Application of Classification Algorithm C4.5 in Recommendations for Natural Tourism Development in District Simalungun

Eka Irawan¹, Sumarno¹, Indra Gunawan¹, Heru Satria Tambunan¹, Hendri Qurniawan¹

¹STIKOM Tunas Bangsa Pematangsiantar, Medan – Indonesia
Email: eka.irawan@amiktunasbangsa.ac.id

Abstract. A tour of human activities to relieve fatigue after a week or a month in carrying out daily activities both as a worker or employee. This activity is carried out to entertain every human being and relax the mind. Besides that, tourism can also help people around the location of tourism objects as a livelihood to meet the needs of life. Simalungun Regency has several natural attractions scattered in several sub-districts which can attract the attention of visitors to travel, but there are still many that have not been well managed by the government or the private sector. To develop the tourism industry in Simalungun Regency, many obstacles faced by the management start from the location, transportation, costs and facilities available and this is also faced by visitors to visit a tourist attraction. In this study, we will choose a location that can be developed based on a questionnaire from several communities/tourists to assess a tourist site that is more supportive to develop. From the results of the election using the C4.5 Algorithm. Then it has been found that the Nature Tourism object can be developed with a level of accuracy that reaches 90% and a recall of 83.33%.

1. Introducing

Tourism objects lately are a necessity for some people who have a busy schedule or are busy in carrying out their daily work to satisfy themselves or pleasure and to spend leisure time. The importance of the role of tourism objects in the surrounding economic development is no longer in doubt for the development of local community activities or in the area of the tourist attraction. In Simalungun district, there are many natural tourism objects in several districts, but in recent years there have been many promotions for these tourism objects using social media. According to Khodyat as a phenomenon arising from human travel and transit, the development of tourism in an area or tourist destination is determined by the following factors [1].

Tourism development will not be separated from physical and non-physical elements. These physical and non-physical elements will be taken into consideration in matters relating to the carrying capacity of the object and consideration of the effects of tourism development. Tourism development in a tourist destination must be based on planning, development and direction of management. Systematic development of tourism and the direction of management itself is in dire need of government attention, as reflected in the formation or recognition of the National Tourism Organization. Local governments have an important role in tourism development, including formulating policies in tourism development and acting as a means of monitoring tourism activities so that they are expected to maximize the potential of tourist destinations.
Development of natural tourism objects can be superior in Simalungun district, this is because natural attractions are more influenced by natural processes, so not all regions have the same characteristics. Simalungun Regency which is designated as a tourist area has 5 superior tourism objects, namely Parapat, Beach in Tigaras, panoramic views of tea gardens, white craters and panarjarunjung peak panorama. But in this study the authors chose 10 natural attractions located in Simalungun district as a natural tourist destination that attracted many visitors to recreation. In this case, there are still many natural tourism objects in Simalungun district that are still being processed by the local and private communities and many ar Using C4.5 algorithm as an algorithm that is used as a classification application to recommend. C4.5 algorithm is the algorithm used to produce a decision tree developed by Ross Qunlan [2]. The basic idea of this algorithm is making a decision tree based on the selection of attributes that have the highest priority or can be called to have the highest gain based on the entropy value of the attribute as the axis of classification attributes [3]. The decision tree produced by the C4.5 algorithm can be used for classification [4]. Decision tree is one of the most popular data mining techniques for knowledge discovery. Systematically analyze and extract rules for classification / prediction purposes [5][6].

In the application and use of C4.5 algorithm, it can be used to predict and classify prospective employees who have the potential to enter the company by making decision trees based on existing data and predicting prospective new employees who want to enter the company [7][8]. In addition, this algorithm is used to classify the predicate of student success in a university. Variables that have the highest priority to the predicate of success of students are students who choose lecture sessions in the morning [9].

Using C4.5 algorithm to do classification in recommending tourist attractions is a good method to do it, so the author takes the title: "Application Of Classification Algorithm C4.5 In Recommendations For Natural Tourism Development In Simalungun District" e still unpublished.

2. Methodology

To do an assessment of the questionnaire data that has been obtained, the author will do the calculation process using the C4.5 classification algorithm to get the highest or best value as a recommendation for the development of tourist attractions in Simalungun district. From the highest / best value, it will be proposed to be developed because the location is more feasible than other locations in order to become a decent tourist destination for tourists / visitors.

In detail this research methodology is designed like a block diagram shown in Figure 1.

![Research Block Diagram](image)

**Figure 1.** Research Block Diagram

For the explanation process of the research block diagram above are as follows:

a. Data collection
   - Questionnaire data collection that has been filled by several visitors who are at the location of the tourism object is then collected which will then be carried out the data analysis process.

b. Data analysis
   - Collecting all the questionnaires that have been distributed to the community and clustering them based on the interests of the community.

c. Process Algorithm C4.5
   - Calculate or assess the questionnaire that has been obtained and then the calculation process will be carried out to recommend tourist attractions to be developed in Simalungun district.

d. Validation and Test Analysis
The results of the calculation will get the best results to determine the natural attractions that might later be developed from this result will be done to test the data. Precision is the percentage of data that is classified as a good model that is actually also good. Recall is a measurement of the actual level of positive recognition

This type of research is quantitative research, because it uses measurement data and objective statistics through scientific calculations derived from community samples by providing answers to a number of questions that are poured into the questionnaire using the C4.5 Algorithm approach method.

3. Results and Discussion

The researcher obtained questionnaire data for tourist recommendations in the development of natural tourism objects, which were made in excel format to be processed to obtain information. The questionnaire data table is as follows:

| Q | Alternatif | Mileage | Transportation | Road conditions | Amenities | View | Security | Decision |
|---|------------|---------|----------------|----------------|-----------|------|----------|----------|
| A1 | BIS        | Far     | There          | Well           | Adequate  | Nice | Secure   | Not      |
|    | KTB        | Near    | There          | broken         | Not       | Nice | Not      | Worthy   |
|    | ATB        | Near    | There          | broken         | Inadequate| Not  | Not      | Not      |
|    | PB         | Near    | Not            | badly damaged  | Not       | Not  | Not      | Worthy   |
|    | PAS        | Far     | Ada            | broken         | Adequate  | Not  | Secure   | Not      |
|    | KA         | Far     | Not            | Well           | Not       | Nice | Not      | Not      |
|    | KP         | Far     | Not            | badly damaged  | Not       | Nice | Not      | Worthy   |
|    | TG         | Far     | There          | Well           | Inadequate| Nice | Secure   | Not      |
|    | ATH        | Far     | Not            | broken         | Not       | Nice | Not      | Worthy   |
|    | AN         | Near    | There          | Well           | Adequate  | Nice | Secure   | Not      |

| Q | Alternatif | Mileage | Transportation | Road conditions | Amenities | View | Security | Decision |
|---|------------|---------|----------------|----------------|-----------|------|----------|----------|
| A2 | BIS        | Far     | There          | Well           | Adequate  | Nice | Secure   | Not      |
|    | KTB        | Near    | There          | broken         | Not       | Nice | Not      | Worthy   |
|    | ATB        | Near    | There          | broken         | Inadequate| Not  | Not      | Not      |
|    | PB         | Near    | Not            | badly damaged  | Not       | Not  | Not      | Worthy   |
|    | PAS        | Far     | Ada            | broken         | Adequate  | Not  | Secure   | Not      |
|    | KA         | Far     | Not            | Well           | Not       | Nice | Not      | Worthy   |
|    | KP         | Far     | Not            | badly damaged  | Not       | Nice | Not      | Worthy   |
|    | TG         | Far     | There          | Well           | Inadequate| Nice | Secure   | Not      |
|    | ATH        | Far     | Not            | broken         | Not       | Nice | Not      | Worthy   |
|    | AN         | Near    | There          | Well           | Adequate  | Nice | Secure   | Not      |

3.1. Data Analysis

The selection of attributes / variables with the consideration that the number of variable values is not so much that it is expected that the prospective partners who are included in one classification of the variable value are quite a lot. From the available data, the variables taken as attributes / variables in the decision tree formation are:

3.2. Process C4.5 algorithm

The C4.5 algorithm process is carried out by changing some attribute values that are initially nominal in value into attribute values that correspond to the data in Table 2 so that the calculation process of C4.5 classification algorithm can be carried out. Entropy is a measure of information theory that can determine the characteristics of the impurity and homogeneity of the data set. From the Entropy value, then the information gain value of each attribute is calculated [11]. Calculation of Entropy values is used in the formula as in Equation (1) [12].
**Entropy(A)** = \( \sum_{i=1}^{n} p_i \log_2 p_i \)  

Information Gain is information obtained from entropy changes in a data set, either through observation or can also be concluded by participating in a data set \[8\].

\[
\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \cdot \text{Entropy}(S_i)
\]

**Table 2.** Entropy and Information Gain of all attributes as a reference for making decision trees by selecting attributes / variables with the biggest information gain as the root node

| Node                  | Number of Cases | Satisfied (S1) | No (S2) | Entropy | Gain        |
|-----------------------|-----------------|----------------|---------|---------|-------------|
| Total                 | 10              | 6              | 4       | 0.970951|             |
| Mileage               |                 |                |         |         | 0.019973    |
| Far                   | 6               | 4              | 2       | 0.918296|             |
| DNear                 | 4               | 2              | 2       | 1       |             |
| Transportation        |                 |                |         |         | 0.419973    |
| There                 | 6               | 2              | 4       | 0.918296|             |
| Not                   | 4               | 4              | 0       | 0       |             |
| Road conditions       |                 |                |         |         | 0.321928    |
| Well                  | 4               | 1              | 3       | 0.811278|             |
| Broken                | 4               | 3              | 1       | 0.811278|             |
| badly damaged         | 2               | 2              | 0       | 0       |             |
| Amenities             |                 |                |         |         | 0.695462    |
| Adequate              | 3               | 1              | 2       | 0.918296|             |
| Inadequate            | 2               | 0              | 2       | 0       |             |
| Not                   | 5               | 5              | 0       | 0       |             |
| View                  |                 |                |         |         | 0.005802    |
| nice                  | 7               | 4              | 3       | 0.985228|             |
| Not                   | 3               | 2              | 1       | 0.918296|             |
| Security              |                 |                |         |         | 0.256426    |
| Secure                | 4               | 1              | 3       | 0.811278|             |
| Not                   | 6               | 5              | 1       | 0.650022|             |

From the calculation results of entropy and information gain obtained then processed into the Decision Tree. Here are the results of the Decision Tree:

**Figure 2.** Decision Tree

From the decision tree (decision tree) formed in Figure 2 above, it is obtained the rules - model rules in determining the recommendations for acceptance of sales partners. There are 4 rules that are formed, can be seen as follows.
If Facilities = No Then Worth (0/5)
If Road Condition = Good Then No (3/0)
If Mileage = Near Then not (1/0)
Else Worth (0/1)

3.3. Validation and Test Analysis
Testing is done by cross validation. One type of cross validation is ten-fold cross validation. Here are the results of ten-fold cross validation.

| PerformanceVector |
|--------------------|
| accuracy: 90.00% |
| ConfusionMatrix:   |
| True: Tidak Layak |
| Tidak: 4 1 |
| Layak: 0 5 |
| precision: 100.00% (positive class: Layak) |
| ConfusionMatrix:   |
| True: Tidak Layak |
| Tidak: 4 1 |
| Layak: 0 5 |
| recall: 83.33% (positive class: Layak) |
| ConfusionMatrix:   |
| True: Tidak Layak |
| Tidak: 4 1 |
| Layak: 0 5 |
| AUC (optimistic): 1.000 (positive class: Layak) |
| AUC: 0.958 (positive class: Layak) |
| AUC (pessimistic): 0.917 (positive class: Layak) |

**Figure 3. Performance Vector**

Based on testing using the Ten-Fold Cross Validation method, it produces nearly perfect accuracy values of 90%, precision of 100% and recall of 83.33%. This shows that, from the classification process carried out, it can be applied to the recommendations for developing tourism objects in Simalungun district.

4. Conclusion
Conclusions of the research carried out are as follows:

a. The application of the C4.5 classification algorithm will be implemented on recommendations for developing Natural Tourism Objects in Simalungun district, judging from the accuracy level of 90% and 83.33% recall, Which states that the calculation will be able to predict and recommend tourist attraction well.

b. Formed 10 rules / model rules can be used as a reference in the design and manufacture of GUI applications.

c. With this algorithm can be used to determine the development of natural attractions in Simalungun district.

References

[1] Kodhyat, H., 1996, “History of Tourism and Its Development in Indonesia”, P.T. Grasindo, Jakarta.

[2] S. S. Nikam, “A Comparative Study of Classification Techniques in Data Mining Algorithms,” *Orient. J. Comput. Sci. Technol.*, vol. 8, no. 1, pp. 13–19, 2015.
[3] L. N. Rani, “Customer Classification Using C4.5 Algorithm As a Basis for Giving Credit,” *J. KomTekInfo Fak. Ilmu Komput.*, vol. 2, no. 2, pp. 33–38, 2015.

[4] H. Widayu, S. Darma, N. Silalahi, and Mesran, “Data Mining to predict the type of customer transactions in savings and loan cooperatives with ...,” *Media Inform. Budidarma*, vol. Vol 1, no. June, p. 7, 2017.

[5] Y. S. Luvia, D. Hartama, A. P. Windarto, and Solikhun, “The application of the C4.5 algorithm for the classification of the predicate of student success in the amic shoots of the nation,” *Juristik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*), vol. Volume 1, no. April 2017, p. 6, 2016.

[6] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. P. Windarto, “Application of Classification Method C4.5 on Selection of Exemplary Teachers,” in *IOP Conference Series*, 2018, pp. 1–6.

[7] Kusrini and E. T. Luthfi, *Algoritma Data Mining. Yogyakarta : Andi*. 2009.

[8] S. Sudirman, A. P. Windarto, and A. Wanto, “Data Mining Tools | RapidMiner: K-Means Method on Clustering of Rice Crops by Province as Efforts to Stabilize Food Crops In Indonesia,” *IOP Conference Series: Materials Science and Engineering*, vol. 420, no. 12089, pp. 1–8, 2018.

[9] K. Adhatrao, A. Gaykar, D. Amiraj, R. Jha, and Honroa Vipul, “Predicting Students’ Performance Using Id3 And C4.5 Classification Algorithms,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 3, no. 5, pp. 39–52, 2013.

[10] Y. Altujjar, W. Altamimi, I. Al-turaiki, and M. Al-razgan, “Predicting Critical Courses Affecting Students Performance : A Case Study,” *Procedia - Procedia Comput. Sci.*, vol. 82, no. March, pp. 65–71, 2016.

[11] K. Adhatrao, A. Gaykar, D. Amiraj, R. Jha, and Honroa Vipul, “Predicting Students’ Performance Using Id3 And C4.5 Classification Algorithms,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 3, no. 5, pp. 39–52, 2013.

[12] P. Premchand and M. ghous. Mohiuddin, “Performance Analysis of Decision Tree Classifiers,” *Int. J. Comput. Sci. Trends Technol.*, vol. 2, no. 2, pp. 62–70, 2014.