall.
To find the value of TP rate in class Yes is

$$\textit{True Positive} = \frac{911}{911 + 4} = 0{,}996$$

To find the TN value in class No is

$$\textit{True Negative} = \frac{18}{18 + 1} = 0{,}947$$

The False Positive (FP) rate is the proportion of samples that are classified as class x, but fall into a different class among all examples that are not class x.

To find the value of FP rate in class Yes is

$$\textit{False Positive} = \frac{1}{1+18} = 0{,}053$$

To find the FN value in class No is

$$\textit{False Negative} = \frac{4}{911+4} = 0{,}004$$

The calculation of the value of Precision Class Yes and No is as follows:

$$\textit{Precision Class Yes} = \frac{911}{911+1} = 0{,}999$$

$$\textit{Precision Class No} = \frac{18}{18+4} = 0{,}818$$

The test results are shown in Figure 9. Confusion Matrix Classification Algorithm C4.5 obtained is an evaluation of the performance of the classification model, and evidence of the occurrence of the results of the classification process that has been provided in it.

```
=== Confusion Matrix ===

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

Figure 9. Confusion Matrix Algorithm C4.5

In addition to accuracy and Confusion Matrix, a classification model can be seen from the recall value and precision. Precision is the probability that a selected item is relevant. While recall is the ratio of the selected relevant items to the total number of relevant items.
The precision and recall results obtained from the above classification model are 0.995 for precision and 0.995 for recall. The results of recall and precision have values between 0-1. The higher the value, the better.

Based on the information above, the process of calculating the average value of the percentage of success accuracy will be carried out using formula (3) and the error rate in the confusion matrix of training data using formula (4).

$$\text{Akurasi} = \frac{\textit{Banyaknya prediksi yang benar}}{\textit{Total banyaknya prediksi}} \quad \text{.....(3)}$$

$$\text{Akurasi} = \frac{911+18}{911+4+1+18} = \frac{929}{934} = 0{,}995$$

Then the Accuracy Percentage Value is

= 0,995 x 100% = 99,5%

$$\textit{Error Rate} = \frac{\textit{Banyaknya prediksi yang salah}}{\textit{Total banyaknya prediksi}} \quad \text{.....(4)}$$

$$\textit{Error Rate} = \frac{4+1}{911+4+1+18} = \frac{5}{934} = 0{,}005$$

Then the Error Rate Percentage Value is

= 0,005 x 100% = 0,5%

From the accuracy value and error rate, the training data using the C4.5 algorithm has an accuracy value of more than 90%, which is 99.5%. This shows that the C4.5 algorithm can be used in the database of baby births at the Population and Civil Registry Office.

CONCLUSION

In detail, the number of True Positive (TP) 911, False Negative (FN) 1, False Positive (TP) 4, True Negative (FN) 18, precision and recall values that are close to 1 indicates that the selected item or attribute is relevant. This is also supported by an accuracy value of 99.5%. The resulting final decision tree shows that there are three areas that are priority locations, namely Jaya Mukti, Buluh Kasap, and Bukit Batrem, so it can be concluded that the C4.5 Algorithm can be implemented to determine priority locations in conducting family planning program counseling.

SUGGESTION

For the next research, optimization can be done at the attribute selection stage so that the complexity of the attributes can be reduced, thus it is expected that the accuracy value will increase.

REFERENCES

Bouchard, K., Bouchard, B., & Bouzouane, A. (2011). A New Qualitative Spatial Recognition Model Based On Egenhofer Topological Approach Using C4. 5 Algorithm: Experiment And Results. *Procedia Computer Science*, *5*, 497-504.

Arifin, N. Y., Kom, S., Kom, M., Tyas, S. S., Sulistiani, H., Kom, M., ... & Kom, M. (2021). *Analisa Perancangan Sistem Informasi*. Cendikia Mulia Mandiri.

Arifin, M. F., & Fitrianah, D. (2018). Penerapan Algoritma Klasifikasi C4. 5 Dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus: PT Atria Artha Persada. *Incomtech: Jurnal Telekomunikasi Dan Komputer*, *8*(2), 87-102.

Putri, R. P. S., & Waspada, I. (2018). Penerapan Algoritma C4. 5 Pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *4*(1), 1-7.

Ermawati, E. (2019). Algoritma Klasifikasi C4. 5 Berbasis Particle Swarm Optimization Untuk Prediksi Penerima Bantuan Pangan Non Tunai. *Sistemasi: Jurnal Sistem Informasi*, *8*(3), 513-528.