JURNAL INFORMATIKA
PROXIES

Vol.9, No.1, July, 2025

Implementation Algorithm C4.5 To Find Recommendation From Gym Exercise Data

Tan Franscisco Angga Setiawan[1] *, Hironimus Leong S.Kom., M.Kom.[2]

[1] Teknik Informatika Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata, 21k10018@student.unika.ac.id
[2] Teknik Informatika Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata, marlon.leong@unika.ac.id

Corresponding Author Email: marlon.leong@unika.ac.id

https://doi.org/10.24167/proxies.v9i1.13050

Received: 2025-01-17
Revised: 2025-08-21
Accepted: 2025-10-22
Available online:2025-10-23

Keywords:
C4.5, Decision Tree, Machine Learning

ABSTRACT

Sport has become an important aspect of human health. One of the most common issues that discovered when participating in sports is a lack of knowledge about the sport itself. The purpose of this research is to generate recommendations from implementing the C4.5 algorithm and the Decision Tree. The research will also carry out several methods such as pre-processing consisting of Data Cleaning, Feature Selection, and Data Transformation to deliver the best data results. Data that used in this research is movement data at the gym. The performance of the C4.5 algorithm is determined by performing Validation and Testing, for this case which is include Accuracy, Precision, and Recall. This research will produce recommendations from the implementation of C4.5, the previously mentioned Decision Tree results will be examined so that a recommendation can be made.

1. INTRODUCTION

Gym or fitness has become a crucial focus in people's lives. Many individuals engage in exercise without knowing exactly what they are doing. Thos occurs due to their lack of awareness and insufficient knowledge during training, coupled with a limited background in sports. Exercise also requires essential focus on the muscles being targeted to ensure optimal effectiveness in muscle building and overall body development. Guidance is necessary for performing exercise movements in a structured manner, reducing the risk of serious injuries to the body.

In Machine Learning, there are various methods available to generate relevant results. The C4.5 algorithm is one such method suitable for recommending gym exercises and can be combined with Decision Tree. The data undergoes a process of selecting desired variables. By utilizing Decision Tree, numerous offspring or decision branches can be generated until there are no more possibilities between one data attribute and another. A total of 2917 name of exercise is used in this case. The dataset variable including exercise movements, explanation of these movements, type of the movement, targeted body parts, the tool to be used in exercise ,and intensity level based on the fitness level of the individual. Inaccurate or useless data entries may be removed from the data using pre-processing.

It Should be mentioned that the information gathered needs to match the standards that are in place in order to perform this kind of execution. The first step is to find the relevant data, the data received is the data from the exercise gym. Then, after the data is collected, apply what is known as pre-processing data, which is trying to fix the data so that the analysis process can keep going. After analyzing the data, use tools to create the decision tree and C4.5 Algorithm. Finally, to ensure that the outcome is accurate and validated, train and test it.

Execution precision and validation is also one of the tests that can verify when some execution that working on is running well or not. Precision and validation using Ten-Fold Cross Validation to calculate the accuracy, validation, and recall from the classification of the processed data. With the presence of validation testing, allows you to offer suggestions as required. 2 Gym recommendations are a suitable option to assist individuals in starting their fitness journey and obtaining valuable guidance. By

receiving recommendations, those looking to exercise can avoid confusion and gain clear answers to their fitness needs. Providing recommendations adds sufficient knowledge to engage in fitness activities. It's important to note that exercise can be done anywhere, not just in the gym.

Several studies demonstrate that the C4.5 algorithm consistently shows better accuracy compared to other algorithms in various use cases. Reynara et al. [1] found that C4.5 performs better in accuracy than CART when Grid Search Optimization and SMOTE are not applied. Suryani et al. [5] confirmed that C4.5 achieves a higher accuracy (95.67%) than CART (95.11%) in stroke disease prediction. Sulistiani and Aldino [4] also reported that C4.5 achieved an accuracy, precision, and recall of 87% in predicting scholarship recipients. Similarly, Muslim et al. [8] demonstrated that combining C4.5 with Particle Swarm Optimization improved the algorithm's accuracy by 0.88%. Even when compared to other classifiers, C4.5 often shows competitive results. For instance, Gerhana et al. [6] reported that Naïve Bayes had a slightly higher accuracy than C4.5, but C4.5 was faster in processing time. Additionally, Gunawan et al. [10] highlighted that while K-Nearest Neighbor and Random Forest achieved higher accuracy, C4.5 remained competitive with 72.45% accuracy, surpassing Naïve Bayes with 71.86% accuracy. These findings suggest that C4.5 is highly effective across different domains, especially when enhanced with optimization techniques. Cherfi et al. [2] the use of mean and median speeds up the binarization process. VFC4.5 is more accurate compared to C4.5 and VFDT on large datasets. Myint and Khaung Tin [3] Weighted CART is more accurate compared to weighted C4.5. C5.0 has the highest accuracy because it is an improvement of C4.5. Teknik et al. [7] The C4.5 algorithm with Adaptive Boosting is superior in precision compared to Random Forest and Catboost, but Adaptive Boosting is less suitable for C4.5. Mardi [9] The C4.5 algorithm is effective for analyzing medical record data, but the research is limited to data from residents of Padang only.

2. RESEARCH METHOD

To find the best results, a process needs to be carried out for C4.5 to provide good recommendations :

1. Conduct a literature study related to the topic discussed
2. Collecting gym exercise dataset from Kaggle platform
3. Preprocessing data using data cleaning, feature selection, data transformation
4. Algorithm modeling using C4.5 with decision tree method
5. Displaying recommendation results from the rules tree
6. Analyzing implementation results and making conclusions

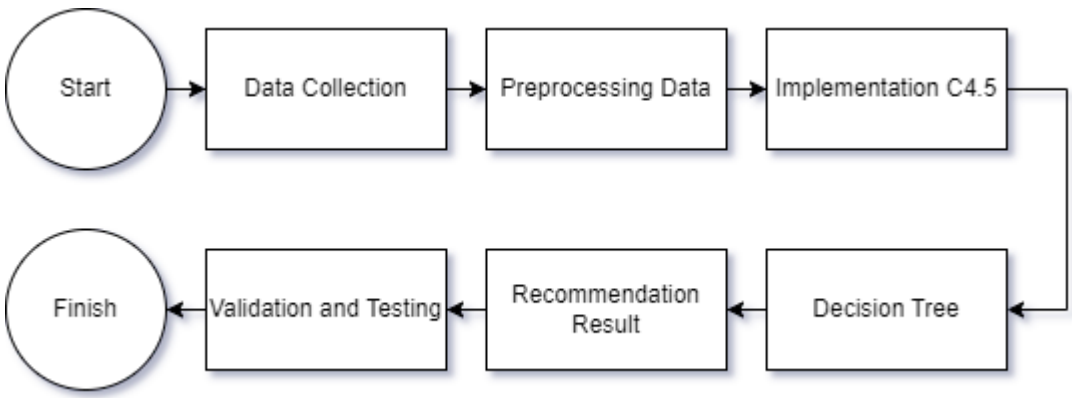

Figure 1. Flow Research.

2.1 Dataset Collection

D Pre-processing Data Pre-processing Data ata Pre-processing Data set taken from Kaggle namely gym exercise data. On that data, contains 8 variable and 2909 data record. Data record about training data i Pre-processing Data n the Gym and some of gym movements.

Table 1. Dataset Variable

No	Name	Information
1	**Title**	**Judul gerakan gym**
2	**Desc**	**Deskripsi gerakan**
3	**Type**	**Tipe gerakan gym**
4	**Equipment**	**Alat gym**
5	**BodyPart**	**Tubuh tujuan**
6	**Level**	**Tingkat gym**
7	**Rating**	**Nilai Gerakan**
8	**Rating Desc**	**Deskripsi Nilai**

2.2 Pre-processing Data

Some preprocessing will be done on the gym exercise data to produce the best results. Data Cleaning, Feature Selection, Data Transformation will be done.

Data Cleaning

Empty value in the dataset will be processed for repair. One way to fill in empty data is by performing data imputation.

Feature Selection

Feature Selection for the implementation is also done in this preprocessing. With the selection of this variable will determine the focus of the data taken.

Data Transformation

Data Transformation is required in this research with the aim of changing the form of data starting from string/ word to numeric by labelling the data like elevation 0, 1, 2 so that it can facilitate the implementation process

2.3 Algorithm C4.5

The Algorithm C4.5 is often combined with the method of using the Decision Tree. So there are some conditions and formulas to be revealed, so that we can produce a Decision Tree.

- Root Node

  First, in making decision tree, determine the root node of the tree, Root is determine by performing the calculation formula of C4.5, calculating the highest Gain value of the selected feature.
- Branches

  The Decision Tree also requires a branches to determine the next decision node. Branches here are determined from the value of a feature.
- Decision Node

  After defining Branches from a feature value. Proceed to selecting the next Node to be the outcome by defining the Gain value as well, but deleting the previous feature with highest gain value.
- Leaf Node

  Leaf Node become the final decision and there will be no any node again. Leaf Node contain value of targeted variable

After performing the above progress, repeat the process until you find the end point using the following formula.

$$Entropy(S) = \sum_{i=1}^{n} - p_i \times log_2 p_i \quad (1)$$

Explanation:

- S : Total Case
- n : Total Partitions on S
- p_i : Ratio of S_i to S

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy\ (S_i) \quad (2)$$

Explanation:

- S : Total Case
- A : Feature
- n : Total Partitions on Feature A
- $|S_i|$: Number of case in i partitions
- $|S|$: Total Case in S

$$SplitInformation(S, A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (3)$$

Explanation :

- S : Total Case

- A : Feature
- S_i : Number of case in i partitions
- n : Total Partitions on Feature A

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (4)$$

Explanation :

- $Gain(S,A)$: Information Gain on Feature A
- $SplitInformation(S,A)$: Split Information on Feature A
- S : Total Case
- A : Feature

2.4 Algorithm CART

CART in here uses a library from Python that is provided with the name library sklearn decisiontreeclassifier, as well as the provided evaluation and testing results. The results provided include accuracy, precision, recall, f1-score, and time processing.

Recommendation Result

In finding this recommendation is the results of a decision tree that is presented for a recommendation that matches the decision that has been made. The results provided are labeled data.

Evaluation

Evaluation and Testing will be the approve that the function is working well or not. Evaluation and Testing shows accuracy, precision, recall, and f1-score for the results of the prediction. Evaluation and Testing will use the Confusion Matrix method that already provided by the python library. Time will also be used as a comparison for the C4.5 and CART.

Results

After implementing the C4.5 modeling, the next step is to process the results of the decision tree rules. Therefore, it is necessary to present the evaluation results after applying the C4.5 algorithm to classify the previously processed data and to present the existing tree rules. Evaluation results such as accuracy, precision, recall, and f1_score will be displayed.

Results C4.5

Several test size ratios were conducted to find the best accuracy results, so they could be implemented in the C4.5 modeling. A 20% Test Size and 80% Train Size were found to yield the best results with the highest accuracy.

Tale 2. C4.5 Modelling Results

C4.5	Precision	Recall	F1 Score	Accuracy	Time
10% Test Size	0.8646	0.8824	0.8667	0.8824	0.0383
20% Test Size	0.8677	0.8841	0.8703	0.8841	0.0365
30% Test Size	0.8354	0.8614	0.8322	0.8614	0.0386

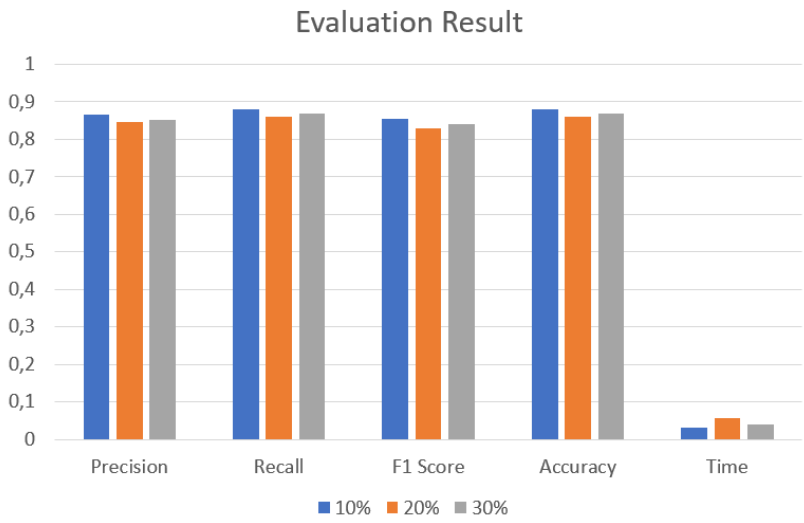

Figure 2. C4.5 Modeling Results

Results Decision Tree

Building a decision tree may be helpful in validating that the achieved results correspond with the expected solution domain in question. The decision tree helps in the understanding of the end results of the model as we are able to view all the steps of attribute selection and data partitioning that are carried out by the model.

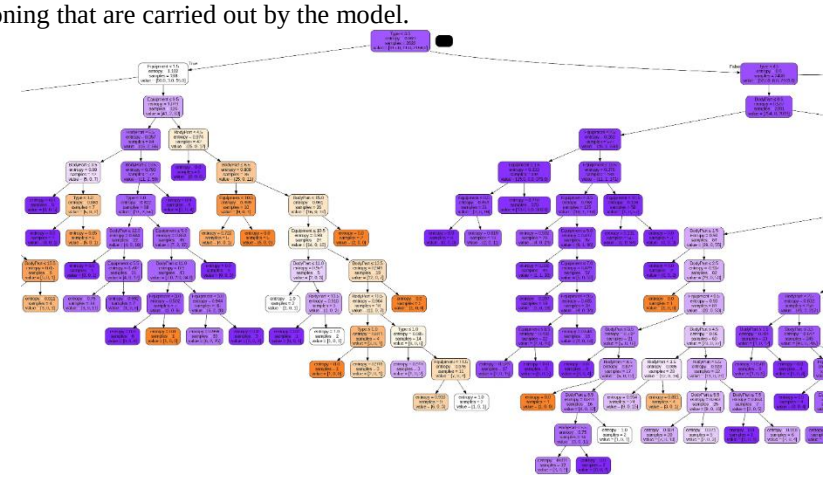

Figure 3. Decision Tree Results

Results Decision Tree Rules

The rules from this decision tree have a good importance for use as a foundation for making the correct recommendations. The rules above are logical in the sense that they are derived from the available data after it has been processed and, therefore, can be helpful in suggesting suitable gym movements depending on different user conditions or needs.

1. **if Type <= 3.5 and Equipment <= 1.5 and BodyPart <= 7.5 and Equipment <= 0.5: Predict: Type 0**
2. **if Type <= 3.5 and Equipment <= 1.5 and BodyPart <= 7.5 and Equipment > 0.5 and Type <= 2.0: Predict: Type 0**
3. **if Type <= 3.5 and Equipment <= 1.5 and BodyPart <= 7.5 and Equipment > 0.5 and Type > 2.0: Predict: Type 2**
4. **if Type <= 3.5 and Equipment <= 1.5 and BodyPart > 7.5 and BodyPart <= 11.5 and Type <= 2.0: Predict: Type 0**

The results here can also be read line by line, for example, taking the earliest rule line.

Discussion

After completing the entire process to find a set of rules/recommendations, there are several things we can discuss from the beginning. First, set up the necessary Library and dataset, Second, proceed to the feature and target selection stage, so it becomes clear what will be predicted in this research. After seeing that there are some data with empty values, several variables were eliminated and not used, because they could ruin the final result. Next, in the preprocessing stage, we use a label encoder. This preprocessing task aims to convert the data to be processed into a more usable form. In this research, the data is transformed into 0, 1, 2,…, which also aims to facilitate the C4.5 modeling to achieve optimal results. It is then followed by the implementation of C4.5 modeling, where three processing stages are carried out to determine the best accuracy. Divided into several ratios, namely 10:90, 20:80, and 30:70 train and test. It was found that the 20:80 ratio had the best evaluation score, so the 20:80 ratio was used.

It was previously mentioned that there would be a comparison between C4.5 and CART. The results from CART show that all accuracy, precision, recall, and f1-score results are the same, but only the 70:30 train test part is different. CART has lower results than C4.5 train test 70:30. C4.5 achieved an accuracy of 86.14% and CART achieved an accuracy of 85.91%. There is a difference of 0.23% between C4.5 and CART. Similarly, the time produced by both algorithms. C4.5 found that the lowest time recorded was 0.0365 seconds in the 80:20 train test, while CART recorded the lowest time of 0.0377 seconds in the 80:20 train test as well. A difference of 0.0012 seconds is also a difference that occurs in processing.

In the end, visualizations were created and the rules from the Decision Tree were presented. In this section, it was also found that the rules from the decision tree can determine recommendations for the necessary gym exercises. Each decision tree pattern was defined and combined into one, resulting in a readable rule. The rules from the decision tree can provide the correct guidance to generate a recommendation.

Conclusion

Towards the end of this research, after processing the C4.5 algorithm by combining it with Decision Tree to find the recommendation results. The processed Gym data can generate existing rules and evaluation results. The evaluation results require the modeling results from the C4.5 Decision Tree, which will be calculated using the confusion matrix formula, and the matrix will appear, thus producing accuracy, precision, recall, and f1-score results. And also in this research, an accuracy comparison algorithm is produced, namely CART. Thus, after obtaining the results from C4.5 and CART, a comparison was made between the accuracy of these two algorithms. It has been found that C4.5 has better results than CART, with an accuracy difference of 0.23% in the 70:30 train-test split, as well as a time difference of 0.0012 seconds from the lowest time of each algorithm. This shows that C4.5 is a better algorithm compared to CART in this research.

Suggestions for further research include adding some data preprocessing to improve the quality of the provided data, Adding more variety data, so that the results provided are also more varied and allowing for more recommendations, More variables on

data can be added to so the decision tree results can be varied. Finding methods to achieve better accuracy, In this research, as a form of recommendation, an adequate display is needed to present the recommendations in a more attractive manner. Therefore, in future research, a display showing the recommendation results can be added, making it a valuable tool for the recommendation system.

REFERENCES

[1] F. J. Reynara, S. Carolina, and I. N. Simbolon, "The Comparison of C4.5 and CART (Classification and Regression Tree) Algorithm in Classification of Occupation for Fresh Graduate," 2022, doi: 10.4108/eai.27-11-2021.2315527.

[2] A. Cherfi, K. Nouira, and A. Ferchichi, "Very Fast C4.5 Decision Tree Algorithm," *Appl. Artif. Intell.*, vol. 32, no. 2, 2018, doi: 10.1080/08839514.2018.1447479.

[3] K. Myint and H. H. Khaung Tin, "Analyzing the Comparison of C4.5, CART and C5.0 Algorithms on Heart Disease Dataset using Decision Tree Method," pp. 1–10, 2021, doi: 10.4108/eai.27-2-2020.2303221.

[4] H. Sulistiani and A. A. Aldino, "Decision Tree C4.5 Algorithm for Tuition Aid Grant Program Classification (Case Study: Department of Information System, Universitas Teknokrat Indonesia)," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 1, 2020, doi: 10.21107/edutic.v7i1.8849.

[5] Suryani, D. Rahmadani, A. A. Muzafar, A. Hamid, R. Annisa, and Mustakim, "Analisis Perbandingan Algoritma C4.5 dan CART untuk Klasifikasi Penyakit Stroke," *SENTIMAS Semin. Nas. Penelit. dan Pengabdi. Masy.*, pp. 197–206, 2022, [Online]. Available: https://journal.irpi.or.id/index.php/sentimas

[6] Y. A. Gerhana, I. Fallah, W. B. Zulfikar, D. S. Maylawati, and M. A. Ramdhani, "Comparison of naive Bayes classifier and C4.5 algorithms in predicting student study period," *J. Phys. Conf. Ser.*, vol. 1280, no. 2, 2019, doi: 10.1088/1742-6596/1280/2/022022.

[7] J. Teknik, I. C. I. T. Medicom, W. Leonardi, H. Weide, D. Cantona, and G. M. Hutagalung, "Comparison of data mining algorithms (random forest , C4 . 5 , catboost) based on adaptive boosting in predicting diabetes mellitus," vol. 16, no. 1, pp. 1–12, 2024.

[8] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, "Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis," *J. Phys. Conf. Ser.*, vol. 983, no. 1, 2018, doi: 10.1088/1742-6596/983/1/012063.

[9] Y. Mardi, "Data Mining Rekam Medis Untuk Menentukan Penyakit Terbanyak Menggunakan Decision Tree C4.5," *J. Sains dan Inform.*, vol. 4, no. 1, pp. 40–53, 2018, doi: 10.22216/jsi.v4i1.3077.

[10] Gunawan, Hanes, and Catherine, "C4.5, K-Nearest Neighbor, Naïve Bayes and Random Forest Algorithms Comparison to Predict Students' On Time Graduation," *Indones. J. Artif. Intell. Data Min.*, vol. 4, no. 2, pp. 62–71, 2021, [Online]. Available: http://dx.doi.org/10.24014/ijaidm.v4i2.10833.