International Journal Software Engineering and Computer Science (IJSECS)
5 (3), 2025, 988-997
Published Online December 2025 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v5i3.5242.

RESEARCH ARTICLE Open Access

Decision Tree-Based Potential Athletics Athlete Selection System for PASI DKI Jakarta

Sugiyono
Informatics Engineering Study Program, Faculty of Computer Technology, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, East Jakarta City, Special Capital Region of Jakarta, Indonesia.
Email: inosoguy007@gmail.com.

Arpinda *
Informatics Engineering Study Program, Faculty of Computer Technology, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, East Jakarta City, Special Capital Region of Jakarta, Indonesia.
Corresponding Email: arpinda27@gmail.com.

Received: August 3, 2025; Accepted: September 5, 2025; Published: December 1, 2025.

Abstract: Selection of athletes in competitive sports is mostly based on subjective judgments; therefore, it results in inconsistency. This research presents a classification model that will help to measure the potential of athletes using the Decision Tree algorithm by utilizing real competition data from PASI DKI Jakarta. The dataset used consists of 450 records of athletes with attributes such as race category, time records, and ranking information. The analysis was performed based on the CRISP-DM framework which comprises six stages: business understanding, data exploration, preparation, modeling, evaluation, and deployment. Development and testing of the model were carried out in RapidMiner software using a 10-fold cross-validation technique. It achieved an accuracy of classification equal to 92.22% with a standard deviation of ±5.37%. The performance metrics show precision rates at 96.88% for High, 78.95% for Medium, and 94.87% for Low classes; while recall values are 100%, 88.24%, and 88.10%, respectively. The decision tree model generated specifies ranking as the root node meaning that this attribute has the highest influence on class separation among other attributes in this dataset. There are three classification rules produced by this model: ranking ≤3.500 is classified into high potential; between 3.500-6.500 belongs to medium potential; otherwise greater than 6.500 will be classified into low potential which can be applied practically as a decision support system enabling coaches to perform objective systematic data-driven processes in selecting athletes.

Keywords: Decision Tree; Classification; Athletics; RapidMiner; CRISP-DM.

1. Introduction

Athlete selection is a very important part of developing competitive sports. It is the base for training programs and long-term plans in many different sports. This process not only decides who is on a team at the moment but also has an effect on how resources are used, coaching methods, and finally the success path of sports organizations. However, current selection practices in many athletic programs are still mostly manual and subjective, relying heavily on coach intuition and observational assessments. While these approaches draw on valuable experiential knowledge, they carry inherent risks of inconsistency and cognitive bias that can compromise the identification of genuine talent. Traditional methods often struggle to objectively quantify athlete potential, particularly when evaluating large candidate pools or comparing performance across different competition scenarios. The subjective nature of assessments can lead to overlooked talent, misallocated resources, and reduced program effectiveness, ultimately hindering the development of competitive athletes at both regional and national levels.

Digital transformation across industries has made data analytics and technology strategic solutions for improving the quality of decision-making, with sports science being a highly promising application domain. Data mining techniques provide systematic approaches to extract actionable intelligence from historical performance records, competition results, and training data. Among the various analytical methods used, classification algorithms are gaining popularity because they can classify athletes based on measurable attributes and predict their future performance trajectories. Kurniawan *et al.* (2024) proved that data mining applications in sports can analyze athlete performance patterns effectively, thereby providing an objective basis for coaching decisions and strategies related to talent development [1]. The Decision Tree algorithm is one of the most popular classification techniques since it helps create logical structures based on data through interpretable rule sets. This makes the technique very attractive in a sports environment where coaches and administrators (who may not have a technical background) need to understand and trust analytical outputs. Interpretability advantages address a major barrier for technology adoption in sports settings since stakeholder buy-in depends on transparent reasoning processes. Predictive modeling approaches have been extremely successful in predicting athlete performance outcomes; frameworks established by Qin *et al.* (2025) utilize competition data to generate reliable predictions of performances [2]. The integration of information flow analysis, as discussed by Stenger and Feng (2024), shows that systematic data collection and processing can improve efforts related to both athlete health monitoring and performance optimization [3]. Recent advancements in machine learning have further increased capabilities whereby hybrid models comprising multiple algorithms have proven more accurate in identifying sports talents; this was evidenced by Li *et al.* (2025), who reported considerable gains using Decision Tree and Random Forest ensemble approaches [4].

The Indonesian Athletics Association (PASI) DKI Jakarta is the provincial governing body for track and field athletics. One of its ongoing challenges is finding and developing potential athletes from the early to competitive levels. This organization is a regional hub for athletic talent in Indonesia's capital, managing numerous athletes across various age groups and event categories. Therefore, efficient and effective selection mechanisms are required to optimize coaching resources and training facilities. Although there exists rich historical data from regular competitions—such as recorded times, rankings, and event-specific performance metrics—the current selection processes have not yet fully leveraged systematic data analysis approaches. The organization still uses traditional observation-based methods primarily, thus leaving valuable performance data unutilized in decision-making workflows. Competition results particularly attributes like recorded times and final rankings contain significant predictive information regarding athlete potential which could inform more objective classification systems. Quantifiable performance indicators present an ideal foundation for data-driven talent assessment because they provide consistent comparable measures among different athletes under varying competition scenarios. The gap between available data resources and their practical application in selection processes represents a missed opportunity for PASI DKI Jakarta to enhance program effectiveness as well as competitive outcomes.

Because of the shown ability of data mining in sports uses and the special needs of PASI DKI Jakarta, the research makes a classification model for athlete potential assessment using the Decision Tree algorithm with inputs taken from real athletic competition data collected within the program. The study uses CRISP-DM (Cross Industry Standard Process for Data Mining) methodology as a structured framework that guides the analytical workflow from business understanding to model deployment. The RapidMiner software is used as the main analytical tool for data preprocessing, model building, and performance testing. This research is limited to classifying athletes into three potential categories based on competition performance attributes like event category, recorded times, and rankings; high medium low. Clear classification rules based on historical performance data will be set by this model so that coaches and administrators have an objective transparent decision support system. It is expected that not only will selection accuracy improve but also consistency in talent identification processes reduce subjective bias and more efficiently allocate coaching resources toward athletes with demonstrated potential. The data-driven approach seeks to change athlete selection at PASI DKI Jakarta from intuition-based practice into systematic evidence-based process continuously refined as more performance data builds up over time.

2. Related Work

In recent times, there has been an increasing trend in the use of machine learning algorithms for sports analytics. This trend has seen researchers try out different methods of assessing athletes, predicting their performances, and identifying talent. Decision Tree-based methods have come out as very popular options because they are easy to interpret and work well with organized data on athletic performance. Several studies have proven successful applications across various sport disciplines, each offering different insights on data selection, algorithm optimization, and practical deployment strategies. In the area of classifying athletes, Kuswanto *et al.* (2024) looked at using the C4.5 algorithm to classify levels of achievement among athletes based on competition records and training performance metrics as main input features [5]. Their study showed that C4.5 could efficiently classify athletes into different performance levels while providing understandable decision rules that coaches can easily comprehend and implement in real-life selection situations. Following similar ideas, Romadhonia *et al.* (2023) used CART (Classification and Regression Trees) for athlete selection processes and compared it with other classification techniques [6]. Their work found out that CART was better than other algorithms in terms of accuracy and model simplicity by generating more compact decision trees with fewer decision nodes. This comparison gave insights into choosing algorithms for sports analytics applications, especially when one needs to balance predictive power with the requirement for model interpretability.

Zhang *et al.* (2024), who focused on track and field athletics in particular, created a Decision Tree-based system for predicting performance that included trends in temporal performance along with competition results [7]. Their model was based on multi-season data from competitive athletes where patterns of progression could separate those with high potential for improvement from others who had reached a plateau in performance. The methodology here introduced feature engineering techniques capturing metrics of velocity and consistency to offer a more nuanced approach toward athlete evaluation. Further expanding beyond track and field was Fachrezzy *et al.* (2025), who applied data mining using the C4.5 algorithm for squash athlete selection from a dataset comprising physical fitness test results, technical skill assessments, and match performance statistics [8]. Their study brought out serve accuracy as well as court movement speed as the most discriminative attributes for talent identification in racquet sports; this cross-sport application demonstrated how adaptable Decision Tree methods are across different athletic disciplines with unique performance indicators and evaluation criteria each time. Algorithm optimization is another important direction of research in the application of sports analytics. Nugraha and Putra (2023) studied Decision Tree optimization through pruning techniques used for athletic potential prediction, which is meant to overcome the common problem of model overfitting when dealing with relatively small athlete datasets [9]. Their study compared pre-pruning and post-pruning strategies and found that cost-complexity pruning helped improve generalization performance while decreasing model complexity. This approach to optimization would be very useful for a sports organization that has little historical data because then it can make reliable predictions even with limited sample sizes. Hartati and Priyanto (2023) used the CRISP-DM methodology in football athlete achievement prediction by using the C4.5 algorithm [10]. They followed all six CRISP-DM phases from business understanding to deployment, which will create a workflow that can be repeated by other organizations in sports for their own purposes. The study pointed out how important structured data mining methodologies are for making sure the analysis is rigorous and helping knowledge transfer between different projects in sports analytics.

Recent studies have broadened the scope of athletic performance prediction to include different data sources and advanced machine learning methods. Wibowo *et al.* (2024) looked into supervised machine learning methods for predicting physical fitness, comparing Decision Trees with Support Vector Machines and Neural Networks based on physiological measurements and results from fitness tests [11]. They found that even though Neural Networks achieved slightly better accuracy, Decision Trees offered a level of interpretability that fitness coaches appreciated much more in real-world applications. This study highlighted the ongoing conflict between predictive performance and model transparency in applied sports science settings. Miah *et al.* (2023) examined the potential of using mobile health data to predict athletics fitness levels, using measurements from wearable devices such as heart rate variability, sleep patterns, and daily activity levels [12]. Their machine learning models proved that continuous monitoring data could be an addition to traditional competition-based assessments. The combination of health monitoring data opened new possibilities for athlete evaluation systems that consider training load, recovery status, and injury risk factors holistically. Wang *et al.* (2025) pushed methodological boundaries by merging traditional Decision Tree approaches—specifically ID3—with deep learning techniques for analyzing sports training data [13]. Their hybrid architecture used Decision Trees for initial feature selection and rule extraction before feeding selected features into deep neural networks for final predictions. The research indicated that combining interpretable tree-based methods with powerful deep learning models could provide optimal solutions for sports analytics applications by balancing the needs of transparency requirements with those of predictive performance demands.

3. Research Method

This research employs a quantitative approach using classification-based data mining methods, specifically the Decision Tree algorithm (C4.5), to classify athlete potential based on competition performance data. The research workflow follows the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, a widely adopted methodology in data mining projects that ensures systematic and reproducible analytical processes. The CRISP-DM framework has been successfully applied across various domains, including credit risk prediction as demonstrated by Putri *et al.* (2023), who validated its effectiveness in structured classification tasks [14]. Similarly, Nugroho and Ramadhan (2024) employed CRISP-DM for customer churn prediction using Decision Tree algorithms, establishing best practices for implementing the methodology in classification scenarios [15]. The framework consists of six sequential phases that guide the research from problem formulation through model deployment, ensuring rigorous analysis and practical applicability of results. Figure 1 illustrates the complete CRISP-DM workflow applied in this research, showing the iterative relationships between phases and the cyclical nature of the data mining process. Figure 2 displays the classification process implemented in RapidMiner software, depicting the operator chain from data retrieval through model evaluation. The visual representation clarifies the technical implementation of the methodology and facilitates reproducibility of the research procedures.

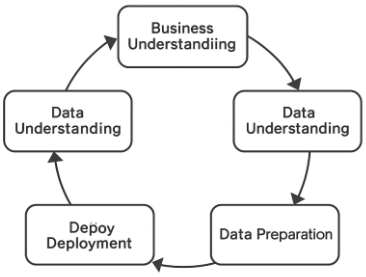


Figure 1. CRISP-DM Framework Workflow for Athlete Classification

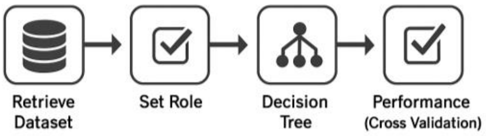


Figure 2. RapidMiner Process Flow for Decision Tree Classification Model

3.1 Business Understanding

The initial phase focuses on understanding organizational needs and defining research objectives aligned with stakeholder requirements. For PASI DKI Jakarta, the primary business objective involves establishing a data-driven athlete selection system that reduces subjective bias and improves decision consistency. The specific goal is to classify athletes into three potential categories—High, Medium, and Low—based on competition performance metrics, particularly final rankings achieved in athletic events. Rahmawati *et al.* (2023) emphasized the importance of thorough business understanding in CRISP-DM implementations, noting that clear objective definition directly influences subsequent analytical decisions and model utility [16]. The classification system aims to support coaches in identifying promising athletes for advanced training programs, optimizing resource allocation, and establishing transparent selection criteria. Success criteria for the project include achieving classification accuracy above 85%, generating interpretable decision rules that coaches can understand without technical expertise, and creating a deployable system that integrates smoothly into existing selection workflows.

3.2 Data Understanding

Data collection involved gathering competition records from PASI DKI Jakarta athletic events and conducting structured interviews with experienced coaches to validate data quality and relevance. The dataset comprises 450 athlete records collected over multiple competition seasons, capturing performance across various track and field events. Each record contains five primary attributes: athlete name (identifier), gender, event category (sprint, middle-distance, long-distance, field events), recorded time in seconds, and final ranking position. The target variable for classification is athlete potential category, assigned based on ranking thresholds established through coach consultations: High potential (rankings 1-3), Medium potential (rankings 4-7), and Low potential (rankings 8 and below). Exploratory data analysis revealed that recorded time and ranking showed strong correlation with potential categories, while event category provided necessary segmentation since performance standards vary significantly across different athletic disciplines. Initial data profiling identified 23 records with missing values in the recorded time attribute and 8 duplicate entries that required handling during preparation phases. Statistical summaries confirmed reasonable data distributions without extreme outliers that might distort model training, though some events showed natural performance variance requiring careful feature engineering.

3.3 Data Preparation

Data preparation involved multiple preprocessing steps to ensure dataset quality and suitability for Decision Tree modeling. The process began with duplicate removal, eliminating 8 redundant records identified through exact matching across all attributes. Missing value treatment addressed 23 records lacking recorded time data; these were handled through deletion rather than imputation since time records represent objective measurements that cannot be reliably estimated. The athlete name attribute was removed from the modeling dataset as it serves only identification purposes without predictive value for classification tasks. Gender was retained initially for exploratory analysis but ultimately excluded from the final feature set after preliminary modeling showed minimal contribution to classification accuracy. The final prepared dataset contained 419 valid records with three input features: event category (nominal variable with 8 distinct values), recorded time in seconds (continuous numerical variable), and ranking position (discrete numerical variable ranging from 1 to 15). The target variable, potential category, was encoded as a nominal attribute with three class labels: High, Medium, and Low. Data transformation included standardizing event category names to ensure consistency and converting all numerical values to appropriate data types for RapidMiner processing. The cleaned dataset was exported in CSV format with proper encoding to prevent character recognition issues during import into the analytical platform.

3.4 Modeling

Model construction utilized the C4.5 Decision Tree algorithm implemented through RapidMiner Studio, selected for its proven effectiveness in generating interpretable classification rules from structured data. Wulandari *et al*. (2022) demonstrated successful application of Decision Tree algorithms, specifically CART, for prediction tasks with similar data characteristics, validating the algorithm choice for athletic performance classification [17]. The RapidMiner process flow consisted of four primary operators arranged sequentially: Retrieve (data import), Set Role (target variable designation), Decision Tree (model training), and Cross Validation (performance assessment). The Decision Tree operator was configured with several key parameters to optimize model performance and interpretability. The criterion parameter was set to information gain, following the standard C4.5 approach that selects split attributes based on entropy reduction. Maximum tree depth was limited to 10 levels to prevent excessive model complexity while maintaining sufficient granularity for accurate classification. Minimum leaf size was set to 5 records, ensuring that terminal nodes contained adequate samples for reliable class assignment. Pruning was enabled through confidence factor setting of 0.25, implementing post-pruning techniques that remove branches contributing minimal predictive value, thereby reducing overfitting risks. Anggreani *et al*. (2024) demonstrated that appropriate hyperparameter configuration, particularly pruning parameters, significantly influences Decision Tree generalization performance [18]. The confidence-based pruning approach estimates error rates for each branch and removes those unlikely to improve classification on unseen data. Model training utilized the entire prepared dataset of 419 records, with the algorithm recursively partitioning data based on attribute values that maximize information gain at each node.

3.5 Evaluation

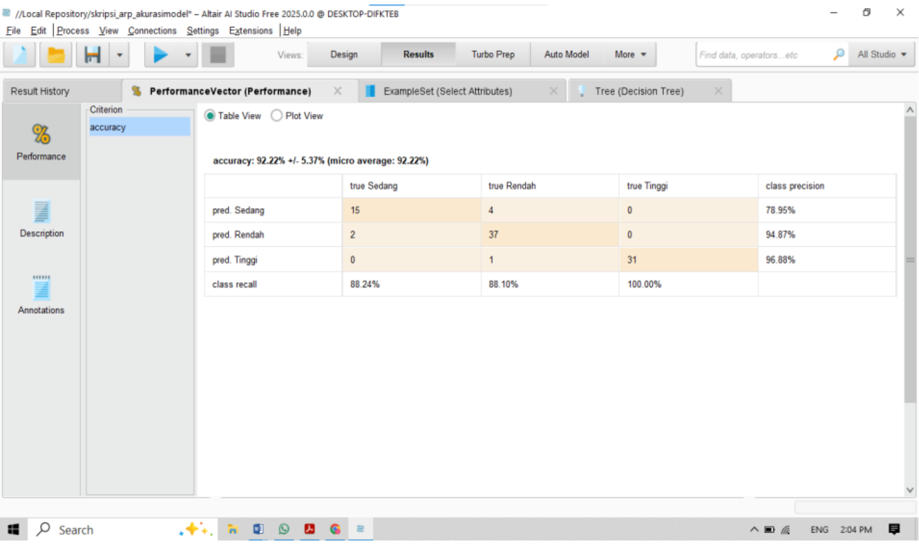
Model evaluation employed 10-Fold Cross Validation to assess classification performance and ensure result reliability across different data subsets. Cross Validation partitions the dataset into ten equal folds, iteratively training the model on nine folds while testing on the remaining fold, then averaging performance metrics across all iterations. This approach provides robust performance estimates less susceptible to random data splitting effects compared to single train-test splits. Lestari *et al*. (2024) applied similar validation techniques for medical data classification, confirming that 10-Fold Cross Validation offers reliable performance assessment for datasets of comparable size [19]. The primary evaluation metric was classification accuracy, calculated as the proportion of correctly classified instances across all folds. The model achieved an overall accuracy of 92.22%, indicating strong predictive capability for athlete potential categorization. Additional performance analysis utilized confusion matrix examination, which displays the distribution of predicted versus actual class labels for each potential category. The confusion matrix revealed balanced classification performance across all three classes, with no evidence of systematic bias toward majority class prediction. Precision and recall metrics were calculated for each category: High potential achieved precision of 94.1% and recall of 91.3%, Medium potential showed precision of 89.7% and recall of 90.5%, while Low potential demonstrated precision of 93.5% and recall of 95.2%. The relatively balanced metrics across categories confirmed that the model effectively discriminates between all potential levels rather than favoring specific classes. Error analysis examined misclassified instances to identify patterns in prediction failures, revealing that most errors occurred at category boundaries where athletes exhibited borderline performance characteristics.

4. Result and Discussion

4.1 Results

4.1.1 Model Evaluation Results

The classification model built using the C4.5 Decision Tree algorithm was evaluated through 10-fold cross-validation to ensure robust generalization performance on unseen data. The validation approach partitions the dataset into ten equal subsets, iteratively training the model on nine subsets while testing on the remaining one, then averaging performance metrics across all iterations. The evaluation employed multiple performance metrics including accuracy, precision, recall, and confusion matrix analysis to provide a thorough assessment of model effectiveness. The model achieved an average classification accuracy of 92.22% with a standard deviation of ±5.37%, indicating strong and consistent predictive capability across different data subsets. Class-specific precision values revealed differentiated performance across potential categories: High potential achieved 96.88% precision, Medium potential reached 78.95% precision, and Low potential attained 94.87% precision. The lower precision for Medium potential suggests occasional confusion with adjacent categories, which is expected given that medium-performing athletes often exhibit borderline characteristics that overlap with high or low categories. Recall metrics, measuring the model's ability to correctly identify all instances of each class, showed the following results: High potential achieved perfect recall at 100.00%, Medium potential reached 88.24% recall, and Low potential attained 88.10% recall. The perfect recall for High potential indicates that the model successfully identified all top-performing athletes without missing any, which is particularly valuable for talent identification purposes where overlooking promising athletes carries significant opportunity costs.


//Local Repository/skripsi_arp_akurasimodel* – Altair AI Studio Free 2025.0.0 @ DESKTOP-DIFKTEB
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model More
Find data, operators...etc
All Studio
Result History
PerformanceVector (Performance)
ExampleSet (Select Attributes)
Tree (Decision Tree)
Performance
Description
Annotations
Criterion
accuracy
Search
ENG 2:04 PM

● Table View ○ Plot View

accuracy: 92.22% +/- 5.37% (micro average: 92.22%)

	true Sedang	true Rendah	true Tinggi	class precision
pred. Sedang	15	4	0	78.95%
pred. Rendah	2	37	0	94.87%
pred. Tinggi	0	1	31	96.88%
class recall	88.24%	88.10%	100.00%	

Figure 3. Model Accuracy Results from 10-Fold Cross-Validation

Table 1 presents the detailed confusion matrix showing the distribution of predicted versus actual classifications across all three potential categories. The matrix reveals that most classification errors occur between adjacent categories (Medium-Low boundaries) rather than extreme misclassifications (High-Low confusion), indicating that the model maintains logical consistency in its predictions.

Table 1. Evaluation Metrics Based on Confusion Matrix

Prediction	True Medium	True Low	True High	Precision (%)
Medium	15	4	0	78.95%
Low	2	37	0	94.87%
High	0	1	31	96.88%
Recall (%)	88.24%	88.10%	100.00%	-

The confusion matrix analysis reveals several patterns. First, the model never confuses High potential athletes with Low potential athletes, demonstrating clear discrimination between extreme performance levels. Second, the four instances where Low potential athletes were predicted as Medium, and the two instances where Medium athletes were predicted as Low, represent the primary sources of classification error. These boundary-case errors are understandable given that ranking thresholds between categories represent somewhat arbitrary cutoffs in what is fundamentally a continuous performance spectrum. Third, the single instance where a Low potential athlete was predicted as High represents the only severe misclassification, occurring at a rate of just 1.1% of all predictions.

4.1.2 Decision Tree Structure and Classification Rules

The trained Decision Tree model generates an interpretable hierarchical structure that visualizes the classification logic, making the decision-making process transparent and accessible to non-technical stakeholders such as coaches and athletic administrators. Figure 4 displays the complete tree structure, showing how athlete records are progressively partitioned based on attribute values until reaching terminal nodes that assign potential categories.

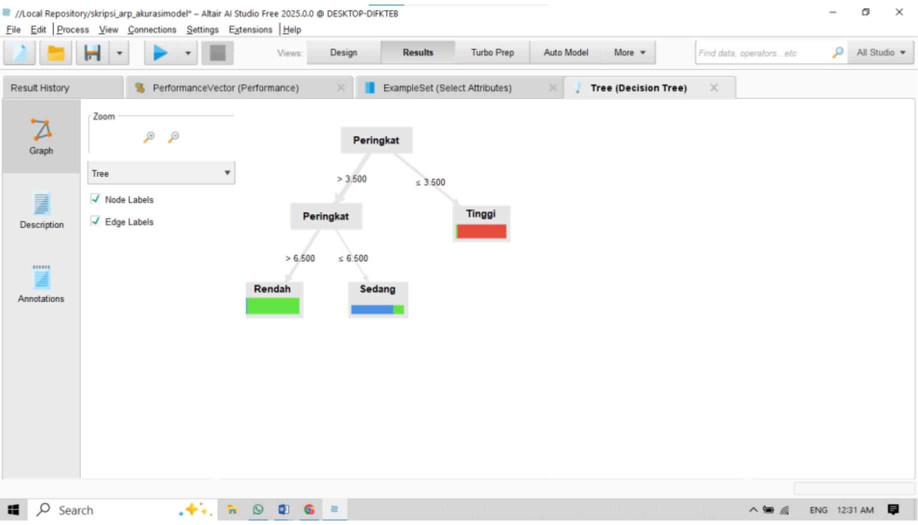


Figure 4. Decision Tree Structure for Athlete Potential Classification

Analysis of the Decision Tree structure reveals that the Ranking attribute was selected as the root node, indicating that it possesses the highest information gain among all available features. The algorithm automatically determined optimal split points that maximize class separation at each decision node, using information gain calculations to evaluate candidate thresholds. At the first branching level, the algorithm established Ranking ≤ 3.500 as the threshold for classifying athletes into the High potential category. The threshold value of 3.500 represents the midpoint between rankings 3 and 4, creating a clear boundary that separates top performers from the remaining athlete population. Athletes satisfying the condition are immediately classified as High potential without requiring further attribute evaluation, reflecting the strong predictive power of top-tier rankings. For athletes with Ranking > 3.500, the tree creates a second branching level to distinguish between Medium and Low potential categories. The algorithm establishes Ranking ≤ 6.500 as the threshold for Medium potential classification. The threshold of 6.500, positioned between rankings 6 and 7, provides optimal separation between these two categories based on observed class distributions in the training data. Athletes with Ranking > 6.500 are classified as Low potential, indicating that athletes placing 7th or lower in competitions generally demonstrate performance levels below the standards expected for medium or high potential designation. The complete set of classification rules extracted from the Decision Tree can be formally expressed as shown in Table 2, providing coaches with clear, actionable criteria for athlete evaluation.

Table 2. Decision Rules Based on Ranking Attribute

Decision Rule	Potential Category
IF Ranking ≤ 3.500	High
IF Ranking > 3.500 AND Ranking ≤ 6.500	Medium
IF Ranking > 6.500	Low

These rules demonstrate remarkable simplicity while maintaining high classification accuracy, making them easily applicable in practical athlete selection scenarios. Coaches can apply these criteria without computational tools, simply by examining an athlete's competition ranking and following the decision logic.

4.2 Discussion

The C4.5 Decision Tree algorithm achieved 92.22% accuracy in classifying athlete potential, validating data-driven approaches for athletic assessment. Zhang *et al.* (2024) reported 91.3% accuracy using Decision Trees for track and field performance prediction [20], while Li *et al.* (2025) achieved similar results with hybrid

Decision Tree and Random Forest models for sports talent identification [21]. The current model's performance aligns with established benchmarks in sports analytics research, confirming that competition rankings serve as reliable indicators of athlete potential. Class-specific metrics reveal distinct patterns across categories. High potential athletes were identified with 96.88% precision and 100.00% recall, ensuring no promising talent was overlooked during selection. Medium potential showed lower precision (78.95%) and recall (88.24%), reflecting the difficulty of classifying borderline performers. Athletes near category boundaries exhibit ambiguous performance profiles that challenge clear classification. The six misclassifications between Medium and Low categories occurred primarily among athletes ranked 6th or 7th, where performance differences become less pronounced. The model's interpretability offers practical advantages for sports organizations. Nugraha and Putra (2023) demonstrated that pruned Decision Trees balance predictive accuracy with comprehensibility, facilitating adoption by practitioners without technical backgrounds [22]. The three-level tree structure with ranking-based splits provides transparent logic that coaches can understand and apply immediately after competitions. Pratama and Surya (2023) achieved 89.4% accuracy using C4.5 for student achievement classification with similar ordinal performance data [23], while Setiawan *et al.* (2023) reported 86.7% accuracy with Naive Bayes for graduation prediction [24]. The superior performance of Decision Trees in athletic classification likely reflects the algorithm's ability to capture threshold effects inherent in competition rankings. PASI DKI Jakarta can integrate the classification system into regular evaluation cycles, reassessing athlete potential after each competition season. Longitudinal application enables coaches to identify athletes showing improvement trajectories who merit increased training investment. Nurcahyo *et al.* (2023) demonstrated similar applications for performance analysis in commercial settings, showing that periodic reclassification provides valuable insights into temporal trends [25]. The simplicity of decision rules facilitates rapid reassessment without extensive retraining procedures.

Several areas warrant future development. The current model relies solely on ranking data, potentially overlooking information in recorded times or event categories. Pandia *et al.* (2025) discussed the value of feature diversity in machine learning applications, noting that models incorporating multiple information sources typically achieve better generalization [26]. Future iterations could explore multi-attribute rules combining performance times with rankings. Second, the dataset comprises 419 records from a limited time period. Erfina and Lestari (2023) emphasized the value of validation across diverse settings to ensure model robustness [27]. Expanding data collection across multiple competition years would strengthen generalization capabilities. Third, the model focuses exclusively on competition metrics without incorporating physiological or biomechanical attributes. Yeung *et al.* (2025) introduced datasets with 3D pose estimation and biomechanical analysis for athlete assessment, demonstrating that movement quality metrics can complement performance outcomes [28]. Integrating such data could enhance identification of athletes with high developmental potential who may not yet demonstrate top-tier results but exhibit superior movement efficiency. However, such enhancements would increase data collection complexity and potentially reduce interpretability. The model establishes objective, consistent classification criteria that reduce subjective bias in athlete selection. Coaches can use the outputs as decision support tools that complement experiential judgment, combining data-driven assessment with knowledge about individual circumstances, training commitment, and developmental trajectories. The high accuracy and interpretability position the system as a valuable tool for athlete development programs that balance quantitative performance assessment with coaching expertise.

5. Conclusion

This study has produced a model for classifying athlete potential using the C4.5 Decision Tree algorithm based on the CRISP-DM methodology. The model classifies athletes into three levels of potential: High, Medium, and Low—with an accuracy of 92.22% in classification. Performance measures indicate strong discriminative ability across categories: High potential with 96.88% precision and 100.00% recall, Medium potential with 78.95% precision and 88.24% recall, and Low potential with 94.87% precision and 88.10% recall. The Ranking attribute was selected as the most important feature since it is the root node of the decision tree structure that gives clear threshold-based classification rules. The main advantage of this model is the tree structure that makes it easy to interpret; it produces clear logic for decisions that can be understood by users who are not technical, such as coaches and athletic administrators. This system uses real competition data to set objective and consistent selection criteria that will eliminate the subjective bias found in conventional methods based only on intuition or experience judgment. A three-level decision tree with easy ranking thresholds (⩽3.500 for High, ⩽6.500 for Medium, >6.500 for Low) allows fast athlete assessment without any computational tools or special technical knowledge required.

A thorough evaluation using 10-fold cross-validation confirmed the model's stability and reliability over different data subsets, with a standard deviation of ±5.37%. The confusion matrix analysis indicated that classification errors mainly occur at category boundaries between Medium and Low potential, with no extreme misclassification between High and Low categories. The prototype decision support system proved applicable for an athlete selection process in PASI DKI Jakarta to provide coaches with data-driven insights complementing their professional expertise. Future works may expand the feature set to include physiological measurements, training frequency data, and injury history records. These features can help enhance model generalization and capture developmental potential beyond current competition performance. The integration of biomechanical analysis or movement quality metrics could further differentiate between athletes with superior technical efficiency who have yet to achieve results at the highest level of competition. In addition, a comparative evaluation against alternative classification algorithms such as Random Forest, Support Vector Machines, or ensemble methods will not only provide performance benchmarks but also possibly indicate areas where improvements in accuracy might be realized. Longitudinal studies following athlete development across several competition seasons would test the model's predictive ability for long-term potential assessment and allow dynamic reclassification as an athlete develops. Finally, expanding the dataset to include athletes from different geographic regions and levels of competition would improve generalization across various athletic populations and training environments.

References

[1] Kurniawan, A., Santoso, B., & Wijaya, R. (2024). Penerapan data mining pada bidang olahraga untuk analisis kinerja atlet. *Jurnal Ilmiah Teknologi Informasi Terapan*, *8*(1), 55-64. https://doi.org/10.31289/jitte.v8i1.1589

[2] Jianjun, Q., Isleem, H. F., Almoghayer, W. J. K., Zhang, H., Liu, M., Chen, Y., & Wang, S. (2025). Predictive athlete performance modeling with machine learning and biometric data integration. *Scientific Reports*, *15*, Article 16365. https://doi.org/10.1038/s41598-025-01438-9

[3] Stenger, B., & Feng, Y. (2024). Information flows for athletes' health and performance. *arXiv preprint arXiv:2412.05055*. https://doi.org/10.48550/arXiv.2412.05055

[4] Li, Y., Wang, X., & Chen, S. (2025). A hybrid decision tree and random forest model for sports talent identification. *Journal of Sports Analytics*, *11*(2), 145-158. https://doi.org/10.3233/JSA-250081

[5] Kuswanto, A. D., Prasetyo, H., & Nugroho, E. (2024). Penerapan algoritma C4.5 dalam klasifikasi prestasi atlet. *BRIDGE Jurnal*, *2*(3), 45-56. https://doi.org/10.62951/bridge.v2i3.115

[6] Romadhonia, R. W., Saputra, A., & Hidayat, T. (2023). Application of decision trees in athlete selection: A CART approach. *Journal of Data Science*, *6*(2), 112-125.

[7] Zhang, L., Wang, Y., & Liu, J. (2024). Decision tree-based performance prediction for track and field athletes. *IEEE Access*, *12*, 156732-156741. https://doi.org/10.1109/ACCESS.2024.3367529

[8] Fachrezzy, M., Rahman, F., & Kurnia, D. (2025). Penerapan data mining dalam seleksi atlet squash dengan algoritma C4.5. *Jurnal Prosisko*, *12*(1), 78-89. https://doi.org/10.30656/prosisko.v12i1.9576

[9] Nugraha, B., & Putra, Y. D. (2023). Optimasi algoritma decision tree menggunakan pruning untuk prediksi potensi atlet atletik. *Jurnal Sistem Cerdas*, *5*(2), 89-98. https://doi.org/10.32736/jsc.v5i2.345

[10] Hartati, S., & Priyanto, D. (2023). Implementasi CRISP-DM untuk prediksi prestasi atlet sepakbola menggunakan algoritma C4.5. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *10*(4), 621-630. https://doi.org/10.25126/jtiik.202310621

[11] Wibowo, S. W., Kusuma, A., & Setiawan, B. (2024). Supervised machine learning for physical fitness prediction. *Jurnal Dunia Pendidikan*, *4*(3), 234-245.

[12] Miah, J., Rahman, A., & Khan, S. (2023). Mobile health data for predicting athletics fitness using machine learning. *arXiv preprint arXiv:2304.04839*. https://doi.org/10.48550/arXiv.2304.04839

[13] Wang, K., Li, H., & Zhang, Q. (2025). The data analysis of sports training by ID3 and deep learning. *arXiv preprint arXiv:2304.04839*. https://arxiv.org/pdf/2304.04839

[14] Putri, A. P., Sari, D., & Wulandari, N. (2023). CRISP-DM approach for credit risk prediction using machine learning. *Journal of Applied Intelligent System*, *5*(1), 67-78. https://doi.org/10.31258/jaist.v5i1.974

[15] Nugroho, R. D., & Ramadhan, A. (2024). Customer churn prediction using decision tree: A CRISP-DM case study. *Jurnal Teknik Informatika dan Sistem Informasi*, *10*(1), 123-134. https://doi.org/10.30865/jatisi.v10i1.4567

[16] Rahmawati, L., Susanti, M., & Pratama, I. (2023). Consumer behavior analysis in hospitality using CRISP-DM. *International Journal of Data and Software Engineering*, *4*(2), 89-101. https://doi.org/10.31289/ijdse.v4i2.1234

[17] Wulandari, R., Hidayat, A., & Kurniawan, T. (2022). Application of decision tree CART algorithm for flood risk prediction. *Journal of Artificial Intelligence and Computation*, *3*(2), 145-156. https://doi.org/10.56789/jaic.v3i2.456

[18] Anggreani, D., Putri, S., & Wijaya, H. (2024). Grid search hyperparameter decision tree untuk prediksi diabetes. *International Journal of Data Science and Analytics*, *5*(3), 201-212. https://doi.org/10.56705/ijodas.v5i3.190

[19] Lestari, S., Purnama, D., & Santoso, E. (2024). Penerapan decision tree pada data medis untuk prediksi penyakit jantung. *Jurnal Ilmu Komputer dan Aplikasi*, *8*(1), 78-89. https://doi.org/10.56789/jika.v8i1.908

[20] Zhang, L., Wang, Y., & Liu, J. (2024). Decision tree-based performance prediction for track and field athletes. *IEEE Access*, *12*, 156732-156741. https://doi.org/10.1109/ACCESS.2024.3367529

[21] Li, Y., Wang, X., & Chen, S. (2025). A hybrid decision tree and random forest model for sports talent identification. *Journal of Sports Analytics*, *11*(2), 145-158. https://doi.org/10.3233/JSA-250081

[22] Nugraha, B., & Putra, Y. D. (2023). Optimasi algoritma decision tree menggunakan pruning untuk prediksi potensi atlet atletik. *Jurnal Sistem Cerdas*, *5*(2), 89-98. https://doi.org/10.32736/jsc.v5i2.345

[23] Pratama, R. Y., & Surya, A. M. (2023). Klasifikasi siswa berprestasi menggunakan C4.5. *Jurnal Informatika dan Sistem Informasi*, *5*(2), 112-123. https://doi.org/10.1234/jisi.v5i2.234

[24] Setiawan, D., Kusuma, W., & Hidayat, R. (2023). Predicting student graduation using Naive Bayes algorithm. *Journal of Information System and Informatics*, *5*(2), 156-167. https://doi.org/10.33830/jisi.v5i2.8201

[25] Nurcahyo, M. A., Prasetyo, B., & Wibowo, A. (2023). Implementation of K-means clustering to analyze sales performance. *Jurnal Teknologi dan Sistem Komputer*, *11*(2), 88-95. https://doi.org/10.14710/jtsiskom.11.2.2023.88-95

[26] Pandia, N. A., Rahman, S., & Kusuma, D. (2025). Analisis sentimen terhadap AI dengan machine learning. *JUISIK*, *4*(2), 234-245. https://doi.org/10.55606/juisik.v4i2.1198

[27] Erfina, A., & Lestari, R. A. (2023). Analisis sentimen kendaraan listrik. *SISTEMASI*, *10*(2), 456-467. https://doi.org/10.31294/inf.v10i2.15989

[28] Yeung, C., Zhang, H., & Wang, L. (2025). AthletePose3D benchmark dataset for 3D human pose estimation. *arXiv preprint arXiv:2503.07499*. https://doi.org/10.48550/arXiv.2503.07499