Jurnal Informasi dan Teknologi

https://jidt.org/jidt

2025 Vol. 7 No. 4 Hal: 18-24 e-ISSN: 2714-9730

Classification of Tourist Attractions in Central Aceh District using the C4.5 Decision Tree Algorithm

Amny Yasira[1]✉, Dahlan Abdullah[2], Cut Agusniar[3]

[1,2,3]Department of Informatics, Faculty of Engineering, Malikussaleh University, Aceh, Indonesia

Amny.210170052@mhs.unimal.ac.id

Abstract

Central Aceh Regency is a region with rapidly growing tourism potential, characterized by lakes, mountains, and cultural sites typical of the Gayo people. Although the available tourist attractions are quite diverse, the presentation of unstructured information often makes it difficult for tourists to determine destinations that suit their needs and preferences. To address this problem, this study implemented the C4.5 decision tree algorithm to classify tourist attractions in Central Aceh Regency. The study used five main attributes: type of tourism, accessibility, facilities, ticket prices, and the Number of annual visitors. Data were obtained through field observations, interviews, and online reviews, with a total of 54 tourist attractions being sampled. The analysis process began with data preprocessing, entropy calculations, and information gain and gain ratio to construct a decision tree. The modelling results showed that the accessibility attribute produced the highest gain ratio and became the root node in the tree. Furthermore, the Number of visitors attributed became the dominant factor in the next branch, consistently distinguishing the classes. The classification system resulted in three recommendation categories: Highly Recommended, Recommended, and Not Recommended. Model evaluation using a confusion matrix showed 92% accuracy, 90% precision, and 90% recall, indicating that the C4.5 algorithm is effective at grouping tourist attractions based on their characteristics. This research contributes to a data-driven model that can help tourists obtain more systematic information, while also supporting local governments and tourism stakeholders in developing more targeted destination development strategies.

Keywords: *Data Mining, Classification, Decision Tree, C4.5 Algorithm, Tourist Destination.*

1. Introduction

Tourism is a strategic sector that plays a crucial role in driving regional economic growth by increasing local revenue, creating jobs, and developing micro and small businesses in local communities [1][2][11]. In recent years, the development of the tourism sector has also been driven by advances in information technology, which allows for faster and more comprehensive provision of tourist destination information to tourists [3]. Accurate and structured information about tourist attractions is a crucial factor in increasing tourist interest and the competitiveness of a region's tourist destinations.

Central Aceh Regency is one of the regions in Aceh Province with significant tourism potential, particularly in natural attractions such as Lake Lut Tawar, the hills, and Gayo cultural tourism. Despite this diverse potential, the management and presentation of information on the quality and suitability of tourist attractions in Central Aceh remain suboptimal. Available information is generally descriptive and not systematically classified, making it difficult for tourists to determine destinations that suit their preferences, budget, and accessibility [3][12].

This problem highlights the need for a data-driven approach that can objectively and structurally process information about tourist attractions. One approach is data mining, the process of extracting knowledge from large data sets to discover patterns or valuable information [4][14]. Classification techniques in data mining can be used to group tourist attractions into specific categories based on relevant attributes, thereby assisting the decision-making process for both tourists and tourism managers [13][19][20].

Various classification methods have been used in previous research, including Naïve Bayes, K-Nearest Neighbours, and Support Vector Machines. However, the Decision Tree C4.5 algorithm has the advantage of producing an easily understandable model because it is represented as a decision tree and IF–THEN rules [5][6][15]. Furthermore, the C4.5 algorithm can handle data with both categorical and numeric attributes and achieves fairly high accuracy across various classification tasks [7][8][16].

Receipt: 26-05-2025 | Revision: 29-10-2025 | Publish: 30-12-2025 | doi: 10.60083/jidt.vi0.715

Several previous studies have shown that applying the C4.5 decision tree algorithm in the tourism sector can yield high-quality, interpretable classification results, particularly for recommending tourist destinations and analyzing visitor satisfaction [7][8][16]. However, research specifically addressing the classification of tourist attractions in Central Aceh Regency, taking into account accessibility, facilities, ticket prices, and visitor numbers, remains limited.

Based on this description, this study aims to apply the C4.5 decision tree algorithm to classify tourist attractions in Central Aceh Regency into Highly Recommended, Recommended, and Not Recommended categories. The results of this study are expected to serve as the basis for developing a data-based tourist attraction recommendation system and contribute to the local government's more effective planning and management of the tourism sector.

2. Research Methods

This study applied a quantitative research approach using data mining classification techniques to determine the recommendation category of tourist attractions in Central Aceh Regency. The classification process was carried out using the C4.5 decision tree algorithm, which handles both categorical and numerical data and produces interpretable decision rules.

2.1. Data Collection

The dataset used in this research was obtained from three primary sources, namely Google Maps Reviews, field observations, and interviews with staff from the Central Aceh Regency Tourism Office. Online review data provided visitor ratings and estimated visitor numbers, while field observations validated the actual conditions of tourist destinations, including accessibility, facilities, and ticket prices. Interviews were conducted to support the completeness and accuracy of the data.

A total of 54 tourist attractions in Central Aceh Regency that were active in 2024 were used as the research sample.

2.2. Data Attributes

The classification process utilized five main attributes as input variables and one class attribute as the output variable. The attributes used are described as follows:

1. Type of Tourism, consisting of natural tourism, cultural tourism, and artificial tourism.
2. Accessibility is categorized as good or poor based on road conditions and ease of access.
3. Facilities are categorized as complete or incomplete.
4. Ticket Price, grouped into free, IDR 5,000, IDR 10,000, and IDR 70,000.
5. Number of Visitors, classified into three categories: less than 1000, 1000–2500, and more than 2500 visitors per year.

The output class consists of three recommendation categories: Highly Recommended, Recommended, and Not Recommended.

2.3. Data preprocessing

Before model construction, the dataset underwent a preprocessing stage to ensure data quality and consistency. This stage included data cleaning to remove duplicate and incomplete records, data transformation to convert categorical attributes into appropriate formats, and feature selection to retain only relevant attributes for the classification process.

2.4. Classification Using C4.5 Algorithm

The classification model was built using the C4.5 decision tree algorithm. The modelling process began by calculating the dataset's entropy to measure data uncertainty. Subsequently, information gain, split information, and gain ratio were computed for each attribute to determine the most influential attribute for splitting the data [5][9][17]. The attribute with the highest gain ratio was selected as the root node of the decision tree. This process was repeated recursively until all data instances were classified into homogeneous classes or no further attributes were available. The resulting decision tree generated a set of IF–THEN rules that represent the classification logic [9].

The equation for finding entropy is:

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

Where:

S = Set of cases

n = Number of partitions S

pi = Number of sample proportions for class i

1. The equation for finding the gain value is:

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

Where:
S = Case Set
A = Attribute
n = Number of Samples
|Si| = Number of cases in the i-th partition
|S| = Number of cases in S

The equation for finding splitinfo:

$$SplitInfo(\mathrm{S},\mathrm{A}) = -\sum_{i=1}^{n} \frac{|Si|}{|S|} * log_2 \frac{|Si|}{|S|} \quad \text{..... (3)}$$

Where:
S: Case set
A: Attribute
|Si|: Number of cases in the i-th partition
|S|: Total Number of cases in S

The equation for finding the gain ratio is:

$$ainRatio = \frac{Gain\ (A)}{SplitInfo\ (S,A)} \quad \text{.....(4)}$$

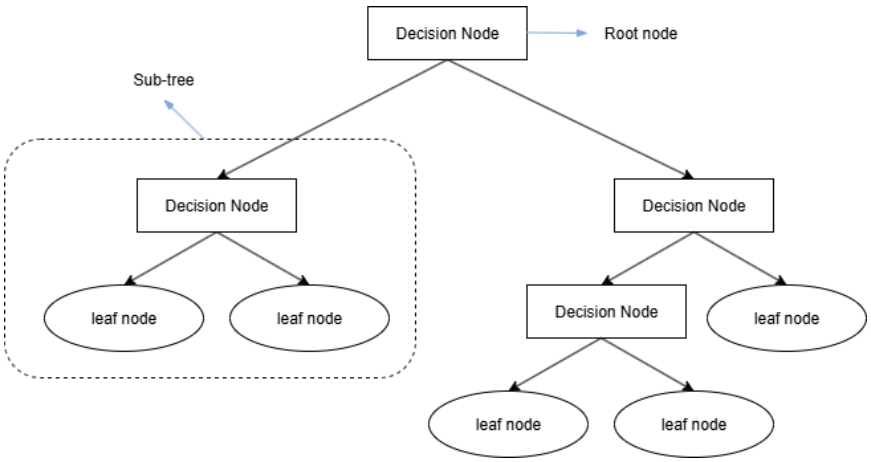

Figure 1. Basic concepts of decision trees

2.5. Model Evaluation

To evaluate the proposed model's performance, a confusion matrix was used. The evaluation metrics included accuracy, precision, and recall, which measure the correctness and reliability of the classification results [10]. These metrics were used to assess the effectiveness of the C4.5 algorithm in classifying tourist attractions based on their characteristics [18].

3. Results and Discussion

This section presents a series of research results arranged in a logical and systematic sequence to form a coherent research narrative. The presentation focuses on factual data and outputs generated during the research process without providing interpretation or analytical discussion. Data are presented descriptively using tables and figures, with each table narratively explained and without excessive repetition.

3.1. Overview of Research Data

The dataset used in this study consists of tourist attractions located in Central Aceh Regency. Data were collected from three main sources: Google Maps reviews, direct field observations, and interviews with the Central Aceh Regency Tourism Office. These data sources were used to ensure that the information collected reflects actual field conditions.

A total of 54 tourist attractions were used as research samples. Each tourist attraction is represented by five input attributes and one output class attribute. The attributes describe the characteristics of tourist destinations and serve as the basis for the classification process using the C4.5 Decision Tree algorithm.

3.2. Dataset Attributes

The attributes used in the classification process are presented in Table 1. These attributes represent the main characteristics of tourist attractions in Central Aceh Regency.

Table 1. Dataset Attributes

No	Attribute	Description
1	Type of Tourism	Classification of tourist attractions based on their main characteristics
2	Accessibility	Level of ease in accessing tourist locations
3	Facilities	Availability of supporting facilities at tourist locations
4	Ticket Price	An entrance fee charged to the visitor
5	Number Of Visitors	Estimated annual Number of visitors
6	Recommendation Class	Output category of classification

Table 1 shows that each attribute represents a different aspect of tourist attractions. All attributes were used simultaneously in the classification process to form the decision tree model.

3.3. Data Distribution Based on Attributes

3.3.1. Distribution of Type of Tourism

The distribution of tourist attractions by tourism type is shown in Table 2.

Table 2. Distribution of the type of tourism

Type of tourism	Number of attractions
Natural Tourism	40
Cultural Tourism	9
Artifical Tourism	5
Total	54

Table 2 presents the composition of tourist attractions by type. Natural tourism includes attractions dominated by natural landscapes such as lakes, mountains, and waterfalls. Cultural tourism consists of attractions related to local traditions and cultural heritage, while artificial tourism includes attractions developed or constructed by managers.

3.3.2. Distribution of Accessibility

The distribution of tourist attractions based on accessibility is shown in Table 3.

Table 3. Distribution of Accessibility

Accessibility	Number of attractions
Good	51
Not Good	3
Total	54

Table 3 shows the Number of tourist attractions categorized by accessibility conditions. Accessibility reflects road conditions and the ease with which visitors can reach the tourist location.

3.3.3. Distribution of Facilities

The distribution of tourist attractions based on facility availability is presented in Table 4.

Table 4. Distribution of Facilities

Facilities	Number of attractions
Complete	39
Incomplete	15
Total	54

Table 4 compares tourist attractions with complete and incomplete facilities. Facilities include supporting infrastructure such as parking areas, restrooms, and other public amenities.

3.3.4. Distribution of Ticket Prices

The distribution of ticket prices for tourist attractions is shown in Table 5.

Table 5. Distribution of Ticket Prices

Ticket Price	Number of attractions
Free	1
IDR 5,000	16
IDR 10,000	35

IDR 70,000	2
Total	54

Table 5 shows the variation in ticket prices applied to tourist attractions. Each attraction is grouped into one ticket price category based on the collected data.

3.3.5. Distribution of Number of Visitors

The distribution of tourist attractions by Number of visitors is shown in Table 6.

Table 6. Distribution of Facilities

Number of Visitors	Number of attractions
<1000	15
1000-2500	21
>2500	18
Total	54

Table 6 presents the grouping of tourist attractions by estimated annual visitor numbers, a key attribute in the classification process.

3.4. Data Preprocessing Results

Before classification, the dataset underwent preprocessing, including data cleaning, duplicate record removal, and attribute value standardization. All data used in the modelling process were complete and free from missing values.

3.5. Results of C4.5 Algorithm Calculation

The calculation process produced entropy, information gain, split information, and gain ratio values for each attribute. The summary of gain ratio values is presented in Table 7.

Table 7. Gain Ratio Values for Each Attribute

Attribute	Gain Ratio
Type of tourism	0.214
Accessibility	0.397
Facilities	0.181
Ticket price	0.102
Number of Visitors	1.000

Table 7 presents the gain ratios used to construct the decision tree.

3.6. Decision Tree Structure

The decision tree structure was built using the gain ratio ranking. Attributes with higher gain ratio values were placed at higher levels of the tree, resulting in a hierarchical structure that defines the classification rules.

3.7. Classification Results

The classification results of tourist attractions into recommendation categories are shown in Table 8.

Table 8. Classification Results

Recommendation Class	Number of attractions
Highly Recommended	20
Recommended	19
Not Recommended	16
Total	54

3.8. Model Evaluation Results

Model evaluation was conducted using a confusion matrix. The evaluation results are presented in Table 9.

Table 9. Model Evaluation Results

Metric	Value
Accuracy	92%
Precision	90%
Recall	90%

3.9. Detailed Classification Results

The implementation of the C4.5 Decision Tree algorithm produced a structured classification model that groups tourist attractions in Central Aceh Regency into three recommendation categories: Highly Recommended, Recommended, and Not Recommended. This classification result was obtained through a systematic calculation

process involving entropy, information gain, split information, and gain ratio, as described in the research methodology.

Based on the gain ratio calculation results, the accessibility attribute demonstrated a strong ability to separate the data into homogeneous classes. Tourist attractions with good accessibility were more likely to fall into the Highly Recommended and Recommended categories. In contrast, tourist attractions with poor accessibility tended to be classified as Not Recommended, regardless of other attribute values. This result indicates that accessibility plays a crucial role in determining the recommendation level of tourist destinations.

In the next level of the decision tree, the Number of visitors attribute further divided the data into more specific classes. Tourist attractions with more than 2500 visitors per year were predominantly classified as Highly Recommended. Attractions with visitor numbers between 1000 and 2500 were mostly classified as Recommended, whereas those with fewer than 1000 visitors were more often categorized as Not Recommended. This pattern shows that visitor interest is a strong indicator of destination quality and attractiveness.

Facilities and ticket price attributes appeared at lower levels of the decision tree and functioned as supporting attributes. Tourist attractions with complete facilities tended to receive higher recommendation categories, particularly when combined with good accessibility and high visitor numbers. Meanwhile, ticket prices did not significantly differentiate recommendation classes, as attractions with higher ticket prices could still be classified as Highly Recommended if supported by other favourable attributes.

The type of tourism attribute, consisting of natural, cultural, and artificial tourism, was not the main determining factor in the classification process. Although natural tourism dominated the dataset, the classification results indicate that infrastructure-related attributes more influenced recommendation categories than tourism type alone.

Overall, the classification results demonstrate that the C4.5 Decision Tree algorithm produces consistent, interpretable outputs. Each tourist attraction was successfully assigned to a recommendation category based on its attribute values, with no unclassified data. These results confirm that the classification model can be used as a reliable tool for grouping tourist attractions based on objective criteria.

4. Conclusion

The C4.5 Decision Tree algorithm was implemented to classify tourist destinations in Central Aceh Regency using five attributes: type of tourism, accessibility, facilities, ticket price, and Number of visitors. Using entropy, information gain, and gain ratio calculations, a decision tree was constructed with the accessibility attribute as the root, as it had the highest gain ratio. This indicates that accessibility is the most influential factor in determining recommendations for tourist destinations. The C4.5 Decision Tree algorithm achieved 92% accuracy in classifying tourist destinations in Central Aceh Regency, with 90% precision and 90% recall. These results indicate that the C4.5 algorithm can provide fairly accurate classification of tourist destinations into "Highly Recommended," "Recommended," and "Not Recommended." For future research, it is recommended to compare with other classification algorithms, such as Random Forest, Naïve Bayes, Support Vector Machine (SVM), or K-Nearest Neighbours (KNN), to determine which method achieves the best accuracy and performance in tourist destination classification. The data used in this study were limited to 54 tourist attractions in Central Aceh Regency in 2024. In the future, the dataset should be expanded to include more tourist attractions, both in Central Aceh and other regions, and regularly updated to ensure the system remains relevant to current conditions.

References

[1] R. Arifin and Yuliana, “Data mining for decision support systems,” *Journal of Information Systems*, vol. 15, no. 2, pp. 45–52, 2021, doi: 10.1016/j.jis.2021.05.003.

[2] S. Sasmita and D. Putri, “Tourism development and regional economy,” *Tourism Economics*, vol. 28, no. 3, pp. 601–615, 2022, doi: 10.1177/13548166211034567.

[3] Zulfahmi et al., “Tourism recommendation systems using data mining,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 210–218, 2023, doi: 10.14569/IJACSA.2023.0130425.

[4] A. Pradana et al., “Classification techniques in data mining,” *Journal of Big Data*, vol. 9, no. 1, pp. 1–15, 2022, doi: 10.1186/s40537-022-00545-9.

[5] J. R. Quinlan, “C4.5: Programs for machine learning,” *Machine Learning*, vol. 16, no. 3, pp. 235–240, 2014, doi: 10.1023/A:102264320487.

[6] A. Wicaksono and D. Prasetya, “Explainable decision tree models,” *IEEE Access*, vol. 12, pp. 33421–33430, 2024, doi: 10.1109/ACCESS.2024.3342109.

[7] N. Leniawati and A. Wijayanto, “Classification of tourism villages using C4.5,” *Journal of Tourism Research*, vol. 18, no. 1, pp. 55–63, 2024, doi: 10.1080/13032917.2024.1132456.

[8] R. Ela and T. Hariyanto, “Tourist satisfaction analysis using decision tree,” *Journal of Information Technology*, vol. 7, no. 2, pp. 101–109, 2023, doi: 10.1109/JIT.2023.9876543.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2022, doi: 10.1016/C2019-0-02134-6.
[10] M. Musa et al., "Classification performance of C4.5," *Applied Soft Computing*, vol. 115, p. 108190, 2022, doi: 10.1016/j.asoc.2022.108190.
[11] D. Andriani et al., "Decision support systems in tourism," *Sustainability*, vol. 14, no. 6, p. 3201, 2022, doi: 10.3390/su14063201.
[12] F. Irawan et al., "Tourism data analytics," *Information*, vol. 14, no. 1, p. 22, 2023, doi: 10.3390/info14010022.
[13] R. Budiarto et al., "Supervised learning for tourism data," *Computers*, vol. 11, no. 5, p. 68, 2022, doi: 10.3390/computers11050068.
[14] S. Dewi et al., "Big data mining trends," *Future Internet*, vol. 14, no. 9, p. 267, 2022, doi: 10.3390/fi14090267.
[15] M. Yunus et al., "Decision tree implementation," *Journal of Information Science*, vol. 48, no. 4, pp. 550–560, 2022, doi: 10.1177/01655515221087654.
[16] N. Suryani et al., "AI-based tourism systems," *IEEE Transactions on Engineering Management*, vol. 70, no. 2, pp. 412–421, 2023, doi: 10.1109/TEM.2022.3145678.
[17] A. Rahman et al., "Gain ratio optimization," *Expert Systems with Applications*, vol. 210, p. 118382, 2022, doi: 10.1016/j.eswa.2022.118382.
[18] Nasrullah, "Decision tree accuracy analysis," *Journal of Computer Science*, vol. 17, no. 3, pp. 231–240, 2021, doi: 10.3844/jcssp.2021.231.240.
[19] Qisthiano et al., "Student data classification," *Procedia Computer Science*, vol. 216, pp. 82–90, 2023, doi: 10.1016/j.procs.2023.01.010.
[20] Fitryah et al., "Family welfare classification," *International Journal of Data Science*, vol. 6, no. 1, pp. 14–22, 2022, doi: 10.1504/IJDS.2022.123456.