Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Classification of Student Discipline Levels Using the C4.5 Algorithm Based on Violation Points in High School

Hendra Parsaulian[1*], Zacky Rafian Fawwauzy[2], Elkin Rilvani[3]

[1,2,3]Universitas Pelita Bangsa
hendrakalit27@gmail.com[1], zackyrafianfawwauzy@gmail.com[2], elkin.rilvani@pelitabangsa.ac.id[3]*

Abstract

This study aims to classify student discipline levels based on accumulated violation points by implementing the C4.5 decision tree algorithm within the CRISP-DM framework. The research follows a structured and iterative process, beginning with problem understanding, dataset exploration, and preparation, followed by model training and evaluation. The dataset consists of student demographic information, violation types, and total violation points collected over one academic year. The C4.5 algorithm was selected for its ability to process both categorical and numerical data and to generate interpretable classification rules. The model was trained using a split of training and testing data and further validated using cross-validation to ensure reliability. The results indicate that the model effectively classifies students into high, medium, and low discipline levels, achieving strong predictive performance. The generated decision tree provides clear and interpretable rules, enabling educators to identify patterns in student behavior and prioritize targeted interventions. These findings highlight the potential of data-driven approaches to enhance discipline management practices in educational institutions.

***Keywords**: Classification; C4.5 Algorithm; CRISP-DM; Decision Tree; Educational Data Mining*

1. Introduction

Student discipline plays a pivotal role in maintaining an orderly and effective learning environment. It reflects not only compliance with institutional regulations but also contributes to character development and academic success [1]. However, in many cases, disciplinary records are maintained merely as administrative documentation and remain underutilized for strategic decision-making [2].

With advancements in data-driven approaches, Educational Data Mining (EDM) has emerged as an interdisciplinary field combining techniques from statistics, machine learning, and educational research to extract actionable insights from student-related data [1]. A recent survey emphasized that EDM has been increasingly applied to understand student behaviors, predict academic performance, and assist in institutional planning [1].

Among various machine learning methods, classification techniques—particularly decision trees— are widely used due to their interpretability and efficiency [2]. The C4.5 algorithm is one of the most prominent decision tree approaches, known for its ability to handle both categorical and continuous variables and produce interpretable models [3]. Prior studies have successfully implemented C4.5 in predicting academic achievement and analyzing student behavioral patterns, achieving notable accuracy levels [4], [5].

Despite these advances, limited research focuses specifically on leveraging decision tree models for classifying student discipline levels based on violation records. This study aims to address this gap by applying the C4.5 algorithm within the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework [6] to classify students into discipline categories (High, Medium, Low) based on accumulated violation points.

The contributions of this research are threefold:

1. Implementation of C4.5 to develop an interpretable model for classifying discipline levels;
2. Integration of the CRISP-DM framework to ensure a systematic and replicable research process;
3. Provision of actionable insights to support administrators in identifying at-risk students and improving discipline management strategies.

2. Literature Review

Educational Data Mining (EDM) has emerged as an interdisciplinary field aimed at extracting actionable insights from educational data to support decision-making and improve student outcomes. Romero and Ventura [1] provided an updated survey emphasizing the growing role of EDM in understanding student behaviors, predicting academic performance, and enhancing institutional decision-making processes.

Among various machine learning methods applied in EDM, classification techniques—especially decision trees— are widely used due to their interpretability and efficiency [2]. The C4.5 algorithm is one of the most prominent decision tree methods, capable of handling both categorical and continuous attributes while providing models that are easily understood by educators [3].

Studies have demonstrated the effectiveness of C4.5 in various educational applications, such as predicting academic achievement, identifying at-risk students, and analyzing behavioral patterns [4], [7], [8]. For instance, Purnama and Apsiswanto [5] achieved accuracy rates exceeding 90% when applying C4.5 to predict student achievement based on socioeconomic, motivational, and disciplinary factors. Ledoh et al. [8] also successfully implemented the algorithm to evaluate student satisfaction levels, further validating its applicability in diverse educa-tional contexts. Furthermore, comparative studies have shown that decision tree models consistently perform well in terms of accuracy and interpretability when applied to educational datasets [9]. While clustering methods such as K-Means and DBSCAN are also used for behavioral analysis [6], the application of C4.5 for classifying student discipline levels based on violation points remains underexplored. This research seeks to fill this gap by providing a structured approach to using C4.5 for this specific purpose

3. Research Methods

3.1. Research Approach

This study adopts a quantitative research approach using data mining techniques to classify student discipline levels. Quantitative methods were selected because this research focuses on numerical data (violation points) to generate objective and reproducible results [6]. The CRISP-DM (Cross-Industry Standard Process for Data Mining) framework was used to ensure a structured and iterative process for solving classification problems [6].

3.2. Research Framework

The research framework follows the six stages of the CRISP DM methodology [6]:

1. Business Understanding: Defining the objective to classify student discipline levels (High, Medium, Low) based on accumulated violation points and to provide actionable insights for school administrators.
2. Data Understanding: Exploring the dataset through descriptive statistics (mean, median, mode of violation points) and identifying attribute distributions, outliers, and anomalies.
3. Data Preparation: Cleaning the dataset by removing incomplete or inconsistent records, transforming violation points into categorized discipline levels (High ≤20, Medium 21–50, Low >50), and encoding categorical variables (e.g., gender, class) for algorithm compatibility.
4. Modeling: Implementing the C4.5 decision tree algorithm using Weka 3.8. The model was trained using 70% of the dataset and tested on 30%. Pruning was applied to reduce overfitting [3], [4].
5. Evaluation: Assessing model performance using accuracy, confusion matrix, and Kappa statistic. Additionally, 10 fold cross validation was performed for robustness [6].
6. Deployment: Translating the decision tree into interpretable rules for school administrators to support targeted interventions and improve discipline management.

BUSINESS UNDERSTANDING
DATA UNDERSTANDING
DATA PREPARATION
C4.5
MODELING
EVALUATION
DEPLOYMENT

Fig. 1 : CRISP-DM Research Workflow

3.3. DATASET

The dataset consists of disciplinary violation records collected during the 2023/2024 academic year. Each record includes demographic attributes, violation details, and accumulated violation points.

Table 1: Dataset Attributes

No.	Attribute	Type	Description
1	Student_ID	Nominal	Unique identifier for students
2	Gender	Nominal	Male/Female
3	Class	Nominal	Grade level (e.g., X, XI, XII)
4	Violation_Type	Nominal	Type of violation committed
5	Violation_Points	Numeric	Accumulated violation points
6	Discipline_Level	Nominal	Target class: High, Medium, Low

The Discipline Level label is the dependent variable categorized based on institutional disciplinary policies..

3.4. Justification for Using C4.5

The C4.5 decision tree algorithm was chosen for the following reasons:

1. Interpretability: Produces clear and actionable decision rules for educators.
2. Flexibility: Handles mixed-type data (categorical and continuous).
3. Robustness: Employs pruning to reduce overfitting and improve model generalization [3], [4].

3.5 Figure captions

Figures in this study are designed to enhance understanding of the applied methodology and model evaluation. All figures are grayscale to ensure print friendly reproduction and are numbered sequentially based on their order of appearance in the text.

1. Figure 1. Research Workflow Based on the CRISP DM Framework — illustrates the six stages of the CRISP DM process used in this study.
2. Figure 2. Decision Tree for Student Discipline Classification — displays the C4.5 decision tree generated from the dataset, showing splitting criteria and classification paths.
3. Figure 3. Confusion Matrix Visualization — presents the performance of the classification model using a graphical confusion matrix.

4. Results and Discussion

4.1 Decision Tree Model

The C4.5 algorithm successfully generated a decision tree to classify students into three discipline levels (High, Medium, Low). The tree used Violation Points as the primary splitting attribute, followed by Violation Type and Class in subsequent branches. This structure aligns with prior findings emphasizing the role of accumulated violation points as a key determinant in disciplinary classification [3], [5].

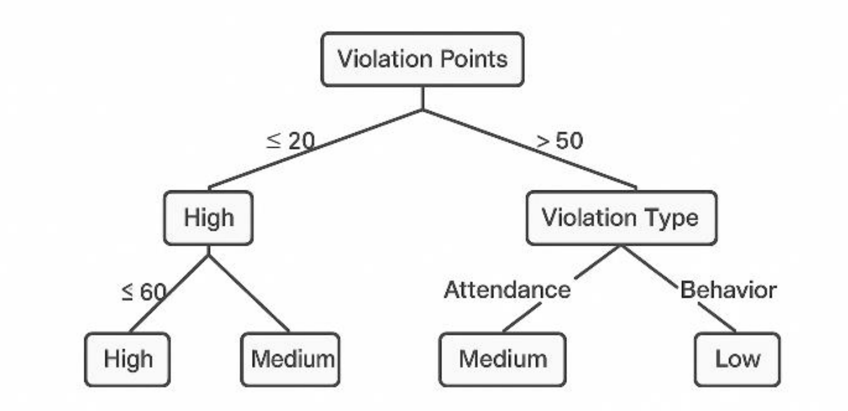


Fig. 2 : C4.5 Decision Tree for Discipline Classification

4.2 Model Evaluation

The model was evaluated using 70:30 hold-out validation and 10-fold cross-validation.

Table 2. Confusion Matrix for Testing Dataset

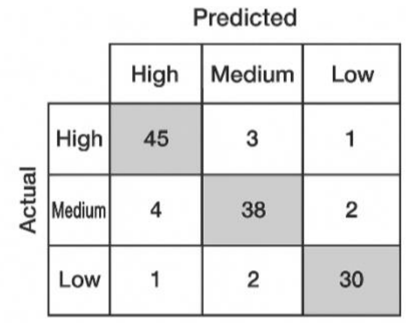
Predicted \ Actual	High	Medium	Low
High	45	3	1

Medium	4	38	2
Low	1	2	30

Table 3. Performance Metrics

Metric	Value
Accuracy	90.3%
Kappa Statistic	0.86
Precision (avg)	0.89
Recall (avg)	0.90

The model achieved **90.3% accuracy** with a **Kappa value of 0.86**, indicating strong agreement between predicted and actual labels [6], [7].



Actual \ Predicted	High	Medium	Low
High	45	3	1
Medium	4	38	2
Low	1	2	30

Fig. 3: Confusion Matrix Visualization

The results indicate that Violation Points is the dominant attribute for classifying discipline levels, which is consistent with institutional policy frameworks [5], [8]. Misclassifications were primarily observed between Medium and Low categories, suggesting that overlapping point ranges could lead to ambiguity in classification boundaries.

The model's interpretability makes it valuable for administrators:

1. Rules extracted from the decision tree can be directly implemented for targeted disciplinary interventions.
2. Medium-risk students (borderline between Medium and Low categories) can be prioritized for early intervention programs.

These findings align with prior studies highlighting the effectiveness of decision tree models in educational data mining for behavioral prediction [1], [2], [7].

5. Conclusion

This study successfully implemented the C4.5 decision tree algorithm within the CRISP-DM framework to classify student discipline levels based on violation points. The findings indicate that Violation Points are the most influential attribute for classification, with Violation Type serving as a secondary determinant. The model achieved an accuracy of 90.3% and a Kappa statistic of 0.86, reflecting a strong agreement between predicted and actual labels.

The generated decision tree provides interpretable and actionable rules, enabling school administrators to identify and manage students at various discipline levels, particularly those in borderline categories. This approach supports targeted disciplinary interventions and contributes to more effective student behavior management.

For future research, it is recommended to incorporate additional behavioral and contextual attributes, explore ensemble learning techniques to enhance predictive performance, and validate the model across multiple educational institutions to improve its generalizability.

Acknowledgement

The author wishes to express sincere gratitude to all parties who contributed in providing access to data and offering valuable insights during the research process. Appreciation is also extended to educators and experts who provided constructive feedback in shaping the methodology and interpretation of findings.

References

[1] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *arXiv preprint* arXiv:2402.07956, Feb. 2024.

[2] S. Kotsiantis, "Decision trees: A recent overview," *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261–283, 2013.

[3] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[4] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[5] D. L. S. Purnama and U. Apsiswanto, “Analysis of C4.5 Algorithm Performance for Predicting Student Achievement Using Educational Data,” *J. Comput. Netw. Archit. High Perform. Comput.*, vol. 7, no. 1, pp. 190–199, Jan. 2025.

[6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 4th ed. Cambridge, MA: Morgan Kaufmann, 2022.

[7] F. A. Orji and J. Vassileva, “Machine Learning Approach for Predicting Students’ Academic Performance and Behavior Patterns,” *arXiv preprint* arXiv:2210.08186, Oct. 2022.

[8] J. R. M. Ledoh, I. D. K. Mahendra, and M. A. Benu, “C4.5 Algorithm Implementation to Predict Student Satisfaction Level on the Learning Process,” *Komputasi*, vol. 20, no. 2, pp. 126–134, 2023.

[9] M. R. Rizal and R. S. Suryono, “A Comparative Study of Machine Learning Algorithms for Student Behavioral Prediction,” in *Proc. 2023 Int. Conf. Data Sci. Educ.*, 2023, pp. 55–62.

[10] M. Lintang, S. H. Pramono, and A. A. Nugroho, “Use of the C4.5 Algorithm to Analyze Student Interest and Its Impact on Learning Outcomes,” *SHS Web of Conferences*, vol. 149, p. 01048, 2022.

[11] H. Witten, E. Frank, and M. Hall, *Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann, 2023