JOURNAL OF DATA SCIENCE | Vol.2023:14
eISSN:2805-5160

Application of Decision Trees in Athlete Selection: A Cart Algorithm Approach

Riska Wahyu Romadhonia[1*], A'yunin Sofro[1], Danang Ariyanto[1], Dimas Avian Maulana[1], Junaidi Budi Prihanto[2]

[1]Department of Mathematics, Universitas Negeri Surabaya, Indonesia
[2]Department of Physical Education, Universitas Negeri Surabaya, Indonesia

[*]**Email:** riskaromadhonia@unesa.ac.id

Abstract

This study investigates the application of Decision Trees (DTs), a non-parametric supervised learning method, renowned for its simplicity, interpretability, and wide applicability in various domains, including machine learning for classification and regression tasks. The focus of this study is on the use of DTs, employing the Classification and Regression Trees (CART) algorithm, in the initial screening of athletes. This involves analyzing 11 sociodemographic and anthropometric variables within a dataset of 113 prospective athletes, encompassing both numerical and categorical data. The DT model exhibits outstanding performance, achieving accuracy and precision rates exceeding 0.8. Further analysis, varying impurity criteria and tree depths, indicates that the Gini index at a depth of 3 optimizes accuracy. Notably, weight, and Body Mass Index (BMI) exhibit the highest significance among the other variables. Looking ahead, future research could explore enhancing DTs' predictive capabilities in athlete selection by incorporating more variables or employing ensemble learning techniques. This study lays the groundwork for further investigations aiming to refine athlete screening processes and broaden the utility of DTs in sports-related predictive modeling.

Keywords

Decision Tree, Athlete Screening, CART, Sociodemographic Data, Anthropometric Data

Introduction

Decision Trees (DTs) represent a versatile and non-parametric supervised learning method extensively utilized in machine learning for both classification and regression tasks. Their forte lies in crafting predictive models grounded in uncomplicated decision rules extracted from data attributes, thereby ensuring lucid interpretability and obviating the necessity for feature scaling (Ochiai, Masuma, Tomii, 2019; Ceballos, 2019). DTs proficiently categorize data and forecast values within distinct segments, proving their efficacy in diverse scenarios.

Despite the merits of Decision Trees, challenges prevailed in predictive modeling, particularly concerning interpretability and adaptability across diverse datasets. Traditional statistical methods grappled with complex decision-making processes, struggling to effectively

Submission: 17 November 2023; **Acceptance:** 30 November 2023

handle both categorical and numerical data simultaneously (Topirceanu, Grosseck, 2017). Deriving decision rules from intricate data attributes posed a significant impediment. Existing models lacked transparency in their decision-making, hindering comprehensive comprehension and explanation of predictive outcomes. Additionally, the necessity for feature scaling in certain methodologies introduced complexities, mandating extensive preprocessing steps that were not universally applicable.

Amidst the evolution of various algorithms, such as C45, CART (Classification and Regression Trees), and CHAID, each algorithm has embraced distinctive approaches to construct decision trees revolving around pivotal attributes (Pedregosa et al., 2011). Notably, the CART algorithm has garnered renown for its binary splitting methodology, effectively partitioning data into two subsets by discerning the optimal split among all variables. This distinguishing feature endows CART with widespread applicability and efficiency in navigating complex datasets. CART stood out for its seamless handling of categorical and numerical data, offering a distinct advantage over traditional methods (Pedregosa et al., 2019). Its binary splitting approach, employing the best possible splits among variables to segment data, not only addressed decision-making complexities but also provided a transparent and interpretable tree structure.

This study applies DTs, specifically the CART algorithm, to the initial screening of athletes, utilizing 11 sociodemographic and anthropometric variables. By constructing a model that generates rules to predict athlete eligibility during preliminary screening, this research highlights the practical utility and adaptability of decision trees in real-world scenarios, emphasizing their importance in diverse applications.

Methodology

Data Collection

This study utilizes a set of feature variables consisting of four anthropometric data, which include physical and morphological measurements, as well as seven sociodemographic data that describe the characteristics of prospective athletes. These variables are detailed in Table 1. Meanwhile, for the class labels in this study represent the outcomes of prospective athletes in the screening process, which fail and pass. The data used for analysis were collected from 113 prospective athletes who underwent the selection process at Universitas Negeri Surabaya.

Table 1. Detailed Antropometric and Sociodemographic data

Data	Variabel	Type
Anthropometric	Height	Numerical
	Weight	Numerical
	Body Mass Index (BMI)	Numerical
	Waist	Numerical
Sociodemographic	Age	Numerical
	Gender	Categorical
	Last Education of Father	Categorical

Data	Variabel	Type
	Last Education of Mother	Categorical
	Occupation of Father	Categorical
	Occupation of Mother	Categorical
	Finance Classification	Categorical

Data Prepocessing

According to Table 1, it has been identified that the dataset comprises six categorical variables. To facilitate further analysis, a preprocessing stage is needed, as library skitlearn only able to process numerical data. This preprocessing step involves the conversion of these categorical variables into numeric representations. The assignment of numerical values needed, ranging from 0 to n, in accordance with the data variations observed within each categorical variable. Detailed outcomes of this preprocessing is shown in Table 2.

Table 2. Result of Preprocessing

Variabel	Before	After
Gender	Female	0
	Male	1
Last Education of Father	High School or lesser	0
	Bachelor or higher	1
Last Education of Mother	High School or lesser	0
	Bachelor or higher	1
Occupation of Father and Occupation of Mother	Unemployed / deceased	1
	Civil Servant	2
	Private Employee	3
	Entrepreneur	4
	Farmer / Cultivator	5
	Laborer	6
	Others	7
Finance Classification (Total Parent Salary)	Less than 3 millions rupiah	0
	Between 3 – 6 millions rupiah	1
	Greater than 6 millions rupiah	2

Utilizing CART method

The Classification and Regression Trees (CART) methodology represents a versatile Decision Tree algorithm designed for both classification and regression tasks. CART's mechanism involves constructing a binary tree where each node signifies a decision based on one of the input

features, branching from a single root into two child nodes at each junction. This binary structure streamlines the decision-making process, enhancing computational efficiency. In CART, the selection of a feature and its corresponding split point is based on achieving the maximum information gain (IG), guided by criteria like the Gini index or entropy. In a general sense, information gain can be defined as:

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right}) \quad (1)$$

where f represent features split on, D represent the dataset, I represent the impurity criterion, and N represent the number of samples

In CART, the selection of features and split points aims to maximize information gain (IG). This selection process is guided by criteria such as the Gini index or entropy. The Gini index measures impurity within decision tree nodes, quantifying it by subtracting the sum of squared class proportions from 1. On the other hand, entropy assesses the disorder or randomness in class distribution within nodes. Lower value indices signify more pure nodes and are preferred for splitting in decision trees.

The equation for the Gini index:

$$I_G = 1 - \sum_{j=1}^{c} p_j^2 \quad (2)$$

The equation for entropy:

$$I_H = -\sum_{j=1}^{c} p_j \log_2(p_j) \quad (3)$$

where p_j is the porportion of the samples that belongs to class c for a particular node.

Evaluating the model

A comprehensive evaluation of the constructed decision tree involves assessing key metrics: accuracy, precision, recall, and the F1 score. Accuracy measures the overall correctness of predictions made by a classification model. Precision quantifies the model's ability to correctly classify positive instances. It measures the ratio of true positive predictions to the total number of instances predicted as positive. Recall measures the model's ability to identify all relevant positive instances. Finally, the F1 score is the harmonic mean of precision and recall. It provides a balanced evaluation of a model's performance, especially when there's an imbalance between the positive and negative classes. These metrics play a crucial role in assessing the performance of a classification model, each focusing on different aspects of correctness, positive instance classification, and the trade-off between precision and recall.

Results and Discussion

This model simulation involves splitting the data samples into two sets: 70% for training and 30% for testing. Subsequently, testing is conducted using the Gini index and Entropy as impurity criteria for various maximum depth values in the decision tree. The results of these tests are shown in Table 3, which includes the accuracy values obtained for each test. Additionally, the simulation outcomes are graphically illustrated in Figure 1 for enhanced interpretability.

Table 3. Evaluation Result for different depth

Depth	Accuracy (%)	
	Gini	Entropi
2	91.18	91.18
3	91.18	85.30
4	85.30	85.30
5	85.30	82.35
max	85.30	82.35

From the graphical analysis, it becomes evident that the Gini index as an impurity measure leads to higher accuracy levels compared to Entropy. Notably, the optimal accuracy is achieved when the tree depth is set at 2 and 3. However, at a depth of 2, although both Gini index and Entropy reach peak accuracies, the decision tree lacks the complexity to effectively differentiate between classes. Therefore, a depth of 3 is considered more suitable for constructing a decision tree that is both representative and precise, with the Gini index serving as the preferred criterion for minimizing impurity. This approach ensures a more accurate and reliable predictive model.

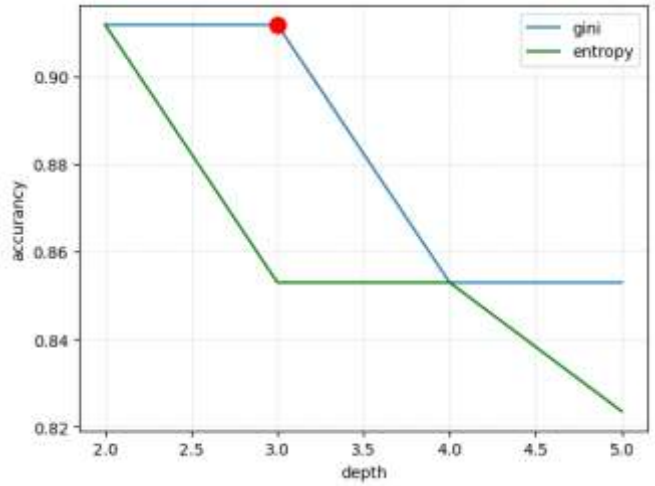

Figure 1. Evaluation Result for different criteria at different depth. Red circle represent the chosen value for next simulation.

Now we can proceed with the comprehensive simulation of the decision tree, following these guidelines: using 70% of the data for training and 30% for testing, taking the Gini index as the impurity criterion, and setting the maximum depth to 3. The resultant decision tree model, visualized in Figure 2 using Python library, was subject to thorough analysis, revealing crucial insights into the classification outcomes. The root node was selected based on the BMI < 30.56 criterion, possessing a Gini Index value of 0.14. This node initially partitioned 79 samples into two categories: 73 “fail” and 6 "pass."

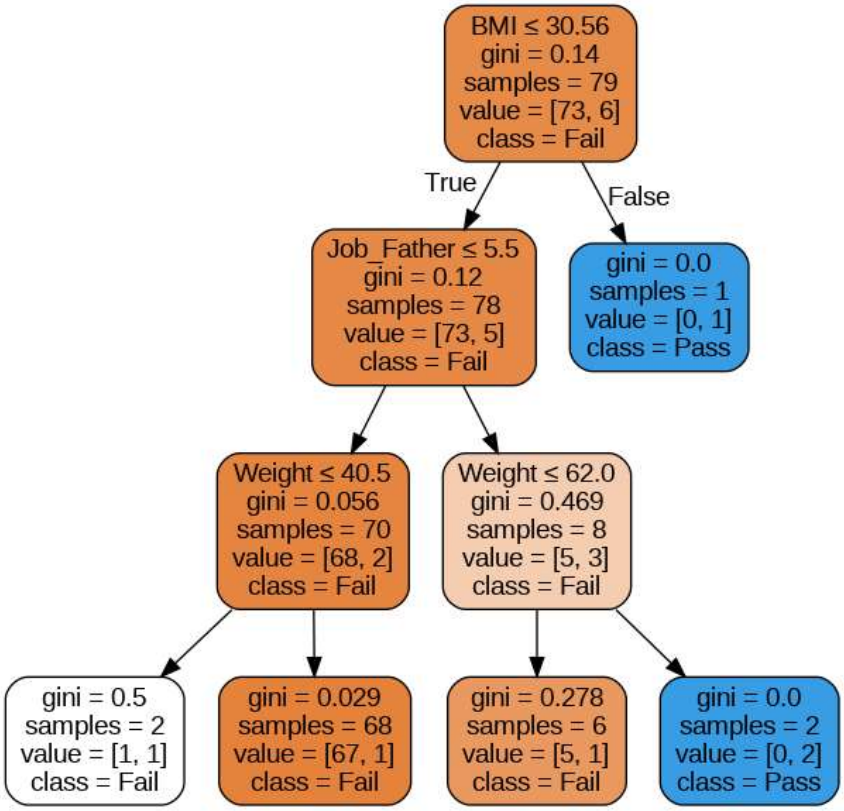

Figure 2. Decision Tree of dataset

Initially, the first branch leaned towards the “pass” class, where the right branch (False) showed a Gini Index of 0.0, containing only one “pass” sample. In contrast, the left branch (True) further split its samples based on the father's occupation (≤ 5.5). This division resulted in varied classification outcomes: the right sub-branch (False) displayed 8 samples, comprising 5 “fail” and 3 “pass” samples. Meanwhile, the left sub-branch (True) consisted of 70 samples, with 68 “fail” and 2 “pass” samples. These sub-branches were subsequently divided based on weight.

Moving to the second level branches, the right branch exhibited a Gini Index of 0.469, dividing further based on weight (≤ 62.0), while the left branch showed a Gini Index of 0.056, dividing based on weight (≤ 40.5). Eventually, at the third level, four leaf nodes emerged: three belonging to the “fail” class with Gini indices of 0.5, 0.029, and 0.278, respectively, and one leaf representing the “pass” class with a Gini Index of 0.0.

Analysis of each branch revealed diverse classification outcomes. Some branches displayed low Gini Index values, indicating purer classifications and higher accuracy, while others exhibited greater uncertainty. This underscores the influence of specific features or conditions in precisely classifying samples and the varied uncertainty levels in their final decisions.

After executing this simulation, we have also computed four evaluation parameters: accuracy, precision, recall, and F1 score. The results of these evaluations are in Table 4. These evaluation values demonstrate good consistency and affirm that the decision tree model has performed effectively in the prediction task on the utilized dataset.

Table 4. Four parameter evaluation

Accuracy	Precision	Recall	F1
91.18%	90.04%	89.75%	92.39%

Furthermore, from the 11 inputs used, we can identify the aspects that are most important in the formation of the decision tree model. The top five aspects with the highest importance are weight, BMI, occupation of father (Job_Father), height, and waist. However, it should be noted from Table 5 that the variables height and waist have an importance percentage of 0.0. This does not mean that these two variables are entirely unimportant for prediction. Instead, it should be understood that in the context of this decision tree model, these variables may not be directly used for splitting at certain levels because other attributes are more informative in class separation. In other words, although they have an importance of 0.0 in splitting at a particular level, the variables height and waist can still provide valuable information at higher levels in tree or in combination with other attributes.

Table 5. Highest Importances of attributes

Attributes	**Importance**
Weight	0.4648
BMI	0.2681
Occupation of Father	0.2671
Height	0.0
Waist	0.0

Conclusion

The construction of a decision tree model, based on a dataset encompassing 113 prospective athlete profiles, has unveiled essential insights into the criteria driving athlete selection. Rooted in the criterion of BMI < 30.56, and in the end categorized five distinct leaf nodes. Among these, three nodes classified athletes as “fail”, while the remaining two nodes categorized athletes as “pass”. This model showcased remarkable performance metrics, boasting accuracy and precision exceeding 0.8, establishing the Gini index as the paramount criterion, particularly at a depth of 3, for achieving optimal accuracy. Emphasizing the significance of weight, BMI, and the father's occupation, the model underscored their substantial influence on athlete selection while acknowledging the distinctive contributions of the remaining eight variables in fine-tuning prediction accuracy. However, areas for refinement persist, primarily concerning the handling of categorical variables and the implementation of advanced feature selection techniques to further enhance predictive accuracy. In summary, while this study demonstrates the efficacy of a decision tree model in athlete selection, opportunities for refinement exist. Future research may focus on refining categorical variable handling and utilizing advanced feature selection techniques to further improve predictive performance.

Acknowledgements

We express our profound gratitude to the Directorate of Research and Community Service, Ministry of Education, Culture, Research, and Technology (Kemendikbudristek), Indonesia, for their financial support through the 2023 research grant (Grant No. SK 1151/UN38/HK/PP/2023). Our sincere appreciation goes to all contributors for their support and the opportunity to advance our research goals.

References

Agresti, A. (1996). An introduction to categorical data analysis. New York: Wiley.

Ceballos, F. (2019). Scikit-Learn Decision Trees Explained - Training, Visualizing, and Making Predictions with Decision Trees.

Breiman L, Friedman J, Olshen R and Stone C J. (1984). Classification and Regression Trees Monterey CA: Wadsworth and Brooks.

Daniya, T., Geetha, M., & Kumar, K. S. (2020). Classification and regression trees with gini index. Advances in Mathematics: Scientific Journal, 9(10)

Ellson, J., Gansner, E., Koutsofios, L., North, S., & Woodhull, G. (2001). Graphviz - Open Source Graph Drawing Tools. In Lecture Notes in Computer Science (pp. 483–484). Springer-Verlag.

James G., Witten D., Hastie T., Tibshirani R. (2013). An Introduction to Statistical Learning, 1st ed., Springer, New York.

Karpowicz, E., & Kaska, A. (2019). Hipertensi on in Athletes. Advances in Experimental Medicine and Biology, 1204, 231-237.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., & Vlahavas, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116

Lemon SC, Roy J, Clark MA, et al (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med, 26(3): 172–181

Loh, W. (2011). Classification and Regression Trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

Lim, S. S., Vos, T., Flaxman, A. D., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet, 380(9859), 2224-2260

Liu, Y. Gastwirth, J.L. (2020). On the Capacity of the Gini Index to Represent Income Distributions. METRON., 78, 61–69

Muhammad, A., Tanveer, Z., Ali khan S. (2020). Decision Tree Classification. Ranking Journals using IGIDI, J Inform Sci., 46(2), 325–339.

Ochiai, Y., Masuma, Y., & Tomii, N. (2019). Improvement of timetable robustness by analysis of drivers' operation based on decision trees. Journal of Rail Transport Planning & Management, 9(March), 57– 65

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Phython. Journal of Machine Learning Research, 12, 2825-2830

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2019). Decision Tree. Journal of Machine Learning Research

Prihanto, Junaidi B., Endang S.W., Faridha N., Ryota M., Miwako T., and Masayuki K. (2021). Health Literacy, Health Behaviors, and Body Mass Index Impacts on Quality of Life: Cross-Sectional Study of University Students in Surabaya, Indonesia. International Journal of Environmental Research and Public Health 18, no. 24: 13132

Retnani L., Emi S.W., Priadhana E.K. (2019) Model Decision Tree untuk Prediksi Jadwal Kerja menggunakan Scikit-Learn.Jurnal UMJ.

Topîrceanu, A., & Grosseck, G. (2017). Decision tree learning used for the classification of student archetypes in online courses. Procedia Computer Science, 112, 51–60.

Wang, J., Perez-R., M. M., Gutiérrez, P. A. (2018). Machine learning in clinical medicine: Current status and future directions. European Journal of Internal Medicine, 52, 16-24