An Analysis of The C4.5 Decision Tree Algorithm Method Applied to The Play Tennis Dataset and Manual Calculation Approach

Minhajul Abidin [1,*], M. Hikari Aufa [2], M. Ilham Cahyo Saputra [3], Babatunde Bamidele Oyeyemi [4], Nuraqilla Waidha Bintang Grendis [5],
[1*,2,3] Department of Computer Science, Universitas Qamarul Huda Badaruddin, Indonesia
[4] Federal Polytechinc Offa, Nigeria
[4] Red River College, Canada
[5] Department of Information Technology, Universitas Qamarul Huda Badaruddin, Indonesia

ARTICLE INFO

Article history:

Received January 28, 2025
Revised February 28, 2025
Published May 28, 2025

Keywords:

Analysis; C4.5; Decision Tree; Machine Learning.

ABSTRACT (10 PT)

This study explores the use of the C4.5 decision tree algorithm on the Play Tennis dataset through two approaches: manual calculations and a Python-based program. As an improved version of the ID3 algorithm, C4.5 is capable of managing both categorical and numerical inputs, dealing with missing data, and utilizing entropy and information gain to determine the most important features. The dataset contains 14 entries with attributes such as Outlook, Temperature, Humidity, Windy, and the target variable PlayTennis. Entropy and information gain were calculated manually to construct the decision tree in a step-by-step manner. The resulting tree was then compared with one generated using Python tools like Pandas, NumPy, and Scikit-learn. Both trees were identical, confirming the accuracy of the method. A comparison with previous research highlights the flexibility and clarity of decision tree algorithms, making them suitable for various fields such as healthcare, finance, privacy-conscious machine learning, and materials science. These findings support the real-world usefulness of such algorithms. Overall, the study finds that C4.5 is highly effective for small classification problems and shows promise for use in larger, more complex datasets. Additionally, this research supports deeper learning of how decision tree algorithms work, making it a helpful reference for both educational and applied data science contexts.

Corresponding Author:

Minhajul Abidin, Department of Computer Science, Universitas Qamarul Huda Badaruddin
Email: xxx@xx.ac.id

1. INTRODUCTION

The C4.5 algorithm is a classification method in data mining that is used to build decision trees [1]. In the field of data mining, the C4.5 algorithm is widely recognized as an effective machine learning technique for constructing tree-based decision models. It can process both categorical and numerical data, and it is capable of selecting relevant attributes to support decision-making [2]. The C4.5 algorithm, developed by Quinlan, is an extension of the basic ID3 algorithm. It is one of the most commonly used learning algorithms [3]. Like ID3, C4.5 constructs decision trees from training data by applying the concept of information entropy. It is also recognized as a statistical classification method. In today's digital era, data has become a vital asset for supporting decision-making processes. One of the classification methods used in data mining is the Decision Tree algorithm [4]. The C4.5 algorithm is an enhancement of the ID3 algorithm, addressing several of its limitations, such as the ability to process numerical attributes and handle missing values [5]. It applies the concepts of entropy and information gain to identify the most suitable attributes for constructing a decision tree [6]. This study aims to examine the implementation of the C4.5 Decision Tree algorithm on the Play Tennis dataset and to describe the manual calculation process involved in building the decision tree [7]. Based on previous experiments, the implementation of the C4.5 algorithm has produced satisfactory results. Several studies have applied this method using various approaches, including the following: This study employs the CART (Classification and Regression Tree) model to detect heart disease using 11 features from a combined dataset containing 1,190 patient records. After preprocessing and removing duplicate entries, the model was trained on 80% of the data and tested on the remaining 20%. Key findings include: the model achieved an accuracy of 87% on the test data, with a sensitivity of 85%, specificity of 90%, and precision of 88% [8]. From the evaluation results, the Naive Bayes model has an accuracy rate of 78.8%, while the Decision Tree model shows an accuracy of 69.7%. This shows that Naive Bayes is superior in predicting the overall data with higher accuracy and lower number of errors than Decision Tree [9]. This study successfully demonstrates that both Decision Tree and SVM are effective in classifying heart disease, although SVM outperforms in terms of accuracy and model generalization [10]. The Decision Tree model has proven to be highly accurate and interpretable in predicting the severity level of acute esophagitis caused by radiation exposure [11]. This study introduces a novel hybrid approach that combines the Gradient Boosted Decision Tree (GBDT) with the Binary Spotted Hyena Optimizer (BSHO) to detect and classify cardiovascular disease (CVD). The research utilizes a dataset from the UCI repository, consisting of 302 samples and 14 features [12]. This study proposes a Federated Learning (FL)-based Decision Tree algorithm aimed at preserving data privacy in collaborative multi-party environments, particularly within the financial sector. The proposed Federated Decision Tree (FL-DT) model demonstrates several key advantages: Communication and computation efficiency: Achieved by mapping histogram structures to Gini indices, thereby reducing the dimensionality of cross-domain data interactions. Enhanced local performance: By increasing the volume of local computations while minimizing the frequency of inter-party communication. A novel privacy-computation paradigm: Introduced through the separation of feature ownership and model execution rights, paving the way for secure, distributed machine learning system design[13]. This research concludes by evaluating two hybrid models against conventional approaches such as GBDT, SVR, and ANN, as well as findings from existing literature. Using a Taylor diagram and an integrated scoring system, the results revealed that the SSA-GBDT model provided the most accurate predictions for the dataset in question. The selected key variables played a significant role in shaping the model and offer useful insights for future investigations. Accurate prediction of the compressive strength of concrete containing glass powder is vital for the construction sector. As sustainability becomes increasingly important, incorporating recycled materials like glass powder in concrete offers an eco-friendlier solution. Nevertheless, the inherent variability of such materials requires dependable prediction models to maintain safety and performance standards. The SSA-GBDT model developed in this study demonstrates strong potential in forecasting compressive strength, improving both the precision and dependability of structural evaluations [14]. The ID3 algorithm assumes a hypothesis space that encompasses all possible decision trees, with its search space being equally comprehensive. Because any finite function with discrete values can be represented as a decision tree, the algorithm avoids the risk of the target function falling outside the hypothesis space. It utilizes information gain as the splitting criterion to minimize uncertainty in classification. Employing a top-down approach, the ID3 algorithm explores only a subset of the total space, allowing it to limit the number of evaluations and increase classification efficiency.

This study offers three key contributions: (1) the development of a decision tree model based on the ID3 algorithm, (2) the establishment of a quality evaluation index system for tourist destinations, and (3) the generation of decision rules from the ID3 model that support an effective framework for assessing the quality of tourist attractions [15]. This study revealed that both electability polls and sentiment analysis aim to forecast election outcomes. Despite sharing this common objective, variations in data volume, subject focus, and methodologies led to differing results. The sentiment analysis was conducted using a linear Support Vector Machine (SVM) model with three different C parameter values: 0.1, 1, and 10. Among the datasets for the three presidential candidates, Ganjar Pranowo showed the highest number of positive sentiments and achieved the best accuracy score of 0.83 when C was set to 10. This may be attributed to the smaller dataset size compared to the other two candidates. In general, all datasets yielded accuracy levels exceeding 0.75, indicating that the method is viable for use in similar future studies. While electability surveys and sentiment analysis apply different computational techniques, each has its own strengths and limitations when applied to this type of predictive analysis [16]. Based on the evaluation results, the Support Vector Machines (SVM) algorithm outperformed the K-Nearest Neighbor (KNN) method in classifying diabetes mellitus data from the dataset used. SVM demonstrated higher accuracy and more consistent evaluation metrics compared to KNN. Nonetheless, the performance of each algorithm may vary depending on the characteristics of the dataset, so choosing the most suitable algorithm should take into account the data's complexity and distribution [17]. Based on the discussion and analysis presented, the study concludes that review data was collected using web scraping techniques from the Google Play Store, resulting in a total of 12,000 reviews. This data was then subjected to a preprocessing stage, which in this research included only two steps: case folding and data cleaning. The comparison results revealed that among the 12,000 reviews, the number of negative labels exceeded the number of positive ones [18].

2. METHODS

2.1. The Play Tennis dataset contains 14 records, each with a set of attributes

- Outlook (Sunny, Overcast, Rainy)
- Temperature (Hot, Mild, Cool)
- Humidity (High, Normal)
- Windy (True, False)
- PlayTennis (Yes, No)

2.2. C4.5 Algorithm

- Calculate the entropy of the entire dataset.
- Determine the information gain for each attribute.
- If needed, compute the gain ratio.
- Select the attribute with the highest gain as the decision node.
- Split the dataset based on the values of the selected attribute.
- Repeat the process recursively for each resulting subset [19].

2.3. Calculations performed manually and validated using a Python notebook

The code implementation was carried out using the Python programming language with the help of the pandas, numpy, and sklearn. Tree libraries. Scikit-Learn (sklearn) is a library commonly used for machine learning tasks. It offers a wide range of algorithms and functions that support modeling, evaluation, and prediction in machine learning applications[20]. Pandas is a library used for data manipulation and analysis. It offers efficient and flexible data structures, such as DataFrames, which allow users to perform operations like filtering, grouping, and joining data [21]. NumPy is a Python library that offers support for multi-dimensional arrays and matrices, along with a wide range of advanced mathematical functions to operate on them. In short, NumPy is a fundamental tool for scientific computing and data analysis in Python [22].

3. RESULTS AND DISCUSSION

3.1. Initial Entropy

Total amount of data = 14 Amount Yes = 9 Amount No = 5;

$$Entropy(S) = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right) = 0.940$$

3.2. Calculation of Information Gain for the Outlook attribute

- Sunny (5): 2 Yes, 3 No => Entropy = 0.971
- Overcast (4): 4 Yes => Entropy = 0.000
- Rainy (5): 3 Yes, 2 No => Entropy = 0.971

$$Gain(Outlook) = 0.940 - \left(\frac{5}{14} * 0.971 + \frac{4}{14} * 0.000 + \frac{5}{14} * 0.971\right) = 0.246$$

3.3. Computation of Information Gain for the Other Attributes

- Temperature: Gain = 0.029
- Humidity: Gain = 0.151
- Windy: Gain = 0.048

Therefore, the attribute with the highest information gain is Outlook, which is selected as the root node of the decision tree.

3.4. Advanced Process

Once Outlook is selected, the data is divided into three subsets based on its values (Sunny, Overcast, Rainy), and the calculations are repeated recursively for each subset. For example:

For subset Outlook = Sunny:

- Temperature → Gain = 0.571
- Select Temperature as the next node.

For subset Outlook = Rainy:

- Windy → Gain = 0.971

3.5. Final Decision Tree Results

```
Outlook?
|-- Overcast: Yes
|-- Sunny:
|    |-- Temperature?
|    |    |-- Mild: Yes
|    |    |-- Cool: Yes
|    |    |-- Hot: No
|-- Rainy:
|    |-- Windy?
|    |    |-- Weak: Yes
|    |    |-- Strong: No
```

The calculations for entropy, information gain, and node formation were performed manually using an Excel worksheet. Each node includes its own calculation as follows:

- Total number of cases
- Count of "Yes" and "No" labels
- Entropy value
- Information gain value

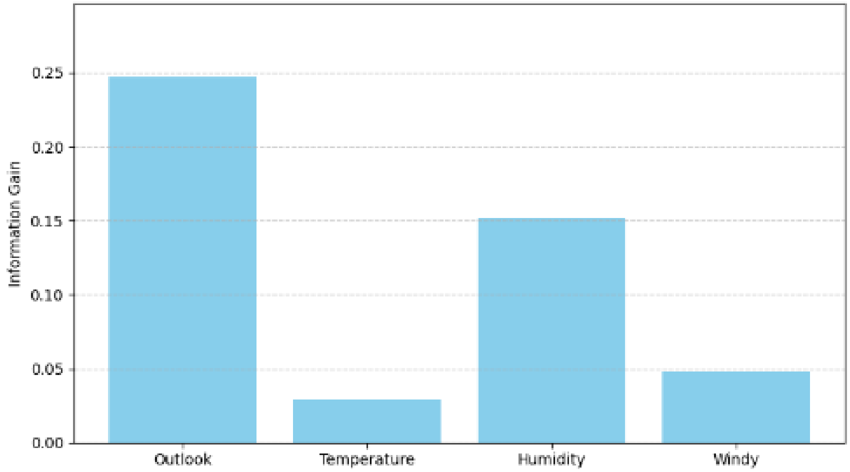

Fig. 1. Information Gain for Each Attribute

This process aligns with the C4.5 algorithm, which recursively splits the data using the attribute with the highest information gain until the entropy reaches zero or further splitting is not possible.

This research effectively illustrates the implementation and validation of the C4.5 decision tree algorithm on the Play Tennis dataset, employing both manual calculations and a Python-based automated approach. By computing entropy and information gain for each attribute, the study identified the most significant features and built the decision tree recursively. The results from the manual process aligned with those generated through Scikit-learn, confirming the correctness of the theoretical method [23].

The C4.5 algorithm demonstrated strong performance in classifying the dataset by prioritizing attributes with the highest gain ratio and effectively managing categorical data. This dual implementation not only verifies the algorithm's accuracy but also deepens the understanding of decision tree construction, particularly for educational or academic contexts [24].

Furthermore, the study includes an overview of recent research involving decision trees in areas such as healthcare, finance, and privacy-focused machine learning. This underscores the adaptability, clarity, and practical value of the algorithm in today's data-driven applications [25].

In summary, the research affirms the capability of the C4.5 algorithm for small-scale classification and lays the groundwork for broader applications in complex, real-world scenarios [26].

4. CONCLUSION

Based on the "Information Gain for Each Attribute" chart, the attribute Outlook demonstrated the highest information gain, approximately 0.25, indicating its dominant role in reducing entropy and improving classification performance in the decision tree. This finding suggests that Outlook is the most informative attribute in the dataset and is therefore selected as the root node during the initial splitting process of the C4.5 algorithm. The attributes were ranked in terms of importance as follows: Outlook, Humidity, Windy, and Temperature, with Temperature contributing the least to the reduction of uncertainty. The results obtained from manual calculations using Microsoft Excel were consistent with those generated through Python's Scikit-learn implementation, thereby validating the accuracy of the theoretical approach. This dual-method analysis confirms that the manual computation of entropy and information gain aligns with the automated implementation, reinforcing the pedagogical value of such an approach in educational and academic settings. Furthermore, the study demonstrates the robustness of the C4.5 algorithm in handling categorical data and constructing effective decision trees for small-scale classification problems. It also highlights the algorithm's adaptability and relevance to broader real-world applications, including in domains such as healthcare, finance, and privacy-aware machine learning.

REFERENCES

[1] N. A. Aziz, A. Manzoor, M. Deedahwar Mazhar Qureshi, M. Atif Qureshi, and W. Rashwan, "Explainable AI in Healthcare: Systematic Review of Clinical Decision Support Systems," 2024, [Online]. Available: https://doi.org/10.1101/2024.08.10.24311735

[2] M. W. Ali, M. Q. Memon, and M. F. Hyder, "Optimization of Fraud Detection Models for Safeguarding Customer Transactions," vol. 5, no. 2, pp. 86–100, 2025.

[3] Y. Mao, B. Legesse, and T. Belsty, "Current Problems in Cardiology Machine learning algorithms for heart disease diagnosis : A systematic review," *Curr. Probl. Cardiol.*, vol. 50, no. 8, p. 103082, 2025, doi: 10.1016/j.cpcardiol.2025.103082.

[4] S. Sreekumari, R. Bhalla, and G. Singh, "Feature Selection and Model Evaluation for Heart Disease Prediction Using Ensemble Methods," *Procedia Comput. Sci.*, vol. 259, pp. 1282–1295, 2025, doi: 10.1016/j.procs.2025.04.083.

[5] G. Dharmarathne, M. Bogahawaththa, U. Rathnayake, and D. P. P. Meddage, "Integrating explainable machine learning and user-centric model for diagnosing cardiovascular disease: A novel approach," *Intell. Syst. with Appl.*, vol. 23, no. August, p. 200428, 2024, doi: 10.1016/j.iswa.2024.200428.

[6] Y. Efe and L. Demir, "The impact of feature selection models on the accuracy of tree-based classification algorithms: Heart disease case," *Procedia Comput. Sci.*, vol. 253, no. 2024, pp. 757–764, 2025, doi: 10.1016/j.procs.2025.01.137.

[7] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. L. Karthikeyan, R. Manikandan, and A. H. Gandomi, "Classification models combined with Boruta feature selection for heart disease prediction," *Informatics Med. Unlocked*, vol. 44, no. December 2023, p. 101442, 2024, doi: 10.1016/j.imu.2023.101442.

[8] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthc. Anal.*, vol. 3, no. July 2022, p. 100130, 2023, doi: 10.1016/j.health.2022.100130.

[9] A. Maulana *et al.*, "Classification of Stunting in Toddlers using Naive Bayes Method and Decision Tree," vol. 1, no. 1, pp. 28–33, 2025.

[10] D. A. Anggoro and D. Permatasari, "Performance Comparison of the Kernels of Support Vector Machine Algorithm for Diabetes Mellitus Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 1, pp. 580–585, 2023, doi: 10.14569/IJACSA.2023.0140163.

[11] M. Alizade-Harakiyan, A. Khodaei, A. Yousefi, H. Zamani, and A. Mesbahi, “Decision tree-based machine learning algorithm for prediction of acute radiation esophagitis,” *Biochem. Biophys. Reports*, vol. 42, no. March, 2025, doi: 10.1016/j.bbrep.2025.101991.

[12] S. Kiran, G. R. Reddy, S. P. Girija, S. Venkatramulu, K. Dorthi, and V. Chandra Shekhar Rao, “A Gradient Boosted Decision Tree with Binary Spotted Hyena Optimizer for cardiovascular disease detection and classification,” *Healthc. Anal.*, vol. 3, no. March, p. 100173, 2023, doi: 10.1016/j.health.2023.100173.

[13] D. Sun, “ScienceDirect Research on Privacy Protection Technology Based on Decision Tree Algorithm,” *Procedia Comput. Sci.*, vol. 262, pp. 1309–1315, 2025, doi: 10.1016/j.procs.2025.05.175.

[14] J. Wu, G. Zhao, M. Wang, Y. Xu, and N. Wang, “Concrete carbonation depth prediction model based on a gradient-boosting decision tree and different metaheuristic algorithms,” *Case Stud. Constr. Mater.*, vol. 21, no. September, p. e03864, 2024, doi: 10.1016/j.cscm.2024.e03864.

[15] M. Qiu, “Path Planning Algorithm and ID3 Decision Tree Model Application of Scenic Intelligent Navigation System,” *Procedia Comput. Sci.*, vol. 247, no. C, pp. 1187–1196, 2023, doi: 10.1016/j.procs.2024.10.143.

[16] A. A. Firdaus, R. A. Faresta, and M. Yunus, “The Role of Sentiment Analysis in Election Predictions Compared to Electability Surveys,” vol. 1, no. 1, pp. 1–8, 2025.

[17] A. Rizky, N. Habibi, I. Sufiyandi, A. K. M. Jayed, and A. M. Nakib, “Diabetes Mellitus Disease Analysis using Support Vector Machines and K-Nearest Neighbor Methods,” vol. 1, no. 1, pp. 22–27, 2025.

[18] R. Hasanah, Z. Hasanah, and S. A. Wijaya, “Play Store Data Scrapping and Preprocessing done as Sentiment Analysis Material,” vol. 1, no. 1, pp. 16–21, 2025.

[19] L. Ji and S. Li, “A dynamic financial risk prediction system for enterprises based on gradient boosting decision tree algorithm,” *Syst. Soft Comput.*, vol. 7, no. December 2024, p. 200189, 2025, doi: 10.1016/j.sasc.2025.200189.

[20] K. Khosravi, A. A. Faroouqe, A. R. Shahvaran, P. Daggupati, S. Heddam, and J. Hatamiafkoueieh, “Beyond conventional modeling: A cutting-edge hybrid IAER-AMT decision-tree-based algorithm for high-resolution river turbidity prediction,” *Ain Shams Eng. J.*, vol. 16, no. 9, p. 103511, 2025, doi: 10.1016/j.asej.2025.103511.

[21] I. Kayan and N. Ayman Oz, “Integrating response surface methodology and decision tree algorithms for valorization of cheese whey wastewater,” *Desalin. Water Treat.*, vol. 322, no. March, p. 101129, 2025, doi: 10.1016/j.dwt.2025.101129.

[22] M. Abidin and M. Muzir, “CLASSIFICATION OF HEART (CARDIOVASCULAR) DISEASE USING THE SVM METHOD,” vol. 1, no. 1, pp. 1–7, 2025.

[23] Z. Zhang and D. Xu, “Enterprise marketing data mining method based on decision tree algorithm,” *Procedia Comput. Sci.*, vol. 261, pp. 1172–1178, 2025, doi: 10.1016/j.procs.2025.04.701.

[24] A. Elhazmi *et al.*, “Machine learning decision tree algorithm role for predicting mortality in critically ill adult COVID-19 patients admitted to the ICU,” *J. Infect. Public Health*, vol. 15, no. 7, pp. 826–834, 2022, doi: 10.1016/j.jiph.2022.06.008.

[25] B. H. De Figueiredo, M. Dos Santos, L. P. L. Fávero, M. Â. L. Moreira, and I. P. De Araújo Costa, “Analysis of maintenance activities in Urban Pavement Management Systems based on Decision Tree Algorithm,” *Procedia Comput. Sci.*, vol. 214, no. C, pp. 712–719, 2022, doi: 10.1016/j.procs.2022.11.233.

[26] A. R. Panhalkar and D. D. Doye, “Optimization of decision trees using modified African buffalo algorithm,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 4763–4772, 2022, doi: 10.1016/j.jksuci.2021.01.011