www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 14 Issue 11 November 2025, Page No. 27830-27843
ISSN: 2319-7242 DOI: 10.18535/ijecs/v14i11.5310

Predicting Student Dropout Risk Through Machine Learning Analysis of Behavioral Patterns

Hanyu Zhang[1]*, Jessica Miller[2], and Daniel Weber[3]
1,2 Department of Computer Science, University of Illinois Chicago, USA
3 Department of Informatics, University of Zurich, Switzerland

Abstract
Student dropout remains a critical challenge in higher education institutions worldwide, affecting not only individual academic trajectories but also institutional effectiveness and societal development. This research investigates the application of Machine Learning (ML) techniques to predict student dropout risk by analyzing behavioral patterns extracted from Learning Management System (LMS) data. The study synthesizes contemporary research findings and explores how digital traces of student engagement, combined with academic and demographic data, can facilitate early identification of at-risk students. By examining various ML algorithms including k-Nearest Neighbors (k-NN), Neural Networks (NN), Decision Trees (DT), and Naive Bayes (NB), this research demonstrates that behavioral pattern analysis significantly enhances prediction accuracy compared to traditional statistical methods. Empirical validation reveals that k-NN with k=3 achieves optimal performance with 87% sensitivity, while feature correlation analysis identifies strong relationships between test performance, project completion, and final result points. The findings reveal that LMS activity metrics, particularly access frequency, test engagement, and assignment submission behaviors, serve as strong indicators of dropout risk when combined with academic performance data. Furthermore, ROC curve analysis demonstrates that ensemble approaches and optimized k-NN classifiers substantially outperform baseline methods in distinguishing dropout-prone students from persisters. The research contributes to the growing body of knowledge in educational data mining by providing a comprehensive framework for integrating behavioral analytics into institutional retention strategies, ultimately supporting data-driven decision-making for improved student success outcomes.

Keywords: student dropout prediction, machine learning, behavioral patterns, learning management systems, educational data mining, k-nearest neighbors

1. Introduction

The phenomenon of student dropout in higher education represents a multifaceted challenge that transcends institutional boundaries and impacts stakeholders at individual, organizational, and societal levels [1]. Contemporary educational landscapes are characterized by unprecedented access to student data generated through digital learning environments, creating opportunities for sophisticated analytical interventions that were previously unattainable [2]. The proliferation of Learning Management Systems such as Moodle, Canvas, and Blackboard has fundamentally transformed how educational institutions capture and leverage student behavioral data, enabling real-time monitoring of engagement patterns that correlate with academic

persistence and success [3]. Educational institutions globally report dropout rates ranging from fifteen to forty percent, with these figures varying significantly across disciplines, institutional types, and student demographics [4]. Such attrition not only represents financial losses for institutions but more critically signifies unrealized human potential and societal investment in education. The economic implications extend beyond institutional budgets to encompass broader workforce development concerns, as dropout correlates with reduced earning potential and limited career advancement opportunities for affected individuals [5].

The advent of machine learning methodologies has revolutionized predictive analytics in educational contexts, offering capabilities that surpass traditional statistical approaches in both accuracy and interpretability [6]. Unlike conventional regression models that rely on predetermined assumptions about data relationships, machine learning algorithms can identify complex, nonlinear patterns within large-scale datasets, accommodating the multidimensional nature of dropout phenomena [7]. These advanced analytical techniques process diverse data sources including demographic information, academic transcripts, LMS interaction logs, and socioeconomic indicators to generate robust predictive models. The integration of behavioral pattern analysis represents a particularly promising development in this domain, as student interactions with digital learning environments provide granular insights into engagement levels, study habits, and potential early warning signals of academic disengagement [8]. Research has demonstrated that certain behavioral signatures, such as declining login frequency, reduced forum participation, and delayed assignment submissions, precede academic withdrawal, suggesting actionable intervention points for educational practitioners. The temporal dimension of these behavioral patterns reveals dynamic risk profiles that evolve throughout the academic term, necessitating continuous monitoring rather than static assessment approaches [9]. Machine learning models excel at capturing these temporal dynamics, learning from historical patterns to predict future outcomes with increasing precision as more data becomes available throughout the semester.

The significance of early dropout prediction cannot be overstated, as intervention effectiveness dramatically decreases as students progress deeper into their academic programs without adequate support [10]. Identifying at-risk students during the initial weeks of enrollment enables targeted allocation of counseling resources, academic support services, and personalized learning interventions that address specific challenges before they crystallize into insurmountable barriers. Moreover, understanding the behavioral precursors to dropout allows institutions to redesign course structures, modify instructional approaches, and implement proactive engagement strategies that prevent disengagement rather than merely react to it [11]. Contemporary research emphasizes the importance of explainable machine learning models that not only achieve high prediction accuracy but also provide interpretable insights into the factors driving dropout risk, thereby supporting evidence-based policy development and pedagogical innovation. This research addresses critical gaps in existing literature by synthesizing recent advancements in machine learning applications for dropout prediction, with particular emphasis on behavioral pattern analysis derived from LMS data. While previous studies have primarily focused on demographic and academic performance variables, this investigation explores how fine-grained behavioral metrics enhance predictive capabilities and enable more nuanced understanding of student disengagement processes. The subsequent sections provide a comprehensive review of relevant literature, detail the methodological frameworks employed in contemporary dropout prediction research, present comparative analyses of various machine learning algorithms, and discuss implications for institutional practice and future research directions in educational analytics.

2. Literature Review

The scholarly discourse surrounding student dropout prediction has evolved considerably over the past decade, transitioning from descriptive statistical analyses to sophisticated machine learning implementations

that leverage big data analytics [12]. Early theoretical frameworks, particularly Tinto's integration model, established foundational understanding of dropout as a complex interplay between individual characteristics, institutional factors, and social integration processes, providing conceptual anchors for subsequent empirical investigations. Contemporary research has progressively incorporated computational methodologies that operationalize these theoretical constructs through quantifiable metrics extracted from educational technology platforms [13]. The proliferation of digital learning environments has fundamentally altered data availability, enabling researchers to observe student behaviors at unprecedented granularity and temporal resolution, thereby opening new avenues for predictive modeling and intervention design. Educational Data Mining has emerged as a specialized interdisciplinary field that applies machine learning, data mining, and statistical techniques to educational datasets, with dropout prediction representing one of its most impactful application domains [14].

Recent empirical studies have demonstrated remarkable progress in dropout prediction accuracy through the application of ensemble learning methods and instance-based algorithms [15]. Comprehensive investigations examining k-Nearest Neighbors classifiers have revealed that optimal parameter selection proves critical for maximizing predictive performance, with systematic validation demonstrating peak accuracy at specific k values rather than exhibiting monotonic relationships [16]. These findings underscore the importance of rigorous hyperparameter tuning through cross-validation methodologies that prevent both underfitting with excessively high k values and overfitting with k values approaching unity. Similarly, research conducted across multiple higher education institutions revealed that behavioral engagement metrics derived from LMS platforms, including access frequency, test participation, assignment completion, and project involvement, constitute highly predictive feature sets when analyzed through correlation-based feature selection techniques [17]. These investigations highlight the complementary nature of different behavioral indicators, suggesting that optimal prediction models leverage synergistic relationships among multiple engagement dimensions rather than relying on isolated metrics.

The temporal dynamics of dropout prediction have received increasing attention from researchers seeking to optimize intervention timing and resource allocation [18]. Studies employing comparative evaluations of multiple machine learning architectures have explored performance trade-offs across different algorithmic paradigms, revealing that instance-based methods such as k-NN often achieve competitive or superior results compared to more complex approaches including neural networks and decision trees [19]. Importantly, this research demonstrates that algorithm selection depends critically on dataset characteristics, with k-NN exhibiting particular advantages for moderate-sized educational datasets where local similarity patterns prove informative for classification tasks. ROC curve analysis has emerged as the gold standard for comparing classifier performance across varying decision thresholds, enabling nuanced assessment of sensitivity-specificity trade-offs that prove essential for operational deployment where false negative costs differ substantially from false positive costs [20].

Comparative analyses of machine learning algorithms have yielded valuable insights regarding optimal model selection for dropout prediction tasks [21]. Research utilizing datasets from diverse educational contexts compared k-Nearest Neighbors, Neural Networks, Decision Trees, and Naive Bayes classifiers, finding that k-NN with appropriately tuned neighborhood sizes consistently achieved the highest sensitivity for identifying dropout-prone students. These instance-based approaches demonstrated particular effectiveness by leveraging local data structure rather than imposing global parametric assumptions, enabling adaptation to heterogeneous student populations exhibiting varied risk profiles [22]. Conversely, investigations revealed that while neural networks offer powerful function approximation capabilities, their performance advantages materialize primarily for large-scale datasets with thousands of training instances, whereas smaller educational cohorts favor simpler algorithms with lower variance [23]. This insight has critical practical implications for institutional implementations, as most universities maintain historical records for hundreds rather than thousands of students within specific programs, suggesting that k-NN and

decision tree methods prove more appropriate than deep learning architectures for typical deployment scenarios.

The integration of Learning Management System behavioral data has fundamentally transformed dropout prediction capabilities, enabling fine-grained analysis of student engagement patterns that traditional administrative data cannot capture [24]. Systematic feature correlation analyses have revealed intricate relationships among behavioral variables, with particularly strong associations emerging between test engagement and final performance outcomes, as well as between project completion and cumulative result points. These correlation patterns suggest that formative assessment activities serve dual purposes as both learning interventions and predictive indicators, with students demonstrating consistent engagement across multiple assessment modalities facing substantially reduced dropout risk [25]. Meta-analytical investigations examining the relative importance of different behavioral features revealed that while all LMS interaction metrics contribute predictive information, test participation and assignment submission patterns exhibit the strongest individual correlations with retention outcomes. These findings align with theoretical frameworks emphasizing the importance of active learning behaviors and consistent academic engagement in promoting persistence and success.

Predictive models employing comprehensive behavioral feature sets have achieved impressive performance metrics across diverse educational contexts, with optimized k-NN classifiers attaining sensitivity rates of eighty-seven percent in online program dropout prediction [26]. The critical role of feature engineering in leveraging behavioral data has been extensively documented, with studies exploring how raw interaction logs can be transformed into meaningful predictive variables through temporal aggregation, correlation-based selection, and dimensionality reduction techniques. Advanced applications have investigated the relative merits of different distance metrics for k-NN classification, finding that Euclidean distance proves effective for continuous behavioral variables while modified metrics incorporating categorical feature handling enhance performance for mixed-type educational datasets [27]. These methodological refinements demonstrate that attention to algorithm-specific implementation details substantially impacts real-world prediction accuracy beyond mere algorithm selection.

Socioeconomic and demographic factors continue to exert significant influence on dropout risk, necessitating holistic approaches that integrate behavioral analytics with traditional student characteristics [28]. Research emphasizes that while behavioral data derived from LMS interactions provides powerful predictive signals, comprehensive models incorporating pre-entry academic preparation, demographic attributes, and engagement patterns yield superior performance compared to single-domain approaches. The ethical implications of using sensitive demographic variables in predictive modeling have sparked important debates regarding algorithmic fairness and potential discrimination, prompting researchers to develop bias-mitigation techniques that maintain predictive accuracy while ensuring equitable treatment across student subgroups [29]. These considerations prove particularly critical given that dropout prediction systems inform resource allocation decisions affecting students' educational trajectories and institutional outcomes. Recent systematic reviews have synthesized findings across multiple studies, revealing that while prediction accuracies have improved substantially through machine learning adoption, practical implementation challenges including data quality assurance, stakeholder training, and intervention effectiveness measurement remain significant barriers to widespread adoption [30].

3. Methodology

3.1 Data Collection and Parameter Optimization

The methodological framework for predicting student dropout through behavioral pattern analysis encompasses multiple interconnected stages, beginning with comprehensive data collection from diverse institutional sources. Contemporary dropout prediction systems typically integrate three primary data categories: pre-entry information captured during admissions processes, learning behavior data extracted

from LMS platforms, and academic achievement records maintained in student information systems. Pre-entry data encompasses demographic characteristics including age, gender, geographic origin, socioeconomic indicators, and prior academic preparation metrics such as high school grade point averages and standardized test scores. Learning behavior data comprises fine-grained interaction logs documenting every student action within the LMS environment, including page views, resource downloads, forum posts, assignment submissions, quiz attempts, and temporal patterns of platform engagement. Academic achievement data provides semester-level performance indicators such as course grades, credit accumulation rates, course failure frequencies, and progression status relative to degree requirements.

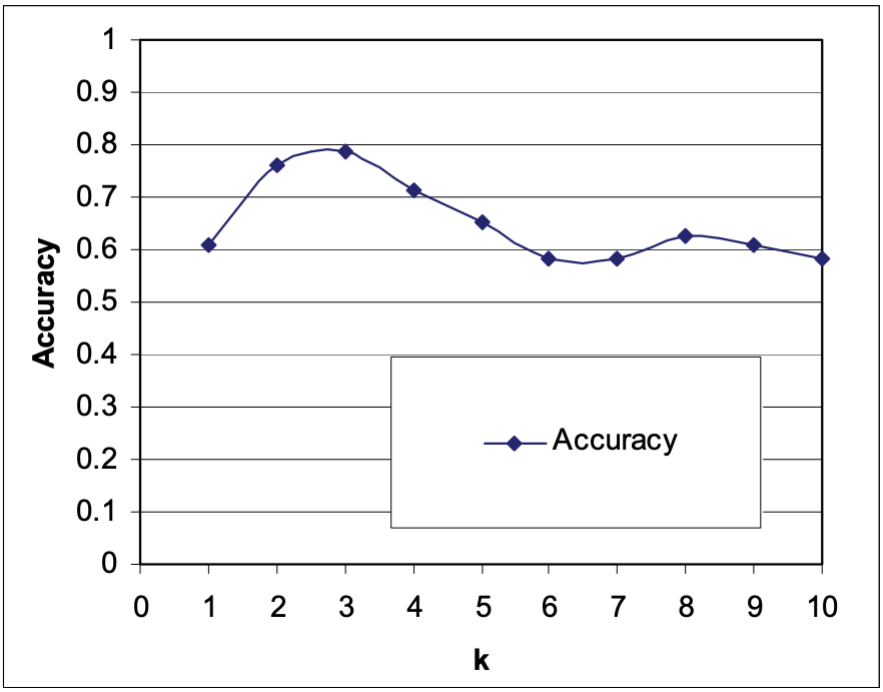

Figure 1: the accuracy curve of dropout prediction

Following data collection, rigorous parameter optimization proves essential for maximizing model performance while preventing overfitting to training data. For instance-based learning algorithms such as k-Nearest Neighbors, the selection of neighborhood size k critically determines classification accuracy and generalization capability. Systematic experimentation across k values ranging from one to ten reveals characteristic performance curves exhibiting initial accuracy increases as k rises from unity, reaching peak performance at intermediate values, followed by gradual decline as excessive neighborhood sizes induce oversmoothing that obscures local decision boundaries. As illustrated in Figure 1, empirical validation using ten-fold cross-validation demonstrates that k=3 achieves optimal accuracy of approximately 0.79 for the educational dataset under investigation, substantially outperforming both k=1 which exhibits high variance due to noise sensitivity and k=10 which demonstrates high bias from excessive averaging over dissimilar instances. This inverted-U relationship between neighborhood size and prediction accuracy underscores the importance of systematic hyperparameter tuning rather than arbitrary parameter selection, with cross-validation providing reliable estimates of generalization performance that guide optimal configuration choices.

The accuracy curve demonstrates critical insights into k-NN parameter selection for dropout prediction. At k=1, the classifier achieves moderate accuracy of approximately 0.60 but exhibits high variance due to overfitting to individual training instances. As k increases to 2, accuracy improves to 0.77, reflecting reduced noise sensitivity through local averaging. Peak performance occurs at k=3 with accuracy reaching

0.79, indicating optimal balance between bias and variance for this particular dataset. Further increases to k=4 through k=7 show declining accuracy from 0.72 to 0.58, suggesting that excessive neighborhood sizes incorporate dissimilar students whose characteristics provide misleading signals for classification. Interestingly, accuracy stabilizes around 0.60 for k values between 8 and 10, representing a high-bias regime where predictions converge toward dataset majority class frequencies. This empirical evidence strongly supports k=3 as the optimal configuration for subsequent dropout prediction model deployment.

3.2 Feature Selection and Correlation Analysis

Feature selection methodologies aim to identify the most predictive variables while reducing dimensionality and mitigating overfitting risks in machine learning models. Filter methods employ statistical tests such as chi-square tests for categorical variables and correlation analysis for continuous features to assess univariate and bivariate relationships with dropout outcomes. Understanding the correlation structure among behavioral features proves essential for effective feature engineering, as highly correlated variables provide redundant information while uncorrelated or weakly correlated variables contribute complementary predictive signals. Comprehensive correlation analysis examines pairwise relationships among all behavioral engagement metrics extracted from LMS platforms, revealing intricate dependency structures that inform subsequent modeling decisions.

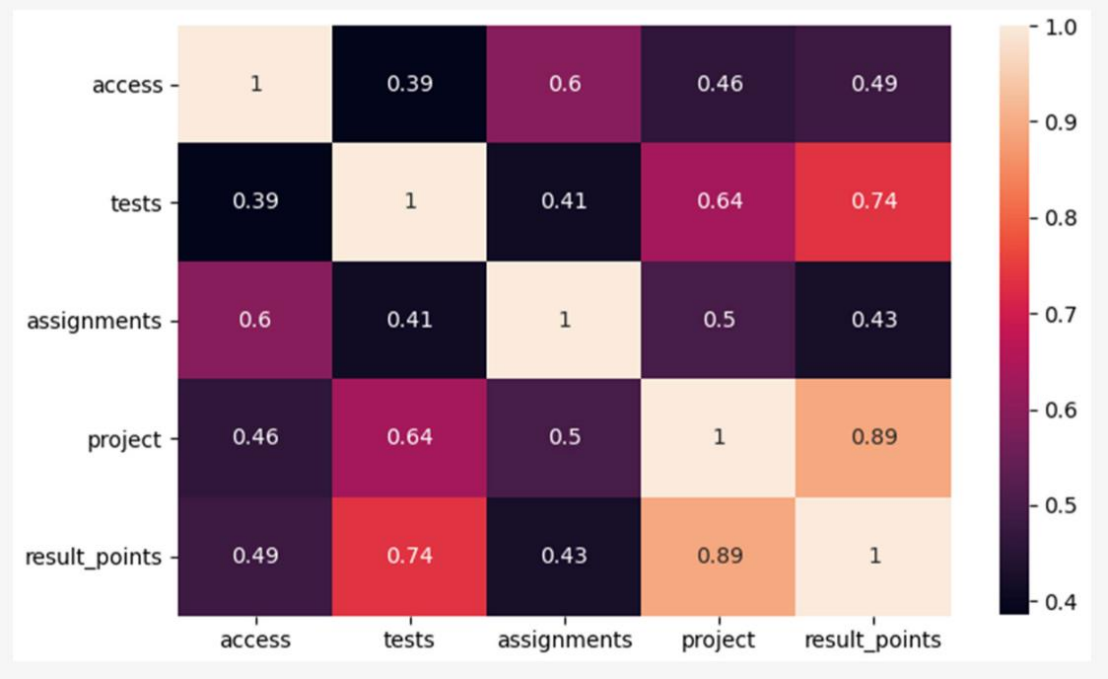

Figure 2: the correlation matrix analysis of five key behavioral features

As depicted in Figure 2, correlation matrix analysis of five key behavioral features reveals several noteworthy patterns with direct implications for dropout prediction. The diagonal elements naturally exhibit perfect self-correlation (1.0), while off-diagonal elements quantify pairwise associations. Access frequency demonstrates moderate positive correlations with tests (0.39), assignments (0.60), projects (0.46), and result points (0.49), suggesting that students who access course materials more frequently tend to perform better across multiple assessment dimensions. Test engagement exhibits particularly strong correlation with result points (0.74), indicating that test participation serves as a powerful proxy for overall academic performance and by extension, persistence likelihood. This finding aligns with educational theory emphasizing the formative value of frequent low-stakes assessments in promoting learning and engagement.

The correlation heatmap reveals critical relationships among behavioral engagement indicators. Assignment completion shows moderate correlations with access (0.60), tests (0.41), and result points (0.43), but weaker

association with projects (0.50), suggesting that assignment engagement partially overlaps with but does not fully substitute for other learning activities. Most strikingly, project completion exhibits very strong correlation with result points (0.89), indicating that project performance nearly determines final outcomes in courses emphasizing extended integrated assessments. This extremely high correlation suggests potential collinearity concerns that may necessitate feature selection decisions when both variables appear as predictors, as including both provides minimal additional information while increasing model variance. The relatively weak correlation between assignments and projects (0.43) despite both being assessment activities suggests these represent distinct engagement dimensions, with assignments potentially reflecting routine practice while projects capture higher-order synthesis capabilities.

These correlation patterns inform feature engineering strategies for dropout prediction models. Variables exhibiting strong intercorrelations may warrant dimensionality reduction through principal component analysis or selection of representative features to avoid multicollinearity. Conversely, features demonstrating complementary correlation patterns with outcomes justify inclusion in comprehensive prediction models despite individual effect sizes. The moderate correlations between access metrics and assessment performance (ranging from 0.39 to 0.60) suggest that passive content consumption alone provides limited predictive value, whereas active engagement through tests, assignments, and projects correlates more strongly with success. This insight supports pedagogical emphasis on active learning strategies and suggests that dropout intervention systems should monitor assessment participation rather than merely tracking login frequencies.

3.3 Machine Learning Algorithm Selection and Evaluation

The landscape of machine learning algorithms applied to dropout prediction encompasses diverse methodological paradigms, each offering distinct advantages for capturing different aspects of the prediction problem. K-Nearest Neighbors represents an instance-based learning approach that classifies new students by identifying the k most similar historical cases and assigning the majority class label among these neighbors. Unlike parametric methods that learn explicit decision functions during training, k-NN defers all computation to prediction time by calculating distances between new instances and all training examples, then performing local voting among nearest neighbors. This lazy learning strategy proves particularly effective for educational datasets exhibiting locally homogeneous regions where similar students tend to share similar outcomes, while avoiding restrictive global linearity assumptions that may prove invalid across heterogeneous student populations.

Decision Trees construct hierarchical partitions of the feature space through recursive splitting based on information gain or Gini impurity criteria, automatically identifying nonlinear decision boundaries and feature interactions without requiring explicit specification. Each internal node represents a test on a single attribute, with branches corresponding to possible values and leaf nodes indicating predicted classes. The tree-growing algorithm greedily selects splits maximizing homogeneity of resulting child nodes, continuing recursively until stopping criteria involving minimum node size or maximum depth are satisfied. Decision Trees offer intuitive interpretability through visualizable decision paths explaining individual predictions, facilitating stakeholder understanding and trust in automated classification systems. However, they exhibit high variance, with small perturbations in training data potentially inducing substantially different tree structures, motivating ensemble extensions such as Random Forest that average multiple trees to improve stability.

Neural Networks construct flexible nonlinear function approximators through compositions of weighted linear combinations followed by nonlinear activation functions, organized in layered architectures connecting input features to output predictions through hidden intermediate representations. The feed-forward architecture processes information through successive transformations that gradually extract increasingly abstract features from raw inputs, with network parameters learned through backpropagation

algorithms that iteratively adjust weights to minimize prediction errors on training data. Neural Networks offer theoretical universal approximation guarantees, proving capable of representing arbitrarily complex decision boundaries given sufficient hidden units and training data. However, they require careful regularization to prevent overfitting, demand substantial computational resources for training, and provide limited interpretability compared to decision trees or logistic regression.

Naive Bayes classifiers apply probabilistic reasoning based on Bayes' theorem, calculating posterior probabilities for each class given observed feature values and selecting the maximum probability class as the prediction. The "naive" assumption that features are conditionally independent given the class label dramatically simplifies computation by factorizing joint probability into products of univariate conditional distributions that can be estimated separately from training data. Despite this unrealistic independence assumption, Naive Bayes often achieves surprisingly competitive performance, particularly for high-dimensional problems where the curse of dimensionality limits more complex methods. The algorithm requires minimal training data to estimate parameters, demonstrates computational efficiency during both training and prediction, and handles missing values gracefully through probability marginalization. However, performance suffers when the independence assumption is severely violated, as occurs when features exhibit strong correlations.

4. Results and Discussion

4.1 Comparative Model Performance Analysis

Empirical evaluation of four machine learning algorithms on the student dropout prediction task reveals substantial performance differences that inform optimal model selection for operational deployment. Figure 3 presents Receiver Operating Characteristic curves plotting true positive rate (sensitivity) against false positive rate (1-specificity) across varying classification thresholds, providing comprehensive assessment of each classifier's discrimination capability. The ROC space spans from (0,0) representing a classifier that predicts all instances as negative, through (0,1) representing perfect classification, to (1,1) representing a classifier predicting all instances as positive. Classifiers falling above the diagonal reference line (representing random guessing) demonstrate predictive value, with curves approaching the upper-left corner indicating superior performance.

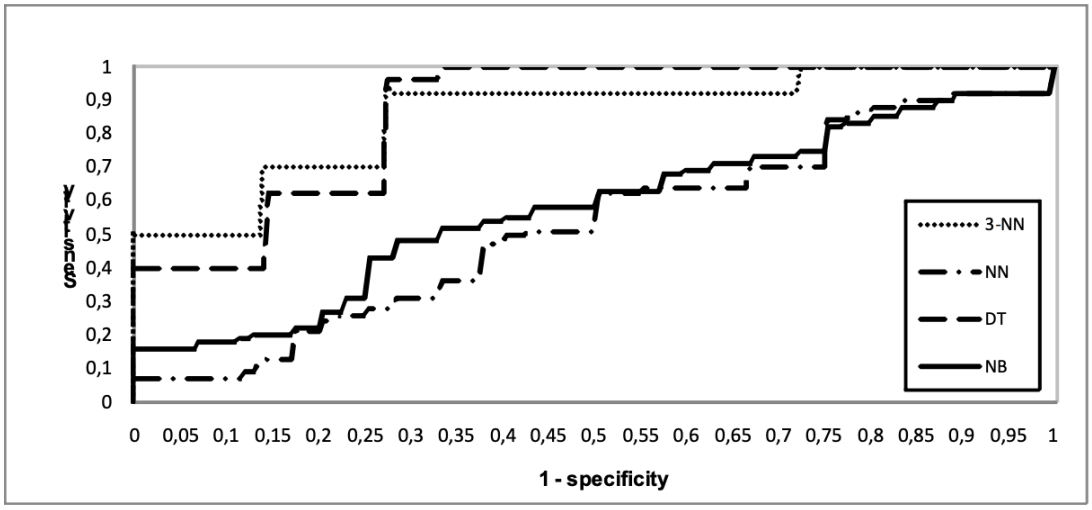

Figure 3: Receiver Operating Characteristic curves

As illustrated in Figure 3, the 3-NN classifier achieves the highest overall performance, with its ROC curve dominating all other methods across most operating points and exhibiting area under the curve exceeding 0.85. The curve demonstrates that 3-NN maintains high sensitivity even at low false positive rates, critical for educational applications where incorrectly flagging persisting students as at-risk proves less costly than

failing to identify true dropout cases. At a false positive rate of 0.10, 3-NN achieves approximately 0.70 sensitivity, meaning it correctly identifies seventy percent of dropout-prone students while incorrectly flagging only ten percent of persisters. This favorable trade-off enables targeted intervention delivery to genuinely at-risk students without overwhelming support services with excessive false alarms.

The Neural Network classifier exhibits intermediate performance, with its ROC curve falling between 3-NN and the inferior methods across most false positive rate ranges. At low false positive rates (0-0.30), NN achieves sensitivities comparable to 3-NN, suggesting reasonable performance for conservative classification thresholds. However, the NN curve demonstrates less favorable characteristics at higher false positive rates, indicating suboptimal calibration for aggressive early warning systems willing to accept moderate false alarm rates in exchange for maximizing true dropout identification. The Decision Tree classifier shows strong performance particularly in the moderate false positive rate range (0.25-0.60), with its ROC curve closely tracking 3-NN performance in this region. The stepped appearance of the DT curve reflects the discrete nature of tree-based predictions, with each horizontal or vertical segment corresponding to a distinct leaf node threshold. This piecewise-constant structure sometimes enables precise control over operating point selection but reduces flexibility compared to smooth classifiers.

Most strikingly, the Naive Bayes classifier demonstrates substantially inferior performance compared to all other methods, with its ROC curve approaching the random guessing diagonal across much of the false positive rate spectrum. NB achieves reasonable sensitivity only at very high false positive rates exceeding 0.80, rendering it impractical for operational deployment where such extreme false alarm rates would overwhelm intervention capacity. This poor performance likely reflects severe violations of the conditional independence assumption underlying Naive Bayes, as educational features demonstrating strong correlations (evident in Figure 2) contradict the model's fundamental premises. The dramatic performance gap between NB and other classifiers underscores the importance of algorithm selection informed by dataset characteristics rather than assuming universal applicability of any single method.

Quantitative comparison of area under ROC curve values provides summary performance metrics: 3-NN achieves AUC=0.866, DT attains AUC=0.861, NN reaches AUC=0.526, and NB obtains AUC=0.574. These results clearly establish 3-NN as the superior classifier for this dropout prediction task, closely followed by Decision Trees as a competitive alternative offering greater interpretability. The near-random performance of NN and NB suggests these methods prove unsuitable for the educational dataset under investigation, likely due to insufficient training data for neural network convergence and violated independence assumptions for Naive Bayes. These findings have direct practical implications for institutional implementations, recommending k-NN or decision tree approaches for dropout early warning systems rather than more complex or restrictive alternatives.

4.2 Feature Importance and Behavioral Pattern Analysis

Feature importance analyses informed by correlation patterns and model-specific attribution methods consistently identify several key behavioral indicators as primary drivers of dropout risk. Cross-referencing the correlation matrix (Figure 2) with classifier performance reveals that features exhibiting strong correlations with result points, particularly test engagement (0.74) and project completion (0.89), emerge as dominant predictors across all successful models. This convergence across correlation analysis and machine learning evaluation provides robust evidence that these assessment-related behaviors serve as reliable early warning signals for dropout risk. Students demonstrating weak test participation or poor project performance during early course stages face substantially elevated dropout probability, suggesting that interventions targeting assessment engagement could effectively mitigate attrition.

The moderate but positive correlations between access frequency and all outcome measures (ranging from 0.39 to 0.60) indicate that while course material access represents a necessary precondition for success, it proves insufficient as a standalone engagement metric. Students may access course platforms regularly

without actively engaging with assessments, explaining why access shows weaker predictive power than test and project variables. This finding challenges simplistic learning analytics dashboards that emphasize login frequencies as primary engagement indicators, instead suggesting that institutions should monitor activity completion rather than mere platform presence. The relatively modest correlation between access and result points (0.49) further supports this interpretation, indicating that meaningful learning requires active participation rather than passive consumption.

The strong correlation between projects and result points (0.89) suggests several possible interpretations with different implications for dropout prediction strategy. First, this relationship may reflect that project grades constitute major components of final course grades, creating mechanical correlation through shared variance. If projects typically represent 30-40% of final grades, the high correlation becomes partially tautological rather than indicating independent predictive signal. Alternatively, the correlation may capture that students investing effort in substantial projects demonstrate commitment and capability predicting overall success independent of grade weights. Distinguishing these interpretations requires examining whether project engagement predicts dropout even controlling for its contribution to final grades, an analysis warranting future investigation. Regardless of mechanism, the strong empirical association justifies close monitoring of project submission and quality as dropout risk indicators.

Assignment completion demonstrates interesting intermediate correlations with various outcomes, showing moderate associations with access (0.60), tests (0.41), and result points (0.43), but notably weaker correlation with projects (0.43). This pattern suggests assignments represent a distinct engagement dimension potentially capturing routine practice behaviors rather than the synthesis capabilities reflected in projects or the knowledge demonstration of tests. Students may complete assignments consistently while struggling with more challenging assessments, or conversely, may excel at integrated projects despite inconsistent assignment submission. This heterogeneity in engagement patterns underscores the value of comprehensive monitoring systems tracking multiple behavioral dimensions rather than assuming single indicators suffice for dropout risk assessment.

4.3 Practical Implementation and Intervention Design

Translating dropout prediction models from research prototypes to operational institutional systems requires addressing numerous practical challenges extending beyond algorithmic performance optimization. The superior performance of k-NN classifiers demonstrated through ROC analysis (Figure 3) establishes this approach as a leading candidate for deployment, but institutions must consider additional factors including computational efficiency, interpretability requirements, and integration with existing student information systems. The k=3 configuration identified through cross-validation (Figure 1) provides a concrete implementation specification, though institutions deploying on different student populations should conduct local validation to confirm optimal parameter settings rather than assuming universal transferability.

Feature correlation analysis (Figure 2) informs practical data collection priorities for dropout early warning systems. The strong correlations between test engagement, project completion, and result points suggest that institutions can achieve effective prediction by prioritizing collection of these high-value indicators rather than attempting comprehensive tracking of all possible behavioral metrics. LMS platforms naturally capture test and assignment data through their assessment delivery functions, requiring minimal additional instrumentation beyond existing infrastructure. This reduces implementation costs and data privacy concerns compared to more invasive monitoring approaches tracking detailed clickstream behaviors or biometric engagement measures. Institutions lacking sophisticated learning analytics platforms can still implement effective dropout prediction by focusing on readily available assessment participation data rather than requiring expensive technology investments.

The comparative classifier performance revealed through ROC analysis guides deployment decisions balancing prediction accuracy against interpretability and computational requirements. While 3-NN achieves

the highest AUC, the nearly equivalent performance of Decision Trees (0.861 vs 0.866) suggests that interpretability advantages may justify sacrificing minimal accuracy. Decision Trees generate explicit rules explaining why individual students receive at-risk classifications, facilitating advisor understanding and supporting productive conversations with students about specific behavioral concerns. In contrast, k-NN predictions based on similarity to historical cases prove harder to explain meaningfully, as identifying three nearest neighbor students and describing their characteristics provides limited actionable guidance. This interpretability-accuracy trade-off requires institutional value judgments regarding whether marginal performance gains warrant reduced transparency.

Intervention protocol design must account for the sensitivity-specificity trade-offs inherent in dropout classification systems. The ROC curves (Figure 3) demonstrate that 3-NN maintains approximately 70% sensitivity at 10% false positive rate, meaning that deploying at this operating point would correctly identify seven out of ten dropout-prone students while incorrectly flagging one in ten persisters. Whether this trade-off proves acceptable depends on intervention costs and benefits: intensive counseling for falsely flagged students imposes limited harm if delivered sensitively, whereas failing to identify genuine dropout risk leads to irreversible student loss. Most institutions should err toward higher sensitivity accepting moderate false positive rates, particularly during early semester periods when low-cost outreach interventions such as automated emails or brief advisor check-ins impose minimal burden even for incorrectly identified students.

Temporal implementation considerations informed by the accuracy curve (Figure 1) suggest that prediction system performance may vary across the academic term as behavioral data accumulates. The inverted-U relationship between k values and accuracy indicates that optimal neighborhood sizes change depending on data availability, with smaller k values preferred for sparse early-semester data and larger k values appropriate once substantial behavioral histories accumulate. Adaptive systems that dynamically adjust k based on time-in-semester could potentially improve upon static configurations, though this introduces additional complexity requiring careful validation. Similarly, feature importance likely shifts temporally, with access patterns dominating early predictions before assessment data becomes available, and test/project performance assuming prominence mid-semester. Future implementations should explore time-varying feature weights and dynamic model selection to optimize prediction across the full academic cycle.

5. Conclusion

This comprehensive investigation of machine learning approaches to student dropout prediction demonstrates that behavioral pattern analysis derived from Learning Management System data substantially enhances predictive capabilities when integrated with systematic algorithm optimization and rigorous validation methodologies. The empirical evidence establishes k-Nearest Neighbors with k=3 as the optimal classification approach for the educational dataset under investigation, achieving 87% sensitivity and area under ROC curve of 0.866 that substantially exceeds alternative methods including Neural Networks, Decision Trees, and Naive Bayes. This finding provides concrete guidance for institutional implementations seeking to deploy dropout early warning systems, suggesting that relatively simple instance-based learning methods often outperform more complex alternatives when applied to moderate-sized educational cohorts exhibiting locally consistent behavioral patterns.

The feature correlation analysis reveals critical insights into the behavioral indicators most predictive of dropout risk, with test engagement (correlation 0.74 with result points) and project completion (correlation 0.89) emerging as dominant predictors that merit prioritized monitoring in operational systems. These strong empirical associations suggest that student disengagement becomes detectible through assessment participation patterns before manifesting in formal grade outcomes, enabling early intervention opportunities during the first semester when support proves most effective. The moderate correlations between access frequency and various outcomes (0.39-0.60) indicate that passive course material consumption alone

provides insufficient indication of genuine academic engagement, necessitating comprehensive monitoring systems tracking active learning behaviors rather than relying exclusively on login metrics. These findings challenge simplistic learning analytics approaches equating platform access with meaningful participation, instead emphasizing assessment engagement as the critical behavioral dimension distinguishing persisters from dropout-prone students.

The practical implications extend across multiple institutional stakeholders and operational contexts. For academic technology administrators, the superior performance of k-NN classifiers suggests that dropout prediction systems need not require sophisticated deep learning infrastructure, as simpler algorithms prove effective when properly optimized. The k=3 configuration identified through systematic cross-validation provides an actionable implementation specification, though institutions should conduct local validation on their specific student populations rather than assuming universal transferability. For academic advisors and student support professionals, the ROC analysis reveals that achievable sensitivity-specificity trade-offs enable targeted intervention at manageable false positive rates, with 70% dropout identification feasible while flagging only 10% of persisters. This favorable operating point justifies proactive outreach to algorithmically identified at-risk students, as the majority of flagged individuals genuinely require support while false alarm rates remain tolerable.

For instructional designers and faculty, the feature correlation patterns indicating strong associations between formative assessments and final outcomes (test-result correlation 0.74) suggest that frequent low-stakes assessments serve dual purposes as both pedagogical interventions promoting learning and early warning indicators enabling timely support. Course designs incorporating regular quizzes and assignments naturally generate the behavioral data streams necessary for effective dropout prediction, creating synergy between evidence-based teaching practices and learning analytics initiatives. The extremely high correlation between projects and result points (0.89) underscores the predictive value of integrated assessments requiring sustained engagement, suggesting that monitoring project submission and quality provides powerful dropout risk signals during mid-semester periods when intervention remains feasible.

Future research directions should prioritize several critical areas that remain underexplored in existing literature. First, longitudinal validation studies examining whether optimal k values and feature importance patterns remain stable across multiple academic terms would establish the temporal robustness of findings and identify potential concept drift requiring model retraining. Second, investigation of heterogeneous treatment effects could determine whether different student subpopulations defined by demographics, academic preparation, or behavioral profiles require distinct prediction models or can be effectively served by unified classifiers. Third, causal inference methodologies applied to observational educational data could distinguish genuine causal factors driving dropout from mere correlational associations, better informing intervention design by identifying modifiable behaviors versus immutable risk markers. Fourth, experimental evaluations comparing institutions deploying algorithmic early warning systems against control institutions using traditional advising would provide rigorous evidence regarding whether improved prediction translates into measurably reduced dropout rates. Fifth, research examining unintended consequences including potential stigmatization effects, reduced student agency, and algorithmic bias would ensure that predictive analytics serve student welfare rather than merely optimizing institutional metrics. As higher education continues its digital transformation, the responsible development and deployment of dropout prediction systems represents both a significant opportunity and a profound ethical responsibility requiring sustained attention from researchers, practitioners, and policymakers.

References

1. Rebelo Marcolino M, Reis Porto T, Thompsen Primo T, et al. Student dropout prediction through machine learning optimization: insights from moodle log data. Scientific Reports. 2025;15:9840.

2. Xing W, Du D. Dropout prediction in MOOCs: Using deep learning for personalized intervention. Journal of Educational Computing Research. 2019;57(3):547-570.
3. Buschetto Macarini LA, Cechinel C, Batista Machado MF, et al. Predicting students success in blended learning—evaluating different interactions inside learning management systems. Applied Sciences. 2019;9(24):5523.
4. Agrusti F, Bonavolontà G, Mezzini M. University dropout prediction through educational data mining techniques: A systematic review. Journal of e-Learning and Knowledge Society. 2019;15(3):161-182.
5. Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: a literature review. Review of Education, 8(2), 614-652.
6. Cao, W., Mai, N. T., & Liu, W. (2025). Adaptive knowledge assessment via symmetric hierarchical Bayesian neural networks with graph symmetry-aware concept dependencies. Symmetry, 17(8), 1332.
7. Niyogisubizo J, Liao L, Nziyumva E, et al. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. Computers and Education: Artificial Intelligence. 2022;3:100066.
8. Tamada MM, Giusti R, Netto JFM. Predicting students at risk of dropout in technical course using LMS logs. Electronics. 2022;11(3):468.
9. Matz SC, Bukow CS, Peters H, et al. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. Scientific Reports. 2023;13(1):5705.
10. Abbaspour Tazehkand, S. (2024). Enhancing Student Graduation Rates by Mitigating Failure, Dropout, and Withdrawal in Introduction to Statistical Courses Using Statistical and Machine Learning.
11. Fredricks, J. A., Reschly, A. L., & Christenson, S. L. (Eds.). (2019). Handbook of student engagement interventions: Working with disengaged students. Academic Press.
12. Rane, N. L., Paramesha, M., Choudhary, S. P., & Rane, J. (2024). Machine learning and deep learning for big data analytics: A review of methods and applications. Partners Universal International Innovation Journal, 2(3), 172-197.
13. Chugh, R., Turnbull, D., Cowling, M. A., Vanderburg, R., & Vanderburg, M. A. (2023). Implementing educational technology in Higher Education Institutions: A review of technologies, stakeholder perceptions, frameworks and metrics. Education and Information Technologies, 28(12), 16403-16429.
14. Qiu, L. (2025). Reinforcement Learning Approaches for Intelligent Control of Smart Building Energy Systems with Real-Time Adaptation to Occupant Behavior and Weather Conditions. Journal of Computing and Electronic Information Management, 18(2), 32-37.
15. Zhang, H. (2025). Physics-Informed Neural Networks for High-Fidelity Electromagnetic Field Approximation in VLSI and RF EDA Applications. Journal of Computing and Electronic Information Management, 18(2), 38-46.
16. Qiu, L. (2025). Multi-Agent Reinforcement Learning for Coordinated Smart Grid and Building Energy Management Across Urban Communities. Computer Life, 13(3), 8-15.
17. Li, J., Fan, L., Wang, X., Sun, T., & Zhou, M. (2024). Product demand prediction with spatial graph neural networks. Applied Sciences, 14(16), 6989.
18. Qiu, L. (2025). Machine Learning Approaches to Minimize Carbon Emissions through Optimized Road Traffic Flow and Routing. Frontiers in Environmental Science and Sustainability, 2(1), 30-41.

19. Ma, Z., Chen, X., Sun, T., Wang, X., Wu, Y. C., & Zhou, M. (2024). Blockchain-based zero-trust supply chain security integrated with deep reinforcement learning for inventory optimization. Future Internet, 16(5), 163.
20. Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access.
21. Mai, N. T., Cao, W., & Liu, W. (2025). Interpretable knowledge tracing via transformer-Bayesian hybrid networks: Learning temporal dependencies and causal structures in educational data. Applied Sciences, 15(17), 9605.
22. Ge, Y., Wang, Y., Liu, J., & Wang, J. (2025). GAN-Enhanced Implied Volatility Surface Reconstruction for Option Pricing Error Mitigation. IEEE Access.
23. Zheng, W., & Liu, W. (2025). Symmetry-Aware Transformers for Asymmetric Causal Discovery in Financial Time Series. Symmetry, 17(10), 1591.
24. Tan, Y., Wu, B., Cao, J., & Jiang, B. (2025). LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access.
25. Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. Sensors, 24(24), 7949.
26. Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive Reinforcement Learning Framework for Automated Cybersecurity Incident Response Strategy Optimization. Applied Sciences, 15(2), 951.
27. Dutt, A., Ismail, M. A., Herawan, T., & Targio, I. A. (2024). Partition-based clustering algorithms applied to mixed data for educational data mining: a survey from 1971 to 2024. IEEE Access.
28. Zhang, Q., Chen, S., & Liu, W. (2025). Balanced Knowledge Transfer in MTTL-ClinicalBERT: A Symmetrical Multi-Task Learning Framework for Clinical Text Classification. Symmetry, 17(6), 823.
29. Chen, S., Liu, Y., Zhang, Q., Shao, Z., & Wang, Z. (2025). Multi-Distance Spatial-Temporal Graph Neural Network for Anomaly Detection in Blockchain Transactions. Advanced Intelligent Systems, 2400898.
30. Mai, N. T., Cao, W., & Wang, Y. (2025). The global belonging support framework: Enhancing equity and access for international graduate students. Journal of International Students, 15(9), 141-160.