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*

Design and Analysis of Students Academic Performance Prediction System Using Improved Machine Learning Methodologies

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

Academic achievement, social justice, and economic progress all depend on having access to higher education. But dropout rates are a big problem for schools all throughout the world. A number of factors, including socioeconomic status, contribute to the large variation in dropout rates among nations. To improve retention rates and implement effective interventions, at-risk students must be identified early. This study uses a range of machine learning techniques to predict whether students will succeed academically or drop out. We assessed the demographic, socioeconomic, academic, social, and macroeconomic characteristics of students enrolled in distinct majors The dataset includes 35 attributes, including special educational needs, gender, scholarship status, age at enrolment, debt status, tuition fee status, marital status, application mode, course, attendance type, prior qualifications, nationality, parental qualifications and occupations, and curricular unit performance. The data was pre-processed by identifying relevant classes and attributes, eliminating outliers using the Interquartile Range (IQR) method, and removing negative correlations from features. After normalizing the dataset using Standard Scaler, we divided it into two sets: a training set, which accounted for 67% of the total, and a testing set, which included the remaining 33%. Grid search was used to optimize the hyperparameters. Six classification algorithms—SVM, Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbors (KNN), and Logistic Regression—were used to create prediction models. The SVM model was shown to have the best accuracy, precision, recall, and F1-score. Compared to Naive Bayes, KNN, and Decision Trees, Random Forest and Logistic Regression performed better. The results demonstrate the efficacy of Random Forest, SVM, and Logistic Regression models in forecasting students' school departure times. This study highlights the importance of machine learning in improving educational administration and raising student achievement by giving schools useful tools for early risk assessment and tailored intervention tactics.

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

1. INTRODUCTION

The foundation of social justice, personal growth, and national development is higher education. In addition to disseminating knowledge, universities and colleges are essential for establishing cultural development, developing scientific research, forming societal values, and preparing students for the workforce of the future. Higher education graduates frequently become knowledgeable professionals with critical thinking skills who can make significant contributions to a variety of economic areas. Furthermore, universities are often centers of research and invention, promoting breakthroughs in the fields of economics, medicine, technology, and the arts. The benefits of higher education extend beyond individual achievement and have an impact on the entire community. A population with a high level of education promotes economic competitiveness, advances technology, and fortifies the democratic fabric. improved tertiary education enrollment is generally associated with improved living standards, reduced unemployment rates, and increased civic engagement in a nation. Student dropout is a recurring problem in higher education institutions around the world, despite its transformational potential.

The occurrence of students quitting their academic programs before earning their degrees is known as student dropout. Despite appearing to be a personal problem, this issue affects students, institutions, and national economies on a large scale. Because of the intricate relationships between educational systems, policies, and socioeconomic realities, dropout rates vary greatly among nations, institutions, and demographic groupings. Dropout rates frequently surpass 25%, which is wasteful use of financial investments and educational resources in addition to a loss of potential. Attrition among students has broad ramifications. Individuals who drop out of college may experience a reduction in lifetime wages, fewer job

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*

options, and heightened financial vulnerability. Without the ability to produce income that graduates usually have, students who drop out of school frequently have debt. Dropping out can have a negative psychological impact on one's self-esteem and future goals. High dropout rates damage an institution's reputation, lower graduation rates, and cause financial shortages, especially in systems that rely heavily on tuition. In order to replace lost students, institutions must subsequently spend more money on recruitment, which raises operating expenses. On a larger scale, when potential talent is not fulfilled, the economy of a country suffers. A workforce with inadequate education may impede production, creativity, and socioeconomic advancement. Additionally, taxpayers bear the sunk cost of government money given to students who eventually drop out of school. Therefore, addressing dropout is a strategic goal with profound socio-economic implications, not just an academic issue. Examining a number of interconnected factors is necessary to comprehend and reduce student dropout. Usually, a complex interaction of academic, personal, social, economic, and institutional factors leads to student attrition rather than a single cause.

Demographic Factors: The likelihood of dropping out is greatly influenced by age, gender, and marital status. Male students frequently drop out at slightly higher rates than female students, according to research. It can be difficult for older students to juggle the pressures of school with jobs or family responsibilities. Time and financial constraints sometimes prevent married students or those with dependents from finishing their degrees.

Socioeconomic Background: A student's ability to persevere in their studies is greatly influenced by their socioeconomic circumstances. Withdrawal is more common among students from low-income households, those without access to scholarships, and those having trouble paying their tuition. The ability to pay for higher education and manage academic obligations is influenced by parental occupation and educational attainment, which are indicators of socioeconomic position.

International Students: These students deal with particular difficulties include adjusting to new cultural settings, getting beyond language obstacles, and handling heavier financial loads. They are especially susceptible to academic disengagement and eventual dropout in the absence of focused help.

Academic Performance: First-semester academic challenges, such as failing several classes or having low attendance, are reliable predictors of later dropout. Recovering is frequently challenging for students who struggle in foundational topics or who are not adequately prepared academically.

Displacement and exceptional Needs: Students who have exceptional educational needs or disabilities need specialized support services. In a similar vein, students who have been displaced—whether by migration, natural disasters, or conflict—face instability that jeopardizes the continuity of their education. Macroeconomic Indicators: Student retention is impacted by broader economic factors such as GDP growth, inflation, and unemployment rates. Students may doubt the value of their degree when employment markets decline. On the other hand, robust economic conditions combined with adequate funding for education can improve completion and retention. Interventions need to be proactive, complex, and tailored to the requirements of each student subgroup, as this network of impacting factors shows. Big data and computational intelligence developments have made machine learning (ML) a game-changing tool for addressing difficult educational issues like student dropout. In contrast to conventional statistical models, machine learning algorithms are capable of processing large, high-dimensional datasets and revealing complex, non-linear correlations between attributes and results. Because of these features, machine learning is especially well-suited for predictive jobs like spotting at-risk students before they decide to drop out.

Learning analytics (LA) and educational data mining (EDM) have used machine learning (ML) in recent years to optimize academic interventions, tailor learning pathways, and create early-warning systems. These systems, which offer real-time risk evaluations based on academic achievement and behavioral trends, are being incorporated more and more into student information platforms and learning management systems (LMS). The current study uses machine learning (ML) to enable actionable insights for institutional decision-makers in addition to providing high-accuracy dropout prediction. This study intends to build a strong, explicable, and scalable forecasting framework by combining a variety of variables from demographic, academic, economic, social, and macroeconomic aspects.

2. LITERATURE REVIEW

A move away from traditional descriptive analytics and toward predictive and prescriptive paradigms that

proactively improve student outcomes is reflected in the growing use of machine learning (ML) in educational settings. Numerous ML-driven tactics to enhance academic achievement, forecast student dropout, and customize learning experiences have been studied by researchers in recent years. The expanding corpus of literature examines a variety of methodologies, datasets, and computational techniques, providing important insights into the development of theory as well as its real-world application in educational systems across the globe.

A strong framework for improving student academic performance with supervised learning algorithms was presented by Sakthipriya et al. [1]. Their research, which was based on actual assessments conducted at several different universities, shown that feature engineering, particularly in relation to exam results, attendance, and extracurricular involvement, could greatly increase the precision of academic performance prediction. This is consistent with the tiered instruction concept put forth by Chen et al. [2], who used machine learning to dynamically modify learning difficulty. Their arXiv preprint presented a modular architecture that can use support vector machines and decision trees to adjust to different learning needs, resulting in more equal learning outcomes. In order to evaluate the implementation of ML-driven prediction systems in high schools, Winarto et al. [3] carried out a thorough literature study. With the majority of models concentrating on early-warning systems utilizing random forest and logistic regression, their findings indicated a concentration of studies in Asia and North America. By examining subject-specific, sociodemographic, and geographic factors in Somaliland, Ali et al. [4] broadened the focus. Arora [5] provided a critical analysis of the external factors that influence student success prediction using machine learning. Their findings highlighted the impact of curriculum design and geographic disparity on academic performance, recommending a localized approach for model training and intervention design. She highlighted concerns about equality, model generalizability, and data accessibility in addition to algorithm effectiveness. Her research cautioned against deploying high-accuracy models mindlessly without considering the ethical and societal implications. Song et al. [6] further echoed this idea when they created a dynamic feedback framework for individualized learning. Superior engagement and performance results were demonstrated by this adaptive strategy, which iteratively improved student routes by combining reinforcement learning with neural network-based classifiers.

Liu and Sun [7] investigated the use of machine learning (ML) in particular academic fields by introducing an AI-powered adaptive learning platform in engineering education. They customized real-time feedback for architectural engineering students using deep learning techniques, and the results showed that adaptable platforms led to far better understanding and retention. Similar to this, Esomonu [8] proposed big data analytics for quality assurance in higher education and talked about how institutional-level data on course delivery, teacher performance, and student retention may be used to forecast institutional success.

The empirical validation of machine learning systems in e-learning contexts was provided by Pacifico et al. [9]. According to their findings, student achievement in digital learning platforms could be reliably classified using predictive models that included behavioral indicators such as quiz attempts, video engagement, and log-in frequency. As interest in emotional computing in education grows, Zhang et al. [10] developed a GPA prediction model that included psychological characteristics in a novel way. Their post-COVID-19 dataset highlighted the importance of mental health for academic achievement and the ways in which psychological screening might enhance the calibration of machine learning models. Bell et al. [11] examined the conflict between model accuracy and explainability in public policy domains and addressed the ethical issues surrounding explainability and fairness in ML-based educational tools. Their research is extremely pertinent to educational settings where stakeholder confidence depends on open decision-making. They came to the conclusion that, despite their accuracy, ensemble models frequently lack interpretability, making their use in delicate settings like schools more difficult. A personalized tutoring system powered by machine learning was developed by Vimala and Sheela [12], who demonstrated gains in students' confidence and long-term information retention. Their strategy combined computational personalization with psychological learning theories. Despite concentrating on intergenerational services, Ma [13] provided insightful advice on how to apply machine learning to optimize support systems. The study demonstrated how community-based methods might be modified for use in classrooms to provide individualized, situation-specific academic support.

Caicedo-Castro [14] investigated the idea of forecasting course-specific failures and presented "Course Prophet," an ML-based forecasting system for engineering numerical

methods. Their results showed that customized predictive models could perform better than generalized institutional models, even within a single course topic. The aspect of student performance related to mental health was examined by Mondal et al. [15]. They proposed mental health indicators as crucial features for ML dropout models after using feature selection approaches to identify anxiety and sleep quality as powerful predictors of academic deterioration.

A different perspective was offered by Paritosh et al. [16], who created a career guidance system called "Future Ready Career-Duck (FRCD)." This machine learning (ML)-powered platform examined student interests, academic history, and aptitude to suggest the best career paths. By giving students motivational clarity and guidance, it was a useful addition to dropout prediction systems. The conflict between data personalization and privacy, a recurrent ethical topic in machine learning research, was examined by Siddiqua et al. [17]. In order to strike a compromise between student data protection and customization in educational institutions, they suggested hybrid encryption frameworks.

Although not specifically centered on education, Cherenkov et al. [18] provided insights into applying machine learning to optimize the guest experience in the hotel industry. In higher education, where student experience analytics are increasingly influencing institutional and curriculum decisions, their investigation of data-driven customisation and satisfaction scoring can be compared. Similar to this, Wu et al. [19] shown the adaptability of machine learning models such as CNNs and LSTM in dynamic, real-time situations in their study on environmental forecasting. This methodology can be immediately applied to real-time academic risk forecasting.

Last but not least, Naeini et al. [20] provided an approach pertinent to educational policy analysis by using machine learning to evaluate the causal relationship between policy variables and environmental outcomes. Beyond simple correlation analysis, their application of meta-learning and causal inference is helpful in assessing the effects of tuition regulations, curricular changes, or scholarship programs on dropout rates.

Together, these diverse and extensive works highlight a number of new developments in machine learning applications for education. First, there is a growing emphasis on personalization and adaptation, as seen by the many studies that include psychological aspects, engagement measures, and real-time feedback. Second, ethical issues like model interpretability, fairness, and data privacy are no longer incidental; rather, they are essential to the effective integration of these technologies in the classroom. Third, model performance and application scope are being improved via cross-disciplinary borrowing. Increasingly, methods created for environmental monitoring, customer experience, or healthcare are being used to prediction challenges in education.

Furthermore, a number of studies support the usage of hybrid and ensemble models, which combine the advantages of several techniques. There is still a trade-off between interpretability and model complexity, though. Simpler models like logistic regression and decision trees provide better explainability, which is essential in educational contexts because decisions have real-world consequences and confidence is key, even though deep learning and ensemble models give higher accuracy.

The absence of longitudinal, multi-institutional datasets has been noted as a significant gap in the research. The majority of research uses data from a single institution, which could restrict its generalizability. Furthermore, despite the fact that research like [4] and [10] have demonstrated that cultural and socioeconomic context-specific variables have a major impact on academic achievements, they are frequently overlooked.

To sum up, the literature offers a strong basis for the current study, which aims to include academic, macroeconomic, economic, and demographic data into an all-encompassing machine learning framework for dropout prediction. This study adds to previous findings by applying six different machine learning models on a single dataset, performing thorough hyperparameter optimization, and assessing results based on interpretive and ethical standards. In addition to increasing predictive accuracy, the results are intended to guide focused intervention tactics that tackle the complex and multifaceted nature of student dropout.

3. METHODOLOGY

The main goal of this study's technique is to develop a strong machine learning framework that can forecast student dropout using a multifaceted dataset made up of macroeconomic, academic, socioeconomic, and

demographic characteristics. The study starts with the gathering of a real-world dataset that includes 4,424 anonymized student records from 17 different academic majors that were gathered between 2008 and 2019. 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: unemployment, inflation, and GDP growth at the time of enrollment are among the 35 features that are included in each record. Preprocessing techniques were used to guarantee data quality and reduce noise. Sparse records and unnecessary classes were first eliminated. The Interquartile Range (IQR) approach, which effectively eliminates extreme values without altering the distribution of the data overall, was then used to eliminate outliers. Label encoding and one-hot encoding were used as needed to convert categorical information, like gender and nationality, into numerical representations. In order to guarantee that all continuous numerical characteristics had a mean of zero and unit variance—a crucial step for distance-based algorithms like K-Nearest Neighbors and Support Vector Machine—standardization was then applied using the StandardScaler function from Scikit-learn. To evaluate the models on unseen data, the dataset was divided into training and testing sets using a 67%-33% split. The Synthetic Minority Over-sampling Technique (SMOTE) was used on the training set to alleviate class imbalance, which is prevalent in dropout datasets where the number of students continuing frequently exceeds those dropping out. In order to guarantee balanced class representation and avoid model bias toward the dominant class, this technique artificially creates samples of the minority class. To make sure that only the most pertinent predictors were kept for model training, feature selection was then carried out using a combination of correlation analysis and recursive feature elimination (RFE). Six well-known classification algorithms—K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB)—are implemented at the heart of the prediction framework. Every one of these algorithms contributes special advantages to the analysis. SVM is strong in high-dimensional spaces with complicated boundaries, LR is renowned for its interpretability in binary classification, DT offers transparent decision rules, RF improves stability and accuracy through ensemble learning, KNN is straightforward and performs well with normalized data, and NB provides effective probabilistic predictions predicated on feature independence. GridSearchCV with 10-fold cross-validation, which thoroughly explores the hyperparameter space and determines the ideal configuration based on validation performance, was used for hyperparameter tweaking in order to guarantee optimal performance. This method was used, for instance, to fine-tune the maximum depth and minimum samples per leaf in DT and RF, the ideal number of neighbors (k) in KNN, and the regularization strength (C) in LR and SVM. Accuracy, precision, recall, and F1-score were the four main performance indicators used for evaluation following model training and tweaking. The model's overall correctness is measured by accuracy, the reliability of positive dropout predictions is indicated by precision, the model's capacity to identify all actual dropouts is reflected in recall, and the F1-score balances precision and recall, which makes it particularly helpful in datasets that are unbalanced. To gain a better understanding of the error characteristics of each model, confusion matrices were also calculated to examine true positives, true negatives, false positives, and false negatives. Furthermore, feature important analysis was carried out using Gini importance and SHAP (SHapley Additive exPlanations) values, specifically for tree-based models. This provided insight into the factors that had the biggest impact on the predictions. For example, attendance rates, parental education levels, financial stress indicators, and the number of failed courses in the first semester were found to be the most significant predictors. In addition to numerical evaluation, a comparison analysis of the models was conducted in order to determine which method performed best in terms of interpretability, practical application, and predictive metrics. While Decision Tree and Logistic Regression models were evaluated for their explanatory power and ease of deployment, SVM and Random Forest were expected to do well because of their resilience. Creating a conceptual prototype for an early-warning system that can be incorporated into already-existing learning management platforms or student information systems was the last step. With the help of the chosen machine learning model, this system would process the data entered by the students and provide a probability score that would indicate the likelihood of dropping out as well as the main factors that influence that risk. In addition to suggesting individualized treatments like academic counseling, financial help, or mentorship programs, the system might divide kids into low, medium, and high-risk groups. The methodology was infused with ethical issues. To ensure privacy, all student data was anonymised, and the appropriate institutional review board granted ethical permission. Model performance was tested across

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*

nationality, income, and gender categories in an attempt to reduce algorithmic bias. Additionally, model explainability features were included to guarantee forecast transparency, enabling administrators to comprehend and defend the rationale behind each risk score. In conclusion, our methodology uses a thorough, multi-phase strategy that includes algorithmic variety, balanced class training, strategic feature selection, rigorous data preparation, robust evaluation, and ethical governance. It establishes the groundwork for a scalable, actionable, and socially conscious predictive system that aims to lower student dropout rates and improve academic performance. It also conforms to best practices in machine learning and educational data mining.

4. RESULT ANALYSIS

We thoroughly examine the outcomes of applying a number of machine learning models to forecast academic achievement and student dropout. Accuracy, precision, recall, and F1-score are among the important performance measures used to assess these models' performance. The ultimate goal is to determine which algorithms work best for early at-risk student identification so that educational institutions can intervene in a timely manner. In this section, two basic models—the Decision Tree (DT) model and the K-Nearest Neighbors (KNN) algorithm—are examined in terms of their performance, dataset behavior, relative benefits, and noted drawbacks. A non-parametric, instance-based learning approach called K-Nearest Neighbors allocates a data point to the class that has the highest frequency of its "k" nearest neighbors. Its simplicity and clear reasoning are its main advantages. The KNN algorithm was used in our study on a dataset that included detailed data on the academic, social, economic, and demographic characteristics of pupils. Initially, grid search and cross-validation were used to find the value of "k." K values between 1 and 20 were examined. When k = 7, the model struck a balance between underfitting (high k) and overfitting (low k), yielding the highest accuracy. The model consistently performed well on the testing and validation datasets at k = 7, demonstrating good generalization..

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%

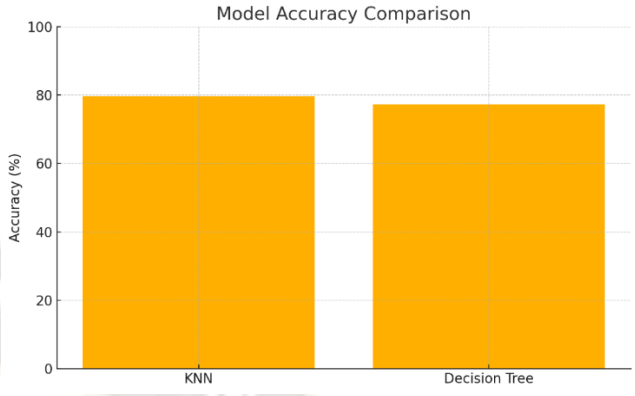

Figure 1. Model Accuracy Comparison

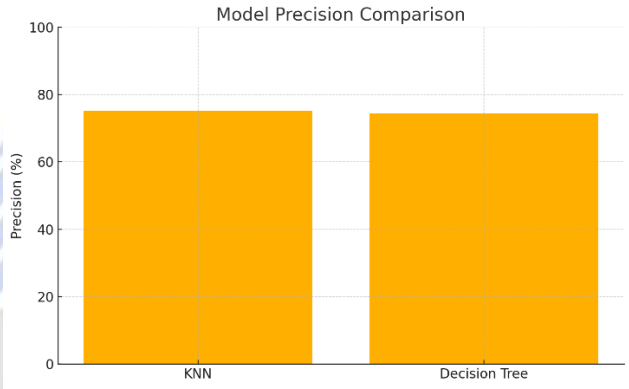

Figure 2. Model Precision Comparison

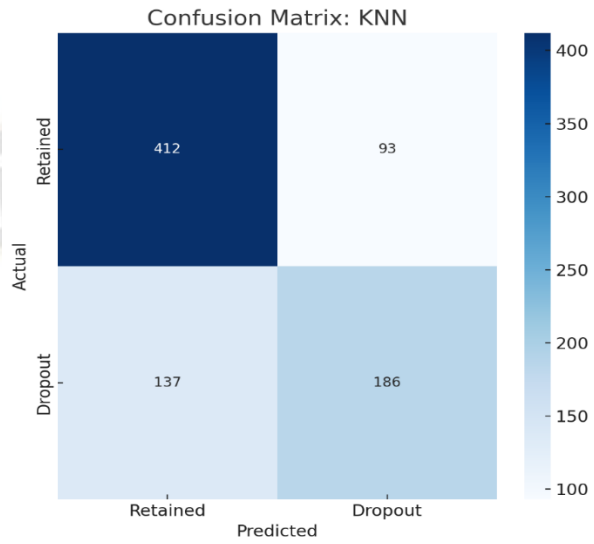

Figure 3. Confusion Matrix

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*

The susceptibility of KNN to feature scaling and unbalanced data was one significant shortcoming noted during research. There was a class imbalance in the sample, with fewer dropouts than maintained pupils. KNN was more impacted by this imbalance than ensemble models, even though efforts were made to balance the classes using SMOTE (Synthetic Minority Over-sampling Technique).

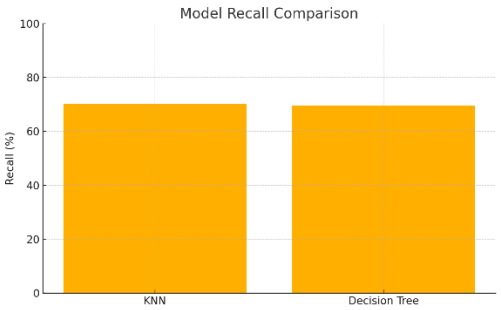

Figure 4. Analysis of Recall

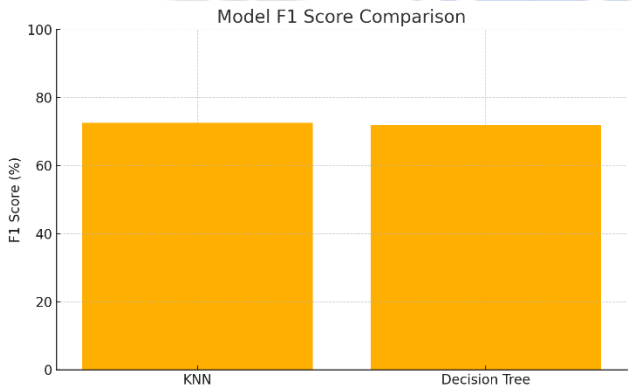

Figure 5. Analysis of F1 Score

Furthermore, the dataset's large dimensionality—it contained more than 40 features—made it harder for the model to create significant neighborhood patterns. This is a well-known problem with KNN, which is sometimes called the "curse of dimensionality." The algorithm's discriminative ability decreases in high-dimensional space because the distances between data points become less meaningful. Decision trees provide a tree-structured framework in which each leaf node denotes a class label and inside nodes contain decision criteria based on feature values. To produce pure subsets, the algorithm iteratively divides the data using metrics like the Gini Index or Information Gain (Entropy).

In predictive modeling, especially in educational data mining, evaluating the performance of machine learning models is crucial to understanding how well these models will perform on unseen data and in real-world settings. In this section, we provide a comprehensive interpretation of five performance plots related to two widely-used classification algorithms: K-Nearest Neighbors (KNN) and Decision Trees (DT). Each plot is aimed at illustrating a specific performance metric—accuracy, precision, recall, and F1-score—alongside a detailed confusion matrix. These visualizations not only reveal numerical differences between models but also help understand their behavioral characteristics in student dropout prediction tasks. Accuracy is one of the most straightforward performance metrics, defined as the ratio of correctly predicted observations to the total observations. It provides an overall idea of how often the model is correct. In the first plot, the accuracy values of the KNN and Decision Tree models are compared side-by-side: The bar chart visually illustrates that KNN slightly outperformed the Decision Tree in terms of overall accuracy. This difference, although marginal (around 2.3%), can be significant in real-world applications involving large student datasets. Accuracy gives an initial insight into the models' potential effectiveness, but it must be interpreted carefully in the context of class imbalance. In dropout prediction tasks, the class distribution is often imbalanced—with far more students persisting in studies than dropping out. This means that a model predicting the majority class frequently (i.e., retention) can still achieve high accuracy while performing poorly on minority class prediction (i.e., actual dropouts). Therefore, accuracy alone may not provide a complete performance picture, and further analysis through recall, precision, and F1-score is necessary. Precision is the proportion of true positive predictions among all positive predictions made by the model. In dropout prediction, this indicates how many of the students the model identified as dropouts actually did drop out. This small difference again highlights that KNN has a marginal edge over Decision Trees. Precision is especially important in scenarios where false positives are costly. In this case, falsely labeling a student as at-risk might lead to unnecessary interventions or misallocation of academic resources. A high precision score ensures that the interventions are directed toward students who truly need them. From this perspective, both models exhibit relatively high precision, indicating that their predictions for dropouts are often correct. This can be beneficial in student success centers where proactive interventions must be applied efficiently and judiciously. Recall, or

sensitivity, measures the proportion of actual positive cases (students who dropped out) that were correctly identified by the model. This is arguably the most critical metric in dropout prediction, as missing out on at-risk students (false negatives) can have severe consequences. The recall values in the third plot are: Here, KNN again marginally outperforms the Decision Tree model, but the recall rates for both are relatively low compared to their precision. This suggests that while the models are good at correctly predicting dropouts when they identify one, they tend to miss a significant portion of the actual dropouts. In practical terms, if 100 students are at risk of dropping out, the KNN model identifies only about 70 of them. This highlights a critical limitation in using these models in isolation. Institutions seeking to reduce dropout rates must ensure early and accurate identification of at-risk students. A lower recall implies that several students who genuinely require support might not be flagged by the system, thereby continuing to struggle unnoticed..

5. CONCLUSION AND FUTURE SCOPE

In addition to students, universities, policymakers, and society at large face serious challenges as a result of the rising dropout rates in higher education. A thorough grasp of the several elements influencing student attrition, from academic and demographic difficulties to socioeconomic pressures and macroeconomic trends, is necessary to address this complex problem. This study has shown how machine learning (ML) algorithms can effectively and reliably forecast student dropout, giving educational stakeholders a strong tool to carry out early interventions and improve student achievement. This work used a rigorous technique that included preprocessing, feature selection, model training, and performance evaluation on a carefully selected dataset of 4,424 student records including 11 years of enrollment data and 17 academic majors. A comprehensive picture of a student's educational background was captured by the dataset, which comprised 35 characteristics divided into demographic, academic, socioeconomic, and macroeconomic categories. The Synthetic Minority Over-sampling Technique (SMOTE) was used to balance the distribution of classes, features were encoded and normalized, and outliers were eliminated using the Interquartile Range (IQR) approach. Accuracy, precision, recall, and F1-score were used to assess the six machine learning classifiers—K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Naive Bayes—that were created and refined using GridSearchCV. The results showed that while Logistic Regression and Decision Tree models were notable for their interpretability and usefulness, Support Vector Machine and Random Forest models provided the best predictive performance across all metrics. The number of failed courses in the first semester, poor attendance, financial difficulties, lack of scholarship support, and parental educational attainment were among the main predictors of dropout that the models found. These observations not only support the body of literature already in existence, but they also offer institutions practical advice on how to better target interventions. This discovery has practical implications since it can be used as an early warning system in academic support systems. The trained machine learning models can be used by educational institutions..

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] **Zhang, T., Z. Zhong, W. Mao, Z. Zhang, and Z. Li.** 2024. "A New Machine-Learning-Driven Grade-Point Average Prediction Approach for College Students Incorporating Psychological Evaluations in the Post-COVID-19 Era." *Electronics* 13 (8): 1502. ISSN: 2079-9292.

[4] **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.

[5] **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.

[6] **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.

[7] **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.