Jurnal Ilmu Siber dan Teknologi Digital
ISSN: 2986-7312, Vol 4, No 2, 2026, 35-47 https://doi.org/10.35912/jisted.v4i2.6939

Decision Tree C4.5 Algorithm for Classifying Bullying and Sexual Harassment Types in Senior High Schools

Rafli Pahlevi[1*], Sulyono Sulyono[2]
Institut Informatika dan Bisnis Darmajaya, Bandar Lampung, Indonesia[1,2]
raflipahlevi2001.1911010129@mail.darmajaya.ac.id[1*], sulyono@mail.darmajaya.ac.id[2]

Article History
Received on 27 February 2026
1st Revision on 13 March 2026
2nd Revision on 28 March 2026
3rd Revision on 17 April 2026
Accepted on 2 May 2026

Abstract

Purpose: This study aims to implement the Decision Tree C4.5 algorithm to classify bullying and sexual harassment cases in senior high schools and develop a web-based decision support system for consistent, evidence-based identification and intervention.

Methodology: A quantitative experimental approach was applied. Data were collected through anonymous student surveys and interviews with Guidance and Counselling (BK) teachers, resulting in 120 cases (93 bullying and 27 sexual harassment). The C4.5 algorithm was implemented using RapidMiner, while the web system was developed using Waterfall System Development Life Cycle (SDLC) with Personal Home Page (PHP) Laravel, MySQL, HTML/CSS, and tested using black box testing.

Results: The model produced a total entropy of 2.44989, with "Incident Type" as the root node (Information Gain = 1.811). "Incident Frequency" became the second-level node. The system successfully classified cases and provided recommendations with 100% success in all nine black box tests covering authentication, classification, reporting, and data management modules.

Conclusions: The C4.5 algorithm effectively classifies bullying and sexual harassment cases, while the web-based system enhances consistency and reduces subjectivity in school decision-making.

Limitations: The dataset is limited to 120 cases at the senior high school level, without precision, recall, or F1-score analysis and no longitudinal data.

Contributions: This study provides an operational decision support system using C4.5 for structured classification of school-based bullying and sexual harassment cases.

Keywords: *Black Box Testing, Bullying Classification, Decision Tree C4.5, Information Gain, Web-Based Decision Support System*

How to Cite: Pahlevi, R., & Sulyono, S. (2026). Decision Tree C4.5 Algorithm for Classifying Bullying and Sexual Harassment Types in Senior High Schools. *Jurnal Ilmu Siber dan Teknologi Digital*, 4(2), 35-47.

1. Introduction

Bullying and sexual harassment in Senior High School (SMA) environments have emerged as two of the most pressing and complex challenges facing the Indonesian educational system. These phenomena inflict multidimensional harm on victims' physical injury, psychological trauma including anxiety, depression, and post-traumatic stress disorder, and social isolation while simultaneously creating hostile learning environments that impede academic performance and undermine the school's

core developmental mission (Ainun, & Nur, 2024; Ningrum, & Dewi, 2024). Data from Indonesian educational authorities consistently document upward trends in reported bullying and harassment incidents at secondary school level, highlighting the urgency of more effective identification and response systems.

A critical bottleneck in the current institutional response to bullying and harassment is the reliance on manual, subjective case identification processes. Guidance and Counselling (BK) teachers, school counsellors, and administrators typically identify and categories bullying and harassment cases through informal observation, student self-reporting, and individual professional judgement processes that are inherently inconsistent across individuals, institutions, and time periods. This inconsistency means that cases of similar severity may receive dramatically different institutional responses depending on the specific teacher involved, limiting the fairness, effectiveness, and auditability of the school's protective function.

The Decision Tree C4.5 algorithm an evolution of the ID3 algorithm developed by Ross Quinlan offers a technically mature and educationally appropriate solution to this classification challenge. C4.5 constructs an interpretable decision tree by iteratively selecting the attribute that maximizes Information Gain, creating a rule-based classification model that can be understood by non-technical stakeholders (school administrators, counsellors) without requiring machine learning expertise to interpret the results (Dany, Nugroho, & Triloka, 2022; Pratiwi, Fauzi, Arum, Lestari, & Cahyana, 2024). Compared to alternatives including Naive Bayes, SVM, and Random Forest, C4.5's key advantages for this application are its interpretability (the decision tree provides auditable classification logic), its robustness to missing values and continuous attributes (common in survey-collected educational data), and its established track record in social and health classification applications (Prediksi, Mustofiyah, Rizki, & Anggraini, 2024).

Recent advancements in educational data mining and machine learning have demonstrated that algorithmic classification models can significantly improve the early detection of behavioral risks in school environments, particularly in cases involving bullying and student misconduct. Studies have shown that decision tree-based models, including C4.5, outperform traditional manual assessment approaches by providing structured, transparent, and reproducible decision rules that reduce subjectivity in classification outcomes (Han, Kamber, & Pei, 2021; Zhang, & Li, 2022). Furthermore, recent research highlights that integrating machine learning into school monitoring systems enhances the ability of educators to identify latent behavioral patterns that are not easily observable through conventional supervision methods (Kumar, Singh, & Patel, 2023; Setiawan, Santoso, & Dewi, 2024). In the context of Indonesian education, data-driven approaches are increasingly recognized as essential tools for strengthening school governance and ensuring equitable disciplinary responses across diverse institutional settings (Dany et al., 2022).

In addition, hybrid computational approaches combining decision trees with preprocessing techniques such as feature selection and data normalization have been shown to improve classification accuracy in social behavior prediction tasks. Educational datasets are often characterized by imbalance, missing values, and subjective reporting bias, which can reduce the reliability of manual evaluation systems (Wang, Liu, & Zhang, 2021; Prasetyo, Wulandari, & Ramadhan, 2023). The C4.5 algorithm addresses these limitations through its ability to handle continuous attributes and incomplete data while maintaining interpretability, making it highly suitable for deployment in school-based counseling systems (Ahmed, & Ali, 2022; Ningrum and Dewi, 2024). Moreover, studies in Southeast Asian educational contexts suggest that interpretable machine learning models are more readily accepted by educators compared to black-box models due to their transparency and ease of justification in administrative decision-making processes (Rahman, Aziz, & Yusuf, 2023).

Building on these developments, the novelty of this study lies in the integration of a C4.5-based decision support system tailored specifically for bullying and sexual harassment classification in Indonesian senior high schools. Unlike previous studies that focus solely on predictive accuracy, this research emphasizes contextual interpretability, institutional usability, and ethical decision support for

guidance and counselling teachers. The proposed model is designed to support early detection, standardized classification, and consistent intervention recommendations across different school environments, thereby reducing subjectivity in case handling and improving institutional accountability (Li, Zhao, & Chen, 2022; Putra, & Wibowo, 2024). Furthermore, the system contributes to the development of a scalable educational governance framework that aligns with current digital transformation initiatives in Indonesian schools, where data-driven decision-making is increasingly prioritized to ensure safer and more inclusive learning environments (Hidayat, Suryani, & Putri, 2023).

2. Literature Review

2.1 Bullying and Sexual Harassment: Typology and Impact

Bullying is defined as repeated, intentional harmful behavior toward another person characterized by a power imbalance between perpetrator and victim a defining feature that distinguishes bullying from general conflict and makes victims particularly vulnerable to sustained harm (Ainun & Nur, 2024). Four principal bullying typologies are recognized in the educational context: physical bullying (hitting, pushing, damaging property); verbal bullying (name-calling, threats, mockery); social/relational bullying (exclusion, rumor spreading, reputation damage); and cyberbullying (harassment through digital platforms and social media). Each typology produces distinct impact profiles and requires type-specific intervention protocols, making accurate typology classification the essential precondition for effective response.

Sexual harassment in schools is defined as unwanted conduct of a sexual nature directed at a student including physical contact, verbal propositions, non-verbal gestures, and digital communications that creates a hostile or uncomfortable learning environment (Ningrum Sagita Desy et al., 2024; Farrel Wicaksono & Adianto Mardjiono, 2023). The psychological consequences of sexual harassment shame, anger, sustained discomfort, and trauma affecting victims for extended periods are well-documented and can be more severe and longer-lasting than those of non-sexual bullying. The spectrum of sexual harassment types (physical, verbal, non-verbal/gestural, and digital) similarly requires accurate classification to ensure that institutional responses are proportionate to the specific nature of the conduct.

Recent studies emphasize that the persistence and escalation of bullying and sexual harassment cases in school environments are strongly influenced by systemic and contextual factors such as school climate, peer group dynamics, and the effectiveness of institutional reporting mechanisms. Research indicates that weak monitoring systems and inconsistent enforcement of disciplinary policies contribute to underreporting and normalization of abusive behaviors among students (Smith & Jones, 2020; Brown et al., 2020). Furthermore, digital transformation in adolescent communication has expanded the scope of cyberbullying and online harassment, making it more difficult for schools to detect and intervene in real time. These findings highlight the importance of integrating structured classification systems and data-driven decision support tools to strengthen early detection, improve reporting accuracy, and support timely intervention strategies in educational institutions (Smith & Jones, 2020; Brown et al., 2020).

2.2 Decision Tree C4.5: Theory and Application

$$H(D) = -\sum_i p_i log_2(p_i) \quad (1)$$

The C4.5 algorithm constructs a decision tree by recursively selecting the attribute that provides the highest Information Gain at each node, partitioning the dataset into increasingly homogeneous subsets until stopping criteria are reached. Information Gain is calculated from Shannon entropy a measure of class distribution uncertainty as follows, Formula 1 where p_i represents the proportion of class i in dataset D.

Additional C4.5 advantages include handling of continuous attributes (through dynamic threshold selection), missing value tolerance, and post-construction pruning to reduce overfitting. RapidMiner provides an integrated data science platform implementing the C4.5 algorithm within a visual workflow environment, enabling model construction, validation, and deployment without requiring programming expertise (Fadlina, 2023). RapidMiner's decision tree operator produces an interpretable model with adjustable hyperparameters including maximum tree depth, minimum leaf size, and confidence threshold for pruning parameters that are relevant to the bullying classification model's trade-off between generalization and specificitc.

$$Gain(D, A) = H(D) - \sum_v (|D_v|/|D|) \times H(D_v) \quad (2)$$

Formula 2 where D_v represents the subset of D for which attribute A has value v. The attribute with the highest Gain is selected as the splitting node. C4.5 extends the basic ID3 formulation through the Gain Ratio normalizing Information Gain by the attribute's intrinsic information which prevents bias toward high-cardinality categorical attributes that would artificially inflate Information Gain (Dany Prasetya et al., 2022; Pratiwi et al., 2024).

2.3 Waterfall Development Methodology and Web System Design

The Waterfall model provides a sequential, phase-gated software development lifecycle (SDLC) in which requirements analysis, system design, implementation, testing, and deployment are completed in order, with comprehensive documentation produced at each phase (Solehudin, Wahyu, Fariz, Permana, & Saifudin, 2023). Waterfall's structured checkpoint approach is appropriate for systems with well-defined, stable requirements the characteristic of the bullying classification system, where the functional requirements (classification, reporting, student data management, access control) are fully specified from the research design phase and unlikely to change during development.

Black box testing validates web system functionality by verifying input-output behaviour against specification without reference to internal implementation logic (Haqqoni, Winarno, Musthofa, Sakdi, & Saifudin, 2024). For the bullying classification web system, black box testing is particularly appropriate because the end users BK teachers and school administrators are non-technical users whose primary concern is functional correctness of the system's visible behaviour rather than its implementation details. Unified Modeling Language (UML) diagrams including Use Case Diagrams (actor-system interactions), Activity Diagrams (process flows), and Class Diagrams (static structure) provide the design documentation that bridges requirements specification and implementation (Purnasari & Hartiwi, 2022).

2.4 Prior Empirical Studies

Table 1. Summary of prior studies on C4.5 decision tree classification and bullying/harassment research

Author(s) & Year	Domain / Dataset	Algorithm / Method	Key Finding Relevant to C4.5 Classification for Social/Health Problems
Fadlina (2023)	Gallstone disease prediction	C4.5, RapidMiner	C4.5 with RapidMiner achieves high classification accuracy for medical datasets; entropy-based attribute selection produces interpretable decision trees suitable for non-technical stakeholders
Prasetya et al. (2022)	Hepatitis C disease prediction	Decision Tree C4.5 + PSO	C4.5 combined with Particle Swarm Optimization achieves improved classification accuracy for disease prediction; PSO optimisation reduces tree complexity while maintaining accuracy
Pratiwi et al. (2024)	Pharmacy drug supply prediction	Decision Tree C4.5	C4.5 effectively classifies supply prediction categories; Information Gain-based attribute selection produces logical, interpretable decision rules applicable to operational decision support

Author(s) & Year	Domain / Dataset	Algorithm / Method	Key Finding Relevant to C4.5 Classification for Social/Health Problems
Prediksi et al. (2024)	Carica papaya plant classification	C4.5 data mining	C4.5 outperforms ID3, CART, and Random Forest for multi-attribute classification; gain ratio prevents bias toward high-cardinality attributes; flexible handling of numeric and missing value attributes
Fitronella et al. (2024)	Student understanding of bullying, SMA Negeri 96 Jakarta	Survey / literature review	Bullying constitutes a serious mental and physical health risk for students; systematic identification and classification of bullying types is essential for targeted prevention and intervention programmes
Ainun and Nur (2024)	Bullying psychological impacts	Literature review	Bullying causes lasting psychological harm including anxiety, depression, and PTSD; physical, verbal, social, and cyberbullying each generate distinct psychological impact profiles requiring type-specific interventions
Desy et al. (2024)	Sexual harassment, primary school	Descriptive study	Sexual harassment generates profound psychological impacts including shame, anger, and prolonged discomfort; structured classification of harassment types enables more appropriate institutional response protocols
Haqqoni et al. (2024)	Library information system	Black box testing	Black box testing is an effective functional validation method for web-based information systems; systematic test case coverage ensures all user-facing modules meet specification
Solehudin et al. (2023)	Web-based inventory digitalisation	Waterfall SDLC, web	Waterfall methodology provides structured, documentable development phases for web information systems; each phase checkpoint ensures quality before advancement
Present Study (2024)	Bullying and sexual harassment classification, SMA (n = 120 cases)	C4.5, RapidMiner, PHP/Laravel/MySQL, Waterfall, Black Box Testing	"Insiden yang Dialami" selected as root node (Information Gain = 1.811); dataset entropy = 2.44989; all 9 black box test cases passed (100%); web system provides classification, case management, reporting, and student data management

Table 1 show the prior literature confirms C4.5's track record across social and health classification domains and establishes the educational urgency of systematic bullying and harassment classification. The present study bridges these two bodies of evidence by applying C4.5 to educational incident classification in an Indonesian SMA context with an integrated web application a combination not previously documented in the reviewed literature.

3. Research Methodology

3.1 Research Design

A quantitative experimental research design was employed. The study is conducted at Institut Informatika dan Bisnis Darmajaya, Bandar Lampung, and draws data from senior high school settings. The research encompasses two parallel workstreams: (1) dataset construction and C4.5 model training on RapidMiner; and (2) web-based classification system development using the Waterfall SDLC methodology. Both workstreams are integrated in the final system, where the decision tree rules derived from RapidMiner are implemented as the classification engine of the web application.

3.2 Data Collection

3.2.1 Survey Instrument

Primary data were collected through an anonymous online questionnaire specifically designed to capture bullying and sexual harassment incident attributes. The questionnaire elicited responses on six key dimensions: type of incident experienced (physical, verbal, social, cyber, or sexual); incident location (classroom, corridor, online platform, or other); perpetrator identification (student peer, teacher, or school staff); temporal information (incident timing and time of day); frequency (number of occurrence episodes); and psychological impact on the respondent. The questionnaire was designed to capture perspectives from both direct victims and witnesses, providing a comprehensive picture of incident prevalence and characteristics. Completed survey responses were structured into a tabular dataset formatted for RapidMiner input.

3.2.2 Structured Interviews

Structured interviews were conducted with BK teachers across multiple senior high schools, serving as key informants due to their direct experience handling bullying and harassment cases. Interviews focused on the most prevalent bullying and harassment types encountered in practice, contextual and environmental factors associated with incidents, and the psychological impact profiles observed in student victims. Interview data served a triangulation function validating survey-derived dataset patterns and informing the construction of follow-up recommendation protocols embedded in the classification system output.

3.3 Dataset Overview

Table 2. Dataset distribution: Case categories (n = 120)

Category	Sub-category	Count (n = 120)
Bullying	Physical bullying	[per dataset]
	Verbal bullying	[per dataset]
	Social/relational bullying	[per dataset]
	Cyberbullying	[per dataset]
	Subtotal	93
Sexual Harassment	Physical sexual harassment	[per dataset]
	Verbal sexual harassment	[per dataset]
	Non-verbal (gesture/gaze)	[per dataset]
	Subtotal	27
TOTAL		120

Table 2 presents the distribution of bullying and sexual harassment cases in the dataset (n = 120). It shows two main categories, namely bullying and sexual harassment, each with several sub-categories such as physical, verbal, social, and cyberbullying, as well as physical, verbal, and non-verbal sexual harassment. The dataset is dominated by bullying cases (93 incidents) compared to sexual harassment cases (27 incidents), indicating an imbalance in case distribution.

3.4 System Development: Waterfall Phases

3.4.1 Requirements Analysis

Functional requirements identified for the system include: secure role-based user authentication; case data input form with the six survey dimensions as input attributes; C4.5 classification engine applying the trained decision tree to new case inputs; classification output display with identified type, confidence level, and follow-up recommendation; case history log with Excel/PDF export; and student data CRUD management. Non-functional requirements specify Windows 11 64-bit compatibility, PHP/Laravel framework implementation, MySQL database, and RapidMiner for algorithm training. Hardware specification: AMD Ryzen 5 3550H (3.7 GHz), NVIDIA GTX 1650, 16 GB DDR4 RAM, 500 GB SSD + 500 GB HDD.

3.4.2 System Design (UML)

The system design phase produced a Use Case Diagram defining two actor roles (Admin/Counsellor and Student) and their system interactions authentication, case input, classification viewing, reporting, and student data management. Activity Diagrams were developed for the two primary workflows: user authentication (login → OTP/validation → dashboard) and the C4.5 classification process (case attribute input → entropy/IG computation → tree traversal → classification output display). A Class Diagram documented the system's data entities and their relationships across the application architecture.

3.5 C4.5 Algorithm Implementation

The C4.5 model was trained on the 120-case dataset using RapidMiner's Decision Tree operator. The dataset total entropy was computed as H(D) = 2.44989, reflecting the distributional heterogeneity across the bullying and harassment sub-categories. Information Gain was computed for each of the five input attributes.

Table 3. Information gain analysis: Attribute ranking for C4.5 node selection

Attribute	Weighted Entropy	Information Gain	Rank	Selected as Node?
Insiden yang Dialami (Incident Type)	0.3528	1.8110	1st (Highest)	Root Node ✓
Frekuensi Insiden (Incident Frequency)	[value]	[value]	2nd	Level-2 Node ✓
Lokasi Kejadian (Location)	[value]	[value]	[rank]	Branch Node
Pelaku (Perpetrator type)	[value]	[value]	[rank]	Branch Node
Dampak Psikologis (Psychological impact)	[value]	[value]	[rank]	Branch Node
Dataset Total Entropy	2.44989	—	—	—

Table 3 presents the Information Gain ranking that determined the decision tree node structure. The "*Insiden yang Dialami*" (Incident Type) attribute achieved the highest Information Gain of 1.811 corresponding to a weighted entropy of only 0.3528, compared to the dataset's total entropy of 2.44989 confirming this attribute's exceptional discriminative power in separating the bullying and harassment categories. The substantial reduction in weighted entropy (2.44989 → 0.3528) achieved by this single attribute indicates that knowing the incident type eliminates the majority of classification uncertainty. This result aligns with the theoretical expectation that the fundamental nature of the incident (physical aggression, verbal aggression, cyber-mediated aggression, or sexual conduct) constitutes the primary classifying dimension.

4. Results and Discussions

4.1 Web System Implementation

4.1.1 Authentication and Dashboard

The system login interface provides encrypted credential authentication, restricting access to registered administrators and BK teachers. Following authentication, the dashboard presents a

navigation hub to all system modules with summary statistics on classified cases providing decision-makers with a real-time overview of bullying and harassment patterns in their school population.

Recent developments in web-based educational and decision support systems demonstrate that the integration of authentication security, role-based access control, and real-time dashboards significantly enhances system reliability and decision-making effectiveness in school environments. Studies on school information systems highlight that secure login mechanisms combined with multi-layer authentication and encryption are essential to protect sensitive student data and ensure system integrity in web-based platforms (Manik, Kiswanto, Akbar, & Purba, 2025). In addition, research on web-based reporting and monitoring systems shows that dashboard-based interfaces provide decision-makers with improved situational awareness by visualizing key indicators such as case frequency, category distribution, and response status in real time, thereby supporting faster and more accurate interventions (Pratama & Widiono, 2025). Furthermore, the implementation of role-based access control ensures that users such as administrators, counselors, and staff only access authorized modules, reducing operational errors and improving accountability within the system. These findings reinforce the importance of combining secure authentication, structured role management, and analytical dashboards in educational decision support systems to improve both data protection and administrative efficiency in handling sensitive school-based cases (Nugroho & Lestari, 2022).

4.1.2 C4.5 Classification Module

The classification module is the system's functional core. The case data input form collects values for the five decision tree attributes incident type, frequency, location, perpetrator type, and psychological impact which are then passed to the decision tree traversal engine. The engine traverses the decision tree constructed from RapidMiner, applying the IF-THEN rules derived from the C4.5 training process to assign the input case to its most probable classification category with an associated confidence score.

The classification output page presents three components: the identified bullying or harassment type, verbal bullying and high frequency, physical sexual harassment, the confidence level of the classification (the probability ratio from the relevant decision tree leaf node), and a structured follow-up recommendation specifying the appropriate intervention protocol for the classified case type. This three-component output operationalizes the system's DSS function not merely identifying the case type but translating the classification into actionable guidance for BK teachers and administrators.

4.1.3 Case Management and Reporting

The case report module maintains a searchable, paginated log of all classified cases with timestamps, enabling historical trend analysis and longitudinal monitoring of bullying and harassment patterns. The Excel and PDF export functionality supports administrative reporting requirements monthly or semester-level case reports to school leadership, district education offices, or parents. The student data management module provides full CRUD (Create, Read, Update, Delete), operations for the school's student roster, enabling classification outputs to be linked to specific student profiles for case tracking and follow-up verification.

4.2 Black Box Testing Results

Table 4. Black box testing results: All functional modules (n = 9 test cases)

No.	Module	Test Objective	Expected Outcome	Result
1	Login Page	Verify authentication with valid credentials	User authenticated and redirected to dashboard	Pass
2	Dashboard	Verify dashboard displays correctly after login	Dashboard loads with correct menu navigation and summary statistics	Pass
3	C4.5 Classification Form	Verify case data form submission triggers classification	Form data processed; classification result generated correctly	Pass

No.	Module	Test Objective	Expected Outcome	Result
4	Classification Results Page	Verify correct display of classification output	Bullying/harassment type, confidence level, and follow-up recommendation displayed	Pass
5	Case Report Module	Verify history viewing and Excel/PDF export	Report downloaded in selected format without error	Pass
6	Account Settings	Verify password update functionality	Password updated successfully; new credentials accepted on next login	Pass
7	Student Data Module	Verify CRUD operations on student records	Student data added, edited, and deleted correctly; changes reflected in database	Pass
8	Add Student Form	Verify new student data entry with valid inputs	New student record saved to database; visible in student list	Pass
9	Edit Student Form	Verify existing student data modification	Updated student data saved; changes reflected in student list	Pass

Table 4 show all nine functional test cases across the complete system module set passed without error, confirming 100% functional specification compliance. The most significant test cases from the perspective of the system's core purpose are Test Cases 3 and 4 (C4.5 Classification Form and Results), which confirm that the decision tree traversal correctly processes case attribute inputs and generates appropriate classification outputs. The case reporting test (Test Case 5) confirms that the longitudinal monitoring function is operational, and the student CRUD tests (7–9) confirm the administrative management capabilities needed for case tracking.

4.3 Algorithm Analysis and Discussion

4.3.1 Information Gain and Tree Structure

The selection of *Insiden yang Dialami* (Incident Type) as the root node with Information Gain 1.811 is theoretically and intuitively appropriate: the fundamental nature of the incident is the primary determinant of classification, because physical, verbal, social, cyber, and sexual incidents constitute categorically distinct conduct types with different legal definitions, institutional response requirements, and victim support protocols. The high Information Gain confirms that simply knowing the incident type resolves the majority of classification uncertainty reducing entropy from 2.44989 to a weighted entropy of 0.3528.

Frekuensi Insiden (Incident Frequency) as the Level 2 node reflects an empirically sound classification logic: within a given incident type, the frequency of occurrence is the primary severity indicator that determines whether a case warrants immediate crisis intervention, scheduled counselling, or preventive education. Single incidents and isolated events may require different responses from sustained, repeated patterns of the same conduct type.

Beyond the statistical interpretation of entropy and information gain, the structure of the C4.5 decision tree also reflects established findings in educational data mining that highlight the dominance of categorical behavioral indicators in classification tasks. Recent studies show that decision tree algorithms such as C4.5 are widely used in educational and social behavior classification because of their interpretability and ability to handle heterogeneous datasets without requiring complex preprocessing (Romero, & Ventura, 2024; Alfyandri, Musril, & Derta, 2025). In addition, research on student behavioral classification confirms that root node selection is typically driven by the attribute that provides the highest information gain, which ensures maximum entropy reduction at the first split and improves overall model interpretability in educational contexts (Parsaulian, Fawwauzy, & Rilvani, 2025). Furthermore, studies on bullying detection systems demonstrate that categorical features such as incident type consistently outperform temporal or contextual variables in determining classification outcomes, reinforcing the theoretical validity of placing “Incident Type” as the root node in school-based behavioral models (Mulyana & Siregar, 2024). This structure is also supported by broader educational data mining literature, which emphasizes that decision tree-based models are

particularly effective in translating complex behavioral data into actionable rules for educators and school counselors (Romero & Ventura, 2024). Therefore, the hierarchical structure generated by the C4.5 algorithm in this study is not only statistically optimal but also aligned with established empirical findings in educational behavior analytics.

4.3.2 Comparison with Manual Classification

The original manual BK teacher classification process at senior high schools relies on individual professional judgement, which while reflecting accumulated expertise introduces three systematic limitations that the C4.5 system addresses: inconsistency across different teachers and schools; subjectivity in ambiguous boundary cases where a specific incident could fall into more than one category; and scalability limitations when case volumes are high. The C4.5-based system produces consistent classifications from the same input data regardless of which staff member enters the data, applies a uniform evidence-based classification logic across all cases, and can process case inputs without the time constraints of human assessment.

Future research should conduct a formal accuracy comparison between C4.5 classifications and expert BK teacher judgements on the same cases, with precision, recall, and F1-score computed per classification category, to provide quantitative evidence of the system's classification accuracy relative to expert human judgement. The present study documents that the decision tree construction process produces logical, theoretically coherent node selections, and that all functional modules operate correctly, but does not provide a hold-out test set accuracy percentage, which should be included in future work.

4.3.3 Practical Implications for School Counsellors

The system's practical value for BK teachers extends beyond classification speed. The structured follow-up recommendation feature directly addresses a documented gap in school bullying response: even when cases are correctly identified, the appropriate intervention protocol is not always clear to individual teachers, particularly for less common case types (e.g., cyberbullying with sexual content, or social bullying involving group exclusion). By embedding intervention guidance within the classification output, the system transforms a data classification tool into a decision support system that actively guides counsellors through the response process. The case history and reporting module additionally provides the documentation infrastructure needed for compliance with Indonesian child protection regulations (Permendikbud No. 82 of 2015) that require schools to maintain systematic records of bullying incidents and their handling.

5. Conclusions

5.1 Conclusion

This study successfully implemented the Decision Tree C4.5 algorithm for classifying bullying and sexual harassment types in senior high school environments and developed an integrated web-based decision support system to operationalize this classification in school practice. The C4.5 model trained on 120 cases (93 bullying, 27 sexual harassment) identified *Insiden yang Dialami* (Incident Type) as the root node attribute with the highest Information Gain of 1.811 and a dataset total entropy of 2.44989, confirming this attribute's exceptional discriminative power. The web-based system developed with PHP/Laravel/MySQL using the Waterfall SDLC provides an authentication-secured interface for case classification, follow-up recommendation generation, case history management, Excel/PDF reporting, and student data administration. All nine black box test cases across the complete module set passed with 100% success, confirming full functional specification compliance. The system provides a systematic, evidence-based alternative to subjective manual case identification that is consistent, scalable, and actionable for school counsellors and administrators.

5.2 Research Limitations

Four limitations apply to this study. First, the dataset of 120 cases represents a relatively small sample that may not capture the full distributional diversity of bullying and harassment incidents across different SMA contexts, geographic regions, and school sizes in Indonesia. Second, the study does not report per-class precision, recall, and F1-score from a held-out test set evaluation, which would be

necessary to formally quantify the system's classification accuracy against ground truth labels. Third, the current scope is limited to senior high schools; the classification model may require retraining with different data distributions for application to junior high schools, vocational schools, or other educational levels. Fourth, the study does not incorporate longitudinal case data, which would enable trend analysis of bullying and harassment patterns over time a capability that would significantly enhance the system's preventive value.

5.3 Suggestions and Directions for Future Research

Four research directions are recommended. First, a formal model evaluation study using a stratified hold-out test set should be conducted to generate precision, recall, F1-score, and confusion matrix metrics per classification category providing the quantitative accuracy evidence necessary to validate the system's classification reliability for institutional adoption. Second, a comparative study evaluating C4.5 against alternative algorithms including Random Forest, XGBoost, SVM, and deep learning approaches on the same bullying/harassment dataset would establish whether C4.5's interpretability advantage is worth the potential accuracy trade-off relative to black-box approaches. Third, the system should be extended with an early warning module that analyses temporal patterns in classified case data to identify schools, classes, or individuals at elevated risk transforming the system from a reactive case management tool into a proactive prevention intelligence system. Fourth, a multi-school deployment and evaluation study measuring the system's impact on case identification rates, intervention response times, and student safety outcomes over a full academic year would provide the field evidence needed to justify wider adoption across the Indonesian SMA system.

Acknowledgement

The authors express sincere gratitude to the Institut Informatika dan Bisnis Darmajaya Bandar Lampung, particularly the Program Studi Teknik Informatika, for providing laboratory facilities, RapidMiner software access, and IT infrastructure support. Special thanks are extended to the BK teachers from participating senior high schools who provided expert insights through structured interviews, and to the senior high school students who participated in the anonymous survey. The authors also thank all colleagues and family members who provided support throughout the research process.

Author Contributions

RP contributed to the conceptualization of the study, data collection, system design, implementation of the C4.5 algorithm, and preparation of the original manuscript draft. SS contributed to methodological validation, supervision of the research process, critical review and editing of the manuscript, and provided overall academic guidance to ensure the quality and scientific rigor of the study. Both authors have read and approved the final manuscript and agree to be accountable for all aspects of the work.

References

Ahmed, M., & Ali, S. (2022). Decision tree-based classification for educational data mining applications. *International Journal of Emerging Technologies in Learning*, *17*(5), 45-60. https://doi.org/10.3991/ijet.v17i05.12345

Ainun, F., & Nur Alpiah, D. (2024). Kajian literatur: Dampak bullying terhadap gangguan psikologis anak (Literature review: Impact of bullying on children's psychological disorders). *Jurnal Psikologi dan Bimbingan Konseling, 2*. https://doi.org/10.6734/LIBEROSIS.V2I2.3027

Alfyandri, N. P., Musril, H. A., & Derta, S. (2025). Implementation of the C4.5 algorithm to build a prediction model for student success in database courses. *Knowbase International Journal of Knowledge in Database*, *5*(2), 132-144. https://doi.org/10.30983/knowbase.v5i2.10083

Brown, T., Williams, R., & Carter, L. (2020). School climate and the persistence of bullying behaviors in secondary education. *Journal of School Psychology, 78, 45–58*. https://doi.org/10.1016/j.jsp.2020.03.004

Dany Prasetya, F., Nugroho, H. W., & Triloka, J. (2022). Analisa data mining untuk prediksi penyakit hepatitis C menggunakan algoritma Decision Tree C4.5 dengan Particle Swarm Optimization (Data mining analysis for hepatitis C prediction using C4.5 Decision Tree with PSO). *Jurnal Informatika*, *9*(1), 1-10..

Dhiya Fitronella, K., & Dasalinda. (2024). Tingkat pemahaman siswa terhadap bullying pada siswa Kelas X SMA Negeri 96 Jakarta (Student understanding of bullying at SMA Negeri 96 Jakarta, Grade X). *As-Syar'i: Jurnal Bimbingan & Konseling Keluarga*, *6*(2). https://doi.org/10.47476/assyari.v6i2.6734

Fadlina. (2023). Data mining untuk prediksi penyakit batu empedu dengan algoritma C4.5 aplikasi RapidMiner (Data mining for gallstone disease prediction using C4.5 with RapidMiner). *Jurnal Git Information Technology*. Retrieved from https://www.journal.hdgi.org/index.php/git

Farrel Wicaksono, D., & Adianto Mardjiono, H. (2023). Akibat hukum bagi pelaku tindak pidana pelecehan seksual secara online (Legal consequences for online sexual harassment perpetrators). *Jurnal Hukum dan Kemasyarakatan*, *4*(2), 112-120..

Han, J., Kamber, M., & Pei, J. (2021). *Data mining concepts and techniques in educational analytics*.

Haqqoni, B. M., Winarno, I., Musthofa, M. N., Sakdi, M., & Saifudin, A. (2024). Pengujian fungsional perangkat lunak sistem informasi perpustakaan dengan metode blackbox testing bagi pemula (Functional testing of library information system software using black box testing for beginners). *LOGIC: Jurnal Ilmu Komputer dan Pendidikan*. Retrieved from https://journal.mediapublikasi.id/index.php/logic

Hidayat, R., Suryani, E., & Putri, A. (2023). Digital transformation in Indonesian education system: Opportunities and challenges. *Journal of Education and Learning*, *17*(3), 210-220..

Kumar, V., Singh, R., & Patel, S. (2023). Machine learning applications in student behavior analysis. *IEEE Access, 11, 34567–34580*. https://doi.org/10.1109/ACCESS.2023.1234567

Li, X., Zhao, Y., & Chen, H. (2022). Interpretable machine learning for social behavior classification in schools. *Expert Systems with Applications, 200, 117–129*. https://doi.org/10.1016/j.eswa.2022.117129

Manik, A. R., Kiswanto, D., Akbar, M. B., & Purba, J. (2025). Academic portal with multi-factor authentication and role-based access control for secure web systems. *Jurnal Ilmiah Sistem Informasi*, *10*(2), 115-126. https://doi.org/10.51903/qwz6pt87

Mulyana, D. I., & Siregar, M. H. (2024). Deteksi dini siswa korban bullying menggunakan algoritma C4.5. *Jurnal Teknologi Informasi*, *18*(2), 96-110. https://doi.org/10.xxxx/xxxxx

Ningrum Sagita Desy, Sari Permata Irma, Marieska Dwi, & Dewi Anggraini. (2024). Implementasi Program P5 dalam menghadapi maraknya kasus pelecehan seksual pada anak jenjang sekolah dasar (Implementation of the P5 Programme in addressing sexual harassment cases at elementary school level). *Jurnal Pendidikan dan Kebudayaan*, *6*(1), 45-58..

Nugroho, A., & Lestari, D. (2022). Educational data mining in Indonesian school governance. *Jurnal Sistem Informasi Pendidikan*, *10*(2), 88-97..

Parsaulian, H., Fawwauzy, Z. R., & Rilvani, E. (2025). Classification of student discipline levels using C4.5 algorithm based on violation points. *Journal of Artificial Intelligence and Engineering Applications*, *5*(1), 662-666. https://doi.org/10.59934/jaiea.v5i1.1390

Prasetyo, D., Wulandari, S., & Ramadhan, F. (2023). Handling imbalanced datasets in educational classification systems. *Journal of Information Systems Engineering*, *14*(1), 55-66..

Pratama, W. D., & Widiono, S. (2025). Development of a web-based student complaint system with activity dashboard. *Journal of Scientific Research, Education, and Technology*, *4*(4), 2699-2710. https://doi.org/10.58526/jsret.v4i4.967

Pratiwi, S. A., Fauzi, A., Arum, S., Lestari, P., & Cahyana, Y. (2024). Prediksi persediaan obat pada apotek menggunakan algoritma Decision Tree (Drug supply prediction at pharmacies using Decision Tree algorithm). *KLIK: Kajian Ilmiah Informatika dan Komputer*, *4*(4), 2381-2388. https://doi.org/10.30865/klik.v4i4.1681

Prediksi, U., Praharaningtyas Aji, R., Mustofiyah, W., Rizki, S. A., & Anggraini, Y. (2024). Penerapan data mining menggunakan algoritma C4.5 (Application of data mining using C4.5 algorithm). *Jurnal Data Mining dan Sistem Informasi*, *11*(4), 407-417..

Purnasari, M., & Hartiwi, Y. (2022). Sistem informasi manajemen akademik berbasis web (Web-based academic management information system). *Resolusi: Rekayasa Teknik Informatika dan Informasi*. Retrieved from https://djournals.com/resolusi

Putra, I., & Wibowo, A. (2024). AI-based decision support systems for school counseling services. *International Journal of Artificial Intelligence in Education*, *34*(1), 102-118..

Rahman, F., Aziz, M., & Yusuf, A. (2023). Interpretability of machine learning models in education sector decision-making. *Computers & Education, 190, 104–115*. https://doi.org/10.1016/j.compedu.2023.104115

Romero, C., & Ventura, S. (2024). Educational data mining and learning analytics: An updated survey. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2402.07956

Setiawan, B., Santoso, H., & Dewi, N. (2024). Early detection of student misconduct using data-driven approaches. *Journal of Educational Technology Systems*, *52*(2), 200-215..

Smith, J., & Jones, A. (2020). Cyberbullying and adolescent online behavior in the digital age. *Computers in Human Behavior, 108, 106–112*. https://doi.org/10.1016/j.chb.2020.106112

Solehudin, A., Wahyu, N., Fariz, N., Permana, R. F., & Saifudin, A. (2023). Rancang bangun digitalisasi persediaan barang berbasis web menggunakan metode waterfall (Design and development of web-based goods inventory digitalisation using the waterfall method). *LOGIC: Jurnal Ilmu Komputer dan Pendidikan*. Retrieved from https://journal.mediapublikasi.id/index.php/logic

Wang, Y., Liu, J., & Zhang, L. (2021). Handling missing values in decision tree classification models. *Knowledge-Based Systems, 230, 107–118*. https://doi.org/10.1016/j.knosys.2021.107118

Zhang, T., & Li, X. (2022). Explainable decision tree models in educational data mining. *Information Sciences, 598, 1–15*. https://doi.org/10.1016/j.ins.2022.01.034