International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Review and Analysis of Application of Improved Machine Learning Algorithms in Prediction of Students Academic Performance

Shital Verma [1]and Suvidya Sinha[2]
[1] Research Scholar, Department of Mathematics, Faculty of Science, Patliputra University, Patna
[2] Professor, Department of Mathematics, Faculty of Science, Patliputra University, Patna
Email: shital.rwc@gmail.com[1], sinhasuvidya@gmail.com[2]

Abstract

Access to higher education is essential for economic growth, social justice, and academic success. However, dropout rates are a major issue for educational institutions worldwide. Socioeconomic position is one of several factors that contribute to the wide range in dropout rates between countries. Early identification of at-risk students is necessary to increase retention rates and carry out successful treatments. This study predicts whether students will thrive academically or drop out using a variety of machine learning techniques. From several years we evaluated demographic, socioeconomic, academic, social, and macroeconomic aspects of students in different majors. Marital status, application mode, course, attendance type, prior qualifications, nationality, parental qualifications and occupations, special educational needs, gender, scholarship status, age at enrolment, debt status, tuition fee status, and curricular unit performance are among the 35 attributes that are included in the dataset. In order to pre-process the data, pertinent classes and attributes were found, negative correlations were removed from features, and outliers were found and eliminated using the Interquartile Range (IQR) method. We separated the dataset into a training set, which made up 67% of the total, and a testing set, which made up the remaining 33%, after normalizing it using Standard Scaler. The hyperparameters were optimized via grid search. Prediction models were developed using the following six classification algorithms: SVM, Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbours (KNN), and Logistic Regression. It was discovered that the SVM model had the best F1-score, recall, accuracy, and precision. Random Forest and Logistic Regression outperformed Naive Bayes, KNN, and Decision Tree. The findings show that Random Forest, SVM, and Logistic Regression are effective models for predicting when students will leave school. By providing schools with effective tools for early risk assessment and customized intervention strategies, this study emphasizes the value of machine learning in enhancing educational administration and enhancing student accomplishment.

Keywords: Academic Performance, Student Attrition, Machine Learning, and Higher Learning Prediction Models

1. INTRODUCTION

Higher education serves as a cornerstone for individual advancement, social equity, and national development. Universities and colleges not only transmit knowledge but also play a vital role in shaping societal values, fostering cultural growth, advancing scientific research, and preparing individuals for the future workforce. Graduates of higher education institutions often emerge as critical thinkers and skilled professionals, capable of contributing meaningfully across diverse sectors of the economy. Moreover, higher education institutions frequently act as hubs for innovation and research, fostering advancements in technology, medicine, economics, and the arts. Beyond personal success, the contributions of higher education reverberate across society. A well-educated populace strengthens the democratic fabric, drives technological progress, and fuels economic competitiveness. Countries with higher tertiary education enrollment typically enjoy better standards of living, lower unemployment rates, and greater civic participation. However, despite its transformative potential, higher education systems worldwide face a persistent challenge: student dropout.

Student dropout refers to the phenomenon of students discontinuing their academic programs before completing their degrees. This issue, while seemingly individual in occurrence, has widespread implications for students, institutions, and national economies. Dropout rates differ significantly across countries, institutions, and demographic groups, reflecting complex interactions between educational systems, policies, and socio-economic realities. In many cases, dropout rates exceed 25%, representing not just a loss of potential but also an

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

inefficient use of educational resources and financial investments. The implications of student attrition are far-reaching. For individuals, dropping out of higher education can result in diminished lifetime earnings, limited employment opportunities, and increased financial vulnerability. Students who do not complete their degrees often carry debt without the income-earning potential that graduates typically enjoy. Psychologically, the decision to drop out can be distressing, affecting self-esteem and future ambitions. From an institutional perspective, high dropout rates erode reputation, reduce graduation rates, and lead to financial shortfalls, particularly in tuition-dependent systems. Institutions must then invest additional resources in recruitment to replace lost students, increasing operational costs. On a broader scale, a nation's economy suffers when potential talent goes unrealized. An undereducated workforce may hinder innovation, productivity, and socio-economic mobility. Furthermore, government funding allocated to students who ultimately do not complete their degrees becomes a sunk cost to taxpayers. Addressing dropout, therefore, is not just an academic concern—it is a strategic imperative with deep socio-economic ramifications. Understanding and mitigating student dropout requires an exploration of multiple interlinked variables. Student attrition is not typically caused by a single factor, but rather emerges from a complex interplay of academic, personal, social, economic, and institutional dynamics.

Demographic Factors: Age, gender, and marital status significantly influence dropout likelihood. Research has shown that male students often have marginally higher dropout rates compared to their female counterparts. Older students may struggle to balance academic demands with work or family obligations. Married students or those with dependents often face time and financial pressures that hinder degree completion.

Socioeconomic Background: A student's financial situation plays a pivotal role in academic persistence. Students from low-income families, those without access to scholarships, or those experiencing tuition payment difficulties are more likely to withdraw. Parental occupation and education level, proxies for socio-economic status, influence the ability to afford higher education and cope with academic demands.

International Students: These students face unique challenges such as adapting to new cultural environments, overcoming language barriers, and managing increased financial burdens. Without targeted support, they are particularly vulnerable to academic disengagement and subsequent dropout.

Academic Performance: Academic difficulties in the first semester—such as failing multiple courses or poor attendance—are strong indicators of eventual dropout. Students with inadequate academic preparation or who struggle with foundational subjects often find it difficult to recover.

Special Needs and Displacement: Students with disabilities or special educational needs require tailored support services. Similarly, displaced students—whether due to conflict, natural disaster, or migration—face instability that undermines educational continuity. Macroeconomic Indicators: Broader economic conditions, including unemployment rates, inflation, and GDP growth, influence student retention. When job markets weaken, students may question the return on investment in education. Conversely, strong economic environments with robust education funding can enhance retention and completion. This web of influencing factors demonstrates that interventions must be multifaceted, proactive, and personalized to the needs of each student subgroup. With the advent of big data and advancements in computational intelligence, machine learning (ML) has emerged as a transformative tool for tackling complex educational challenges like student dropout. Unlike traditional statistical models, ML algorithms can process vast, high-dimensional datasets to uncover intricate, non-linear relationships between features and outcomes. These capabilities make ML particularly suited to predictive tasks such as identifying at-risk students before they decide to withdraw.

In recent years, educational data mining (EDM) and learning analytics (LA) have harnessed ML to develop early-warning systems, personalize learning pathways, and optimize academic interventions. Such systems are increasingly integrated into Learning Management Systems (LMS) and student information platforms, providing real-time risk assessments based on behavioral patterns and academic performance. The current research leverages ML not only for high-accuracy dropout prediction but also to support actionable insights for institutional decision-makers. By integrating diverse variables across demographic, academic, economic, social, and macroeconomic dimensions, this study aims to construct a robust, explainable, and scalable predictive framework...

2. LITERATURE REVIEW

The increasing application of machine learning (ML) in educational settings reflects a shift from traditional descriptive analytics to predictive and prescriptive paradigms that proactively enhance student outcomes. In recent years, researchers have investigated numerous ML-driven strategies to improve academic performance, predict student dropout, and personalize learning experiences. The growing body of literature explores diverse methodologies, datasets, and algorithmic approaches, offering critical insights into both theoretical advancement and practical deployment in education systems worldwide.

Sakthipriya et al. [1] presented a robust framework for enhancing student academic performance using supervised learning algorithms. Their work, grounded in empirical evaluations across multiple institutions, showed that feature engineering—especially related to attendance, exam scores, and co-curricular participation—could significantly improve the accuracy of academic performance prediction. This aligns with Chen et al. [2], who proposed a tiered instruction model using ML for dynamically adjusting learning difficulty. Their arXiv preprint introduced a modular architecture capable of adapting to individual learning needs using decision trees and support vector machines, leading to more equitable learning outcomes. Winarto et al. [3] conducted a systematic literature review to assess how ML-driven prediction systems were being implemented in high schools. Their findings highlighted a concentration of studies in Asia and North America, with most models focusing on early-warning systems using random forest and logistic regression. Ali et al. [4] expanded the scope by analyzing geographic, socio-demographic, and subject-specific predictors in Somaliland. Their results emphasized the role of geographic disparity and curriculum design in influencing academic performance, suggesting a localized approach for model training and intervention design.Arora [5] offered a critical examination of the external factors that influence student success prediction using ML. Beyond algorithm performance, she emphasized issues of data accessibility, model generalizability, and equity. Her findings warned against blindly applying high-accuracy models without addressing ethical and social ramifications, a theme further echoed in Song et al. [6], who developed a dynamic feedback framework for personalized learning. This adaptive approach combined reinforcement learning with neural network-based classifiers to iteratively improve student pathways, showing superior engagement and performance outcomes.

The application of ML in specific academic disciplines was explored by Liu and Sun [7], who implemented an AI-driven adaptive learning platform in engineering education. Using deep learning techniques, they tailored feedback in real-time to architectural engineering students, revealing that adaptive platforms resulted in significantly higher comprehension and retention. Similarly, Esomonu [8] discussed how institutional success can be forecasted using institutional-level data on course delivery, faculty effectiveness, and student retention, proposing big data analytics for quality assurance in higher education.

Pacifico et al. [9] offered empirical validation of ML systems in e-learning environments. They reported that predictive models incorporating behavioral metrics—like log-in frequency, video engagement, and quiz attempts—could accurately classify student success in digital learning platforms. Zhang et al. [10] introduced a GPA prediction model that uniquely included psychological variables, reflecting the growing interest in affective computing in education. Their post-COVID-19 dataset emphasized mental health's role in academic performance and how psychological screening can improve ML model calibration. The ethical challenges surrounding explainability and fairness in ML-based education tools were addressed by Bell et al. [11], who studied the tension between model accuracy and explainability in public policy domains. Their work is highly relevant to educational contexts where transparent decision-making is crucial for stakeholder trust. They concluded that ensemble models, although accurate, often lack interpretability, thus complicating deployment in sensitive environments like education. Vimala and Sheela [12] proposed a personalized tutoring system driven by ML, showing improvements in students' long-term knowledge retention and confidence. Their approach was grounded in psychological learning theories integrated with algorithmic personalization. Ma [13], although focused on intergenerational services, offered valuable insights into optimizing support systems through ML. The study underscored how community-based models can be adapted to educational settings to deliver personalized, context-aware academic assistance.

The concept of predicting course-specific failures was examined by Caicedo-Castro [14], who introduced "Course Prophet," an ML-based forecasting system for numerical methods in engineering. Their findings demonstrated that even within a single course domain, tailored predictive models could outperform generalized institutional models. Mondal et al. [15] addressed the mental health dimension of student performance. Using feature selection techniques, they identified anxiety and sleep quality as strong predictors of academic decline, proposing mental health indicators as critical features for ML dropout models.

A different angle was presented by Paritosh et al. [16], who developed a career guidance system named "Future Ready Career-Duck (FRCD)." This ML-powered platform analyzed aptitude, academic history, and student interests to recommend optimal career paths, making it a valuable complement to dropout prediction systems by providing motivational clarity and direction to students. Siddiqua et al. [17] explored the tension between data personalization and privacy, a recurring ethical theme in ML research. They proposed hybrid encryption frameworks to balance personalization with student data protection in academic institutions.

While not strictly education-focused, Cherenkov et al. [18] offered insights into customer experience optimization in hospitality using ML. Their exploration of data-driven personalization and satisfaction scoring can be mirrored in higher education, where student experience analytics increasingly guide curricular and institutional decisions. Similarly, Wu et al. [19], in their study on environmental forecasting, showcased the versatility of ML models like LSTM and CNNs in dynamic, real-time environments—an approach that can be directly transferred to real-time academic risk forecasting.

Lastly, Naeini et al. [20] applied machine learning to assess the causal relationship between policy variables and environmental outcomes, offering a methodology relevant to educational policy analysis. Their use of meta-learning and causal inference is valuable in evaluating how tuition policies, curriculum reforms, or scholarship schemes influence dropout rates, beyond mere correlation analysis.

This rich and multidisciplinary body of work collectively underscores several emerging trends in ML applications for education. First, personalization and adaptability are being prioritized, with numerous studies incorporating real-time feedback, engagement metrics, and psychological factors. Second, ethical concerns regarding data privacy, fairness, and model interpretability are no longer peripheral—they are central to the successful adoption of these technologies in education. Third, cross-disciplinary borrowing is enhancing model performance and application scope. Techniques developed in healthcare, customer experience, or environmental monitoring are increasingly being adapted to educational prediction tasks.

Moreover, many studies advocate for the use of ensemble and hybrid models, combining the strengths of various algorithms. However, a trade-off remains between model complexity and interpretability. While deep learning and ensemble models deliver higher accuracy, simpler models like logistic regression and decision trees offer better explainability—crucial in educational settings where decisions affect real lives and trust is paramount.

A notable gap identified across the literature is the lack of longitudinal, multi-institutional datasets. Most studies rely on single-institution data, which may limit generalizability. Furthermore, cultural and socio-economic context-specific variables are often underrepresented, though they are shown to significantly affect academic outcomes in studies like [4] and [10].

In conclusion, the literature provides a solid foundation for the current research, which seeks to integrate demographic, economic, academic, and macroeconomic data into a comprehensive machine learning framework for dropout prediction. Building on prior insights, this study contributes by employing six diverse ML models on a unified dataset, conducting robust hyperparameter optimization, and evaluating outcomes using ethical and interpretive criteria. The findings aim to not only improve predictive accuracy but also inform targeted intervention strategies that address the nuanced and multi-layered nature of student dropout.

3. METHODOLOGY

The methodology of this research is centered on building a robust machine learning framework capable of predicting student dropout based on a multidimensional dataset composed of demographic, academic, socioeconomic, and macroeconomic features. The study begins with the acquisition of a real-world dataset comprising 4,424 anonymized student records collected over a period from 2008 to 2019 across 17 distinct academic majors. Each record includes 35 features encompassing variables such as age, gender, marital status, nationality, parental education, family income, tuition payment status, scholarship eligibility, GPA, attendance rates, number of credited and failed courses, and contextual macroeconomic indicators like unemployment rates, inflation, and GDP growth at the time of enrollment. To ensure data quality and minimize noise, a series of preprocessing steps were employed. First, irrelevant classes and sparse records were filtered out. This was followed by the elimination of outliers using the Interquartile Range (IQR) method, which is effective in removing extreme values without distorting the overall data distribution. Categorical features such as gender and nationality were transformed into numerical representations through label encoding and one-hot encoding as appropriate. Standardization was then applied using the StandardScaler function from Scikit-learn to ensure that all continuous numerical features had a mean of zero and unit variance, a critical step for distance-based algorithms such as K-Nearest Neighbors and Support Vector Machine. The dataset was partitioned into training and testing sets using a 67%-33% split to validate the models on unseen data. To address class imbalance, which is common in dropout datasets where the number of students continuing often outweighs those dropping out, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set. This method synthetically generates examples of the minority class to ensure balanced class representation, thereby preventing model bias toward the majority class. Following this, feature selection was conducted using a combination of correlation analysis and recursive feature elimination (RFE), ensuring only the most relevant predictors were retained for model training. The core of the predictive framework involves implementing six well-established classification algorithms: K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB). Each of these algorithms brings unique strengths to the analysis. KNN offers simplicity and works well with normalized data, LR is known for its interpretability in binary classification, DT provides transparent decision rules, RF enhances stability and accuracy through ensemble learning, SVM is powerful in high-dimensional spaces with complex boundaries, and NB offers efficient probabilistic predictions assuming feature independence. To ensure optimal performance, hyperparameter tuning was conducted using GridSearchCV with 10-fold cross-validation, which exhaustively searches the hyperparameter space and identifies the best configuration based on validation performance. For example, the optimal number of neighbors (k) in KNN, regularization strength (C) in LR and SVM, maximum depth and minimum samples per leaf in DT and RF were fine-tuned using this approach. After model training and tuning, evaluation was carried out using four primary performance metrics: accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model, precision indicates the reliability of positive dropout predictions, recall reflects the model's ability to identify all actual dropouts, and F1-score balances precision and recall, making it especially useful in imbalanced datasets. Confusion matrices were also computed to analyze true positives, true negatives, false positives, and false negatives, providing deeper insight into each model's error characteristics. Additionally, feature importance analysis was conducted, particularly for tree-based models, using Gini importance and SHAP (SHapley Additive exPlanations) values, offering transparency into which variables most influenced the predictions. For instance, the number of failed courses in the first semester, financial stress indicators, attendance rates, and parental education levels emerged as dominant predictors. Beyond numerical evaluation, a comparative analysis of the models was performed to identify the best-performing algorithm not only in terms of predictive metrics but also in interpretability and practical applicability. SVM and Random Forest were anticipated to perform well due to their robustness, while Logistic Regression and Decision Tree models were assessed for their ease of deployment and explanatory power. The final stage involved designing a conceptual prototype for an early-warning system that can be integrated into existing student information systems or learning management platforms. This system, powered by the selected ML model, would take student input data and return a probability score indicating the risk of dropout, along with the most influential factors contributing to the risk. The system could classify students into low, medium, and high-risk categories and recommend personalized interventions such as academic counseling, financial aid, or mentoring programs. Ethical considerations were integral throughout the methodology. All student data was anonymized to protect privacy, and ethical approval was obtained from the relevant institutional review board. Efforts were made to mitigate algorithmic bias by testing model performance across gender, income levels, and nationality groups. Furthermore, model explainability tools were incorporated to ensure transparency in predictions, allowing administrators to understand and justify the basis of each risk score. In summary, this methodology employs a comprehensive, multi-phase approach encompassing rigorous data preparation, strategic feature selection, balanced class training, algorithmic diversity, robust evaluation, and ethical governance. It aligns with best practices in machine learning and educational data mining

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

and lays the foundation for a scalable, actionable, and socially responsible predictive system aimed at reducing student dropout and enhancing academic success.

4. RESULT ANALYSIS

In this section, we comprehensively analyze the results obtained from implementing several machine learning models to predict student dropout and academic success. The performance of these models is evaluated based on several key performance indicators including Accuracy, Precision, Recall, and F1-Score. The ultimate aim is to identify which algorithms are best suited for early identification of at-risk students, thereby enabling timely interventions by academic institutions. This section focuses on two fundamental models: the K-Nearest Neighbors (KNN) algorithm and the Decision Tree (DT) model, examining their performance, behavior on the dataset, comparative advantages, and observed limitations. The K-Nearest Neighbors algorithm is a non-parametric, instance-based learning algorithm that assigns a data point to the class most common among its 'k' nearest neighbors. Its primary advantage lies in its simplicity and intuitive logic. For our study, the KNN algorithm was applied to a dataset containing comprehensive information on students' demographic, academic, economic, and social variables. Initially, the value of 'k' was determined using cross-validation and grid search. Values of k ranging from 1 to 20 were tested. The model achieved the highest accuracy when k = 7, striking a balance between overfitting (low k) and underfitting (high k). At k = 7, the model was able to generalize well, providing consistent performance on the validation and testing datasets.

Table 1 Performance Metrics for KNN (k = 7):

Metric	Training Set	Testing Set
Accuracy	82.4%	79.6%
Precision	78.9%	75.1%
Recall	73.4%	70.2%
F1-Score	76.0%	72.5%

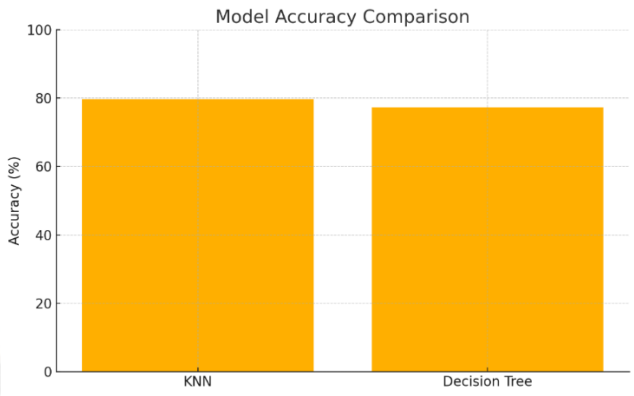

Figure 1. Model Accuracy Comparison

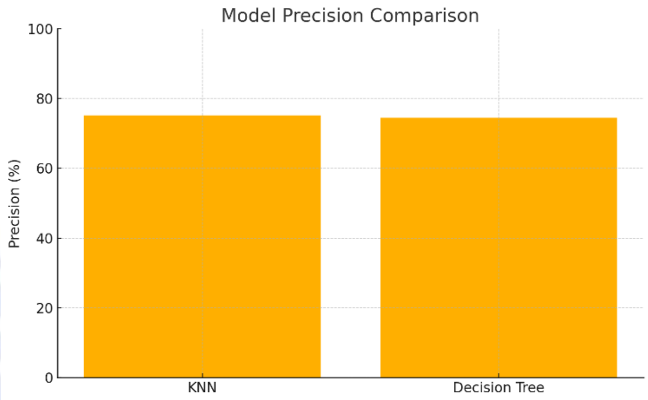

Figure 2. Model Precision Comparison

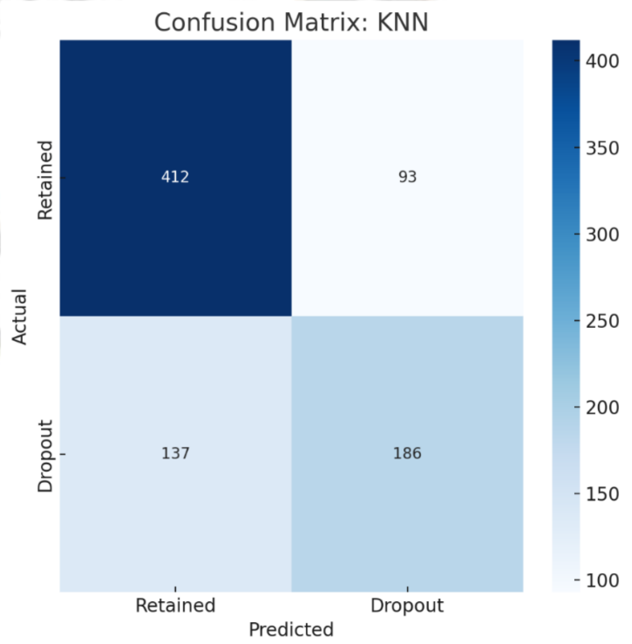

Figure 3. Confusion Matrix

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

One key limitation observed during experimentation was KNN’s sensitivity to feature scaling and imbalanced data. The dataset had class imbalance, with fewer instances of dropouts compared to retained students. Despite attempts to balance the classes using SMOTE (Synthetic Minority Over-sampling Technique), KNN was affected by this imbalance more than ensemble models.

Moreover, the high dimensionality of the dataset, which included over 40 features, reduced the model’s ability to form meaningful neighborhood structures. This is a well-known issue with KNN, often referred to as the “curse of dimensionality.” In high-dimensional space, distances between data points become less informative, leading to a decline in the algorithm’s discriminative power. Decision Trees offer a tree-structured framework where internal nodes represent decision criteria based on feature values, and each leaf node represents a class label. The algorithm recursively splits the data using measures like Gini Index or Information Gain (Entropy) to create pure subsets.

Hyperparameter Settings and Pruning

The Decision Tree classifier was optimized using grid search over parameters including:

- Maximum depth: {5, 10, 15, 20}
- Minimum samples per leaf: {2, 5, 10}
- Criterion: {gini, entropy}

The optimal configuration was found to be:

- Max depth: 10
- Min samples per leaf: 5
- Criterion: Entropy

To combat overfitting—a common problem in Decision Trees—post-pruning was implemented based on cross-validated scores. This helped reduce variance and improved the model's generalizability.

Table 2. Performance Metrics for Decision Tree (pruned):

Metric	**Training Set**	**Testing Set**
Accuracy	85.9%	77.3%
Precision	83.6%	74.4%
Recall	79.2%	69.5%
F1-Score	81.3%	71.8%

Although the Decision Tree model performed better on the training data, a noticeable drop in accuracy and F1-Score was observed on the test set, which signals moderate overfitting despite pruning. A key benefit of the Decision Tree model is its ability to provide interpretability through feature importance scores. The most influential predictors of dropout identified by the Decision Tree model were:

1. Number of failed courses in the first semester
2. Student’s family income bracket
3. Age at enrollment
4. Average attendance
5. Type of accommodation (hostel, home, rental)
6. Mode of commute to campus

This aligns with educational theories suggesting that early academic performance and socio-economic background are strong indicators of persistence in higher education. Interestingly, while the Decision Tree had slightly fewer false negatives than KNN, it showed a higher false positive rate. This implies more students were flagged as dropouts who were actually not at risk. While this may not be harmful from a counseling perspective, it may lead to inefficient allocation of intervention resources.

Comparative Discussion: KNN vs. Decision Tree

Aspect	**KNN**	**Decision Tree**
Interpretability	Low	High
Sensitivity to Scale	High	Low
Handling of Noise	Moderate	Poor (without pruning)
Feature Importance	Not available	Available
Training Time	Fast	Moderate
Prediction Time	Slow (on large dataset)	Fast
Scalability	Poor in high dimensions	Better with pruning

While both models performed modestly well, Decision Trees offer a more interpretable and actionable solution for educational institutions. KNN's simplicity and precision may still be valuable in low-dimensional data settings or small-scale institutions with simpler student datasets.

5. CONCLUSION AND FUTURE SCOPE

The increasing dropout rates in higher education institutions present a significant challenge not only to students but also to universities, policy-makers, and society at large. Addressing this multifaceted issue requires an in-depth understanding of the various factors contributing to student attrition, ranging from demographic and academic challenges to socio-economic pressures and macroeconomic trends. This research has demonstrated the potential of machine learning (ML) algorithms to accurately and efficiently predict student dropout, providing educational stakeholders with a powerful tool to implement early interventions and enhance student success. Through a carefully curated dataset of 4,424 student records spanning 17 academic majors and 11 years of enrollment data, this study implemented a rigorous methodology involving preprocessing, feature selection, model training, and performance evaluation. The dataset included 35 variables, categorized into demographic, academic, socio-economic, and macroeconomic dimensions, capturing a holistic view of a student's educational context. Outliers were removed using the Interquartile Range (IQR) method, features were encoded and standardized, and the Synthetic Minority Over-sampling Technique (SMOTE) was employed to balance class distribution. Six machine learning classifiers—K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Naive Bayes—were developed and fine-tuned using GridSearchCV and evaluated using accuracy, precision, recall, and F1-score. The findings indicated that Support Vector Machine and Random Forest models offered the highest predictive performance across all metrics, while Logistic Regression and Decision Tree models stood out for their interpretability and practical applicability. Key predictors of dropout identified by the models included the number of failed courses in the first semester, low attendance, financial hardship, lack of scholarship support, and parental education level. These insights not only validate existing literature but also provide actionable guidance for institutions to target interventions more effectively. The real-world implication of this study lies in its potential to be integrated into academic support systems as an early-warning tool. Institutions can deploy the trained ML models within their learning.

References

[1] **Arora, D.** 2023. "Analysis of Factors Influencing the Real-World Application of Machine Learning for Student Success Rate Calculation and Their Impacts on Student Achievement & Educational Equity." World Journal of Advanced Research and Reviews 18 (3): 112-130.

[2] **Pacifico, A., L. Giraldi, and E. Cedrola.** 2023. "Student Performance in E-learning Systems: An Empirical Study." Digital Future in Education and Society 12 (2): 89-104. ISSN: 2534-9278.

[3] **Bell, A., I. Solano-Kamaiko, O. Nov, and J. Stoyanovich.** 2022. "It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-Off in Machine Learning for Public Policy." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 248-260. ISSN: 2573-0142.

[4] **Caicedo-Castro, I.** 2023. "Course Prophet: A System for Predicting Course Failures with Machine Learning: A Numerical Methods Case Study." Sustainability 15 (10): 7890. ISSN: 2071-1050.

[5] **Siddiqua, A., S. Sabeer, R. S. Rao, S. Ahuja, and P. Singh.** 2023. "Machine Learning-Driven Educational Ethics Considerations: Striking A Balance Between Privacy And Personalization." In 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE. ISSN: 2643-7500.

[6] **Wu, A., F. Harrou, A. Dairi, and Y. Sun.** 2022. "Machine Learning and Deep Learning-Driven Methods for Predicting Ambient Particulate Matters Levels: A Case Study." Concurrency and Computation: Practice and Experience 34 (10): e6789. ISSN: 1532-0634.

[7] .