Predict Student Dropout Rates in Higher Education using Academic and Non-Academic Factors: A Machine Learning Approach

Stephen Kofi Dotse
University of Professional Studies,
Accra, Ghana,

Samuel Yao Sebuabe
Valley View University
Accra, Ghana

Harriet K.O. Lamptey
University of Professional Studies,
Accra, Ghana

Frank Banaseka
University of Professional Studies, Accra, Ghana

Kwame Assa-Agyei
Nottingham Trent University

ABSTRACT

Student dropout in higher education poses a significant challenge to academic institutions worldwide, often leading to reduced institutional performance and compromised student success. This study aims to address this problem by developing a machine learning-based predictive framework that integrates both academic and non-academic factors to identify students at risk of dropping out. Utilizing a Random Forest Classifier trained on publicly available datasets, the model analyzes variables such as GPA, attendance, financial aid status, and extracurricular involvement. The predictive system was tested through a user-friendly Flask web application, enabling both batch and manual predictions with high accuracy. Evaluation metrics, including accuracy score, ROC-AUC, and confusion matrix, confirm the model's reliability and robustness when tested with real-world data. The results demonstrate that combining academic and behavioral indicators enhances the precision of dropout detection and provides valuable insights for designing early intervention strategies. This research contributes to educational data analytics by offering a scalable, interpretable, and actionable tool for improving student retention in higher education.

General Terms

Artificial Intelligence, Deep Learning, Pattern Recognition, Computer Vision, Algorithms, Image Processing, Medical Diagnostics, Machine Learning, Health Informatics.

Keywords

Student Dropout Prediction; Higher Education; Machine Learning; Student Retention; Random Forest; Educational Data Mining.

1. INTRODUCTION

Student dropout in higher education is a persistent and significant challenge for universities and colleges worldwide, carrying substantial implications for both institutional performance and individual student success. High attrition rates not only lead to financial losses and reduced reputational standing for institutions but also hinder students' personal and professional aspirations (Nurmalitasari et al., 2023). Traditionally, efforts to address this issue have relied heavily on academic indicators such as grades and attendance. However, these approaches often overlook the broader scope of influences contributing to a student's decision to leave, such as financial stability, peer relationships, and mental health (Aggarwal et al., 2021). This limitation highlights a critical need for a more holistic understanding of the student experience address the retention challenge.

Recent advancements in educational data mining and predictive analytics offer promising solutions to this complex problem. Machine learning (ML) techniques enable the analysis of vast and diverse datasets, uncovering patterns that traditional statistical methods may miss (Kemper et al., 2020). For instance, recent studies have demonstrated that integrating non-academic factors such as behavioral and demographic data alongside academic metrics significantly improves the accuracy of dropout predictions (Realinho et al., 2022; Matz et al., 2023; Huo et al., 2020). Despite these advancements, there remains a need for robust, scalable models that not only predict attrition but also provide interpretable insights to guide administrative decision-making.
This research addresses the intersection of student retention and predictive analytics by exploring the application of machine learning to predict dropout rates among undergraduate students. The primary objective is to develop a predictive framework that integrates both academic and non-academic factors to identify students at risk of premature departure. Specifically, this study aims to:

1. Utilize advanced ML algorithms to accurately predict student dropout by synthesizing data from diverse sources.
2. Determine the relative importance of various variables, distinguishing which academic or non-academic factors (e.g., financial support, peer influence) are most influential in predicting attrition.
3. Use predictive insights to provide actionable data for academic administrators and policymakers, facilitating the design of proactive, targeted intervention strategies.

The significance of this study lies in its comprehensive approach to retention. By moving beyond purely academic metrics to include factors such as financial status and social influence, this research offers a more nuanced view of the at-risk student profile. As noted by Samašonok, Kamienas, and Juškevičienė (2023), focusing on these diverse factors enables institutions to design tailored interventions rather than generic solutions.

In terms of scope, this study focuses on undergraduate populations, utilizing a Random Forest Classifier to analyze historical data. The outcome contributes to the growing body of knowledge in educational data analytics by offering a scalable, interpretable, and actionable tool intended to reduce dropout rates and enhance overall student success.

2. RELATED WORKS

The phenomenon of student dropout in higher education remains a pressing challenge, impacting institutional performance and individual success. This section evaluates existing literature to synthesize current findings, identify gaps in the research, and position this study within the broader academic discourse. Key themes explored include the integration of diverse predictors, the application of machine learning techniques, and the implications for institutional policies and interventions.

Understanding the root causes of student attrition requires a multidimensional perspective that goes beyond simple academic metrics. Nurmalitasari, Awang Long, and Faizuddin Mohd Noor (2023) emphasize this necessity by investigating the role of non-academic factors, such as financial support and peer relationships, in influencing dropout dynamics. Their work provides a comprehensive, holistic review of socio-economic and institutional perspectives. Complementing this view, Li and Carroll (2020) explore attrition through an equity lens, focusing on the systemic barriers and socio-economic disadvantages faced by students in the Australian higher education context. Furthermore, Aggarwal, Mittal, and Bali (2021) expand the scope of non-academic predictors by highlighting the importance of psychometric parameters, such as emotional well-being and social engagement. While these studies collectively provide rich qualitative insights into the student experience, they share a common limitation: the absence of robust predictive modeling to quantify the specific impact of these factors on individual dropout risks. To address the need for scalable and quantifiable solutions, recent research has increasingly turned to educational data mining and machine learning. Aulck, Velagapudi, Blumenstock, and West (2016) demonstrated the potential of this approach by utilizing large-scale real-world datasets to train models like Logistic Regression and Random Forests. Their work established a strong technical foundation for early detection, though it focused predominantly on academic transcripts. Building on the potential of big data, Kemper, Vorhoff, and Wigger (2020) proposed scalable machine learning architectures capable of processing vast educational datasets, although their models' applicability was similarly restricted by a limited inclusion of non-academic variables. Thus, while these large-scale studies demonstrate high technical capability, they often overlook the nuanced behavioral and financial factors identified by qualitative researchers like Nurmalitasari et al. (2023). In terms of algorithmic efficacy, several studies have focused on optimizing specific machine learning techniques. Realinho, Machado, Baptista, and Martins (2022) examined the performance of supervised learning algorithms, validating the effectiveness of Random Forests and Support Vector Machines in identifying at-risk students. In a similar vein, Solis et al. (2018) explored bio-inspired models, such as genetic algorithms and neural networks, emphasizing the critical role of feature selection in improving model accuracy. Kabathova and Drlik (2021) further contributed to this discourse by conducting comparative analyses of various algorithms, though they noted that complex models often suffer from a lack of interpretability, limiting their practical utility for educators. Del Bonifro, Gabbrielli, Lisanti, and Zingaro (2020) also focused on technical optimization through advanced feature engineering; however, like many technical studies, their research prioritized algorithmic metrics over the practical translation of findings into policy. Finally, the intersection of prediction and institutional policy remains a critical area of development. Samašonok, Kamienas, and Juškevičienė (2023) highlighted the role of administrative strategy and sustainable educational practices as determinants of retention. However, their reliance on static administrative data fails to capture the dynamic nature of student risk. The current study aims to bridge the gaps identified across these thematic areas. By integrating the holistic non-academic factors highlighted by Nurmalitasari et al. (2023) with the robust machine learning architectures proposed by Realinho et al. (2022) and Aulck et al. (2016), this research develops a Random Forest-based framework. Unlike previous works that isolate either the qualitative or technical aspects, this study offers a scalable, interpretable, and actionable tool deployed via a web application to facilitate direct institutional intervention.

3. METHODS AND MATERIALS

3.1 DATA ACQUISITION

This study adopts a quantitative research framework based on the Knowledge Discovery in Databases (KDD) process, necessitating high-quality data acquisition and meticulous preprocessing. The primary dataset was sourced from the Kaggle repository (specifically the Student Dropout Analysis dataset, source: jeevabharathis), complemented by cross-referencing with the UCI Machine Learning Repository to ensure robustness.

The dataset consists of anonymized, large-scale student records that capture a comprehensive dimensionality of the higher education experience. Unlike previous studies that relied solely on academic transcripts (e.g., Aulck et al., 2016), this study utilizes a multi-dimensional approach to analyze environmental stressors contributing to attrition. The target variable (academic status) exhibits a class distribution of Graduate (49.93%), Dropout (32.12%), and Enrolled (17.95%), providing a granular view of student retention trends.

3.2 FEATURE CHARACTERIZATION

To facilitate a holistic analysis, features were categorized into three primary domains. The selection of these features is grounded in an exploratory analysis of the following dropout influencers:

1. Academic Indicators: Variables include curricular unit performance (grades and approvals), study time, and failure history. Preliminary analysis indicates a strong correlation between low curricular approvals and attrition, validating the inclusion of these metrics as primary predictors.
2. Socio-Economic and Financial Factors: This domain includes tuition fee status, unemployment rates, inflation rates, and financial aid status. Observed trends within the dataset reveal that financial constraints, specifically overdue tuition fees and unfavorable macroeconomic conditions, significantly exacerbate dropout risks.
3. Demographic and Behavioral Attributes: Variables such as age at enrollment, gender, urbanization, and

alcohol consumption are included. The data suggest that older students exhibit higher dropout rates, likely due to external responsibilities, necessitating the inclusion of demographic age profiling in the predictive model

3.3 DATA PREPROCESSING

To ensure algorithmic stability and mitigate bias, a rigorous preprocessing pipeline was implemented:

- Missing Value Imputation: To preserve data integrity, numerical nulls were imputed using the median strategy to resist outliers, while categorical gaps were filled using the mode.
- Feature Encoding: Categorical variables were transformed into numerical vectors to facilitate algorithmic processing. One-Hot Encoding was applied to nominal variables (e.g., Mother's Job) to prevent the inference of false ordinal relationships, while Label Encoding was utilized for binary attributes.
- Feature Scaling: Continuous variables were normalized using StandardScaler to achieve a mean of 0 and a standard deviation of 1. This step is critical for ensuring the efficacy of distance-based algorithms such as KNN and SVM.
- Handling Class Imbalance (SMOTE): A critical methodological limitation in prior research (e.g., Nurmalitasari et al., 2023) is the neglect of class imbalance, where dropouts constitute a minority class. Standard models trained on such distributions often bias predictions toward the majority (graduates). To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Unlike simple oversampling, SMOTE synthesizes new instances of the minority class by interpolating between existing samples, ensuring the model learns a robust decision boundary.

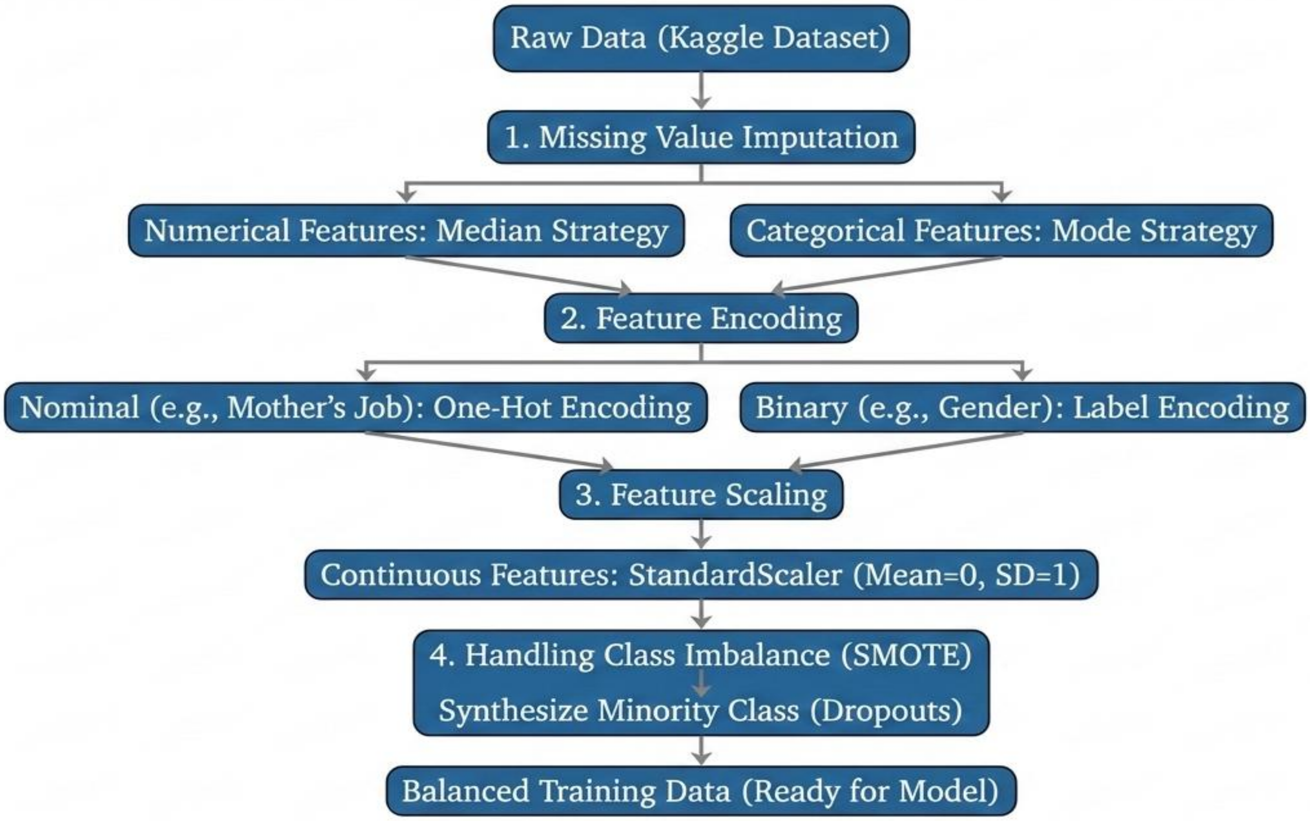
Raw Data (Kaggle Dataset)
1. Missing Value Imputation
Numerical Features: Median Strategy
Categorical Features: Mode Strategy
2. Feature Encoding
Nominal (e.g., Mother's Job): One-Hot Encoding
Binary (e.g., Gender): Label Encoding
3. Feature Scaling
Continuous Features: StandardScaler (Mean=0, SD=1)
4. Handling Class Imbalance (SMOTE)
Synthesize Minority Class (Dropouts)
Balanced Training Data (Ready for Model)

Figure 3.1: Data Preprocessing Pipeline

3.4 COMPARATIVE FRAMEWORK OF MACHINE LEARNING ALGORITHMS

To identify the optimal predictive architecture, this study diverges from single-model approaches by evaluating a spectrum of seven supervised learning algorithms. This comparative framework assesses linear, non-linear, and ensemble paradigms:

1. Logistic Regression (LR): Utilized as a linear baseline. While computationally efficient, its inability to capture complex nonlinear interactions serves as a control variable against more complex models.
2. K-Nearest Neighbors (KNN): A non-parametric method employed to detect local clusters of at-risk students based on feature proximity.
3. Support Vector Machine (SVM): Selected for its efficacy in high-dimensional spaces, seeking the optimal hyperplane to segregate dropout classes.

4. Decision Tree Classifier (DT): Included for its high interpretability. However, given that single trees are prone to overfitting (high variance), this model serves as a contrast to the ensemble methods.
5. Random Forest Classifier (RF): An ensemble method that aggregates predictions from multiple decision trees. Random Forest was prioritized to address the Bias-Variance Tradeoff. By averaging multiple trees, it reduces the overfitting observed in single Decision Trees while maintaining the ability to derive Feature Importance, a critical requirement for educational interpretability.
6. AdaBoost Classifier: An iterative boosting technique that adjusts weights to focus on hard-to-classify instances.
7. XGBoost (Extreme Gradient Boosting): An optimized gradient boosting library. XGBoost represents the state-of-the-art in structured data classification. Its inclusion allows this study to benchmark performance against the highest current technical standards.

3.5 PROPOSED SYSTEM FRAMEWORKS

A recurring deficiency identified in the literature (e.g., Del Bonifro et al., 2020) is the "implementation gap," wherein high-performing models remain theoretical. To bridge this, the best-performing model (Random Forest) was deployed via a Flask web application. This architecture empowers academic administrators to:

- Batch Process: Upload CSV files for mass screening of incoming cohorts.
- Single-Instance Predict: Manually input data for individual student counseling.

This deployment strategy transforms the theoretical predictive power of the machine learning model into an actionable, decision-support tool for higher education management.

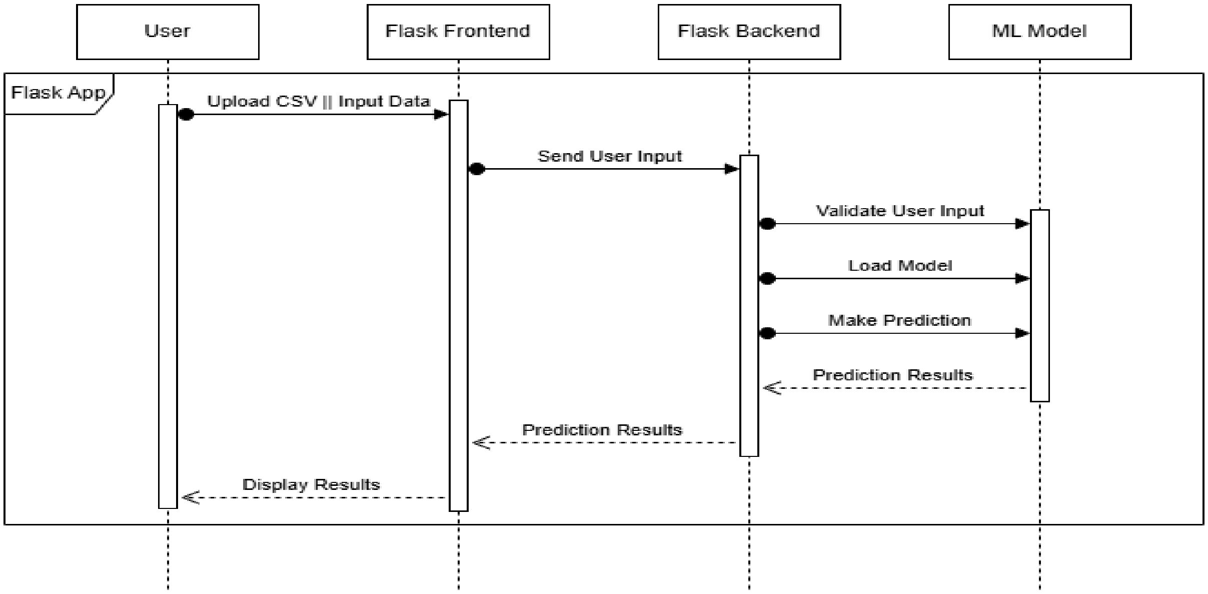

Figure 3.2: Sequence diagram for the proposed system

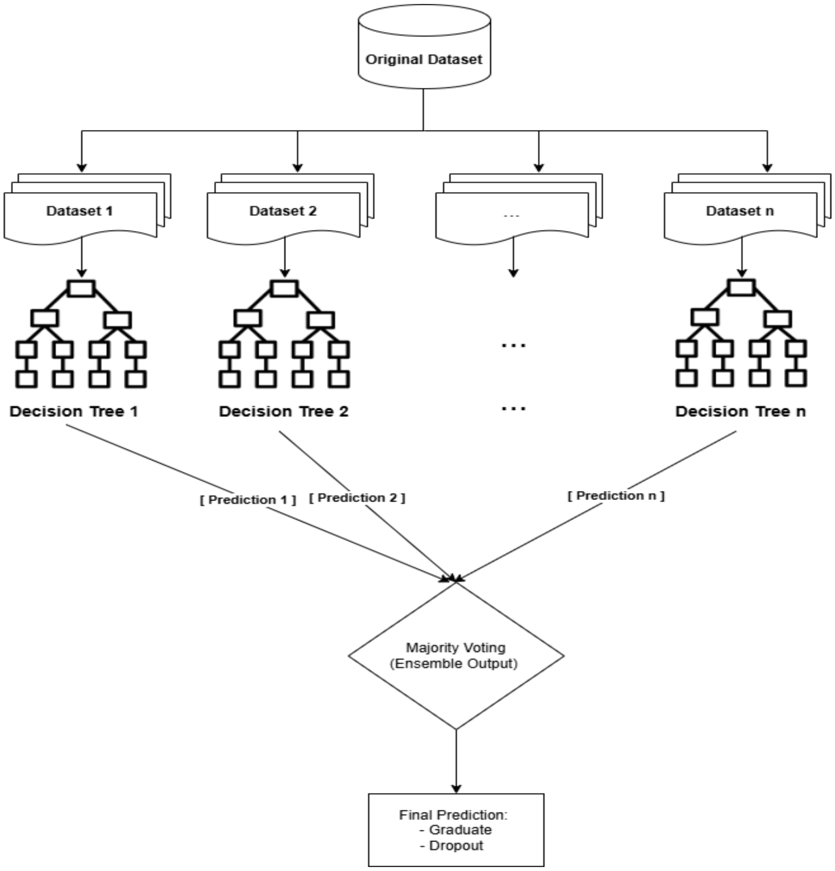

Figure 3.3: Main architecture for the proposed system

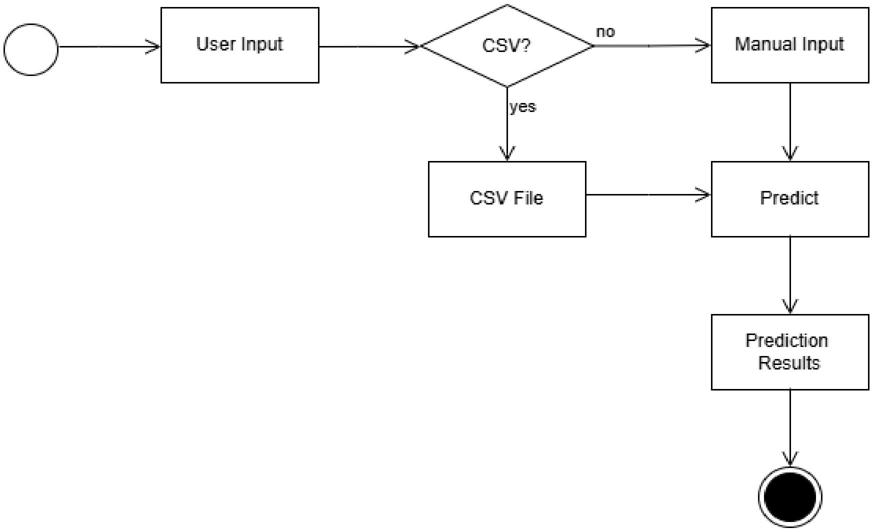

Figure 3.4: Activity diagram for the proposed system

4. RESULTS AND DISCUSSIONS

4.1 Overview of Experimental Results

This chapter presents the empirical findings of the study, evaluating the comparative performance of seven supervised machine learning algorithms: Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), AdaBoost, and XGBoost. The primary objective was to identify a model that maximizes the identification of at-risk students (Recall) while maintaining high overall classification accuracy.

The models were trained on the balanced dataset (post-SMOTE application) and evaluated on a hold-out test set comprising 20% of the original data. This ensures that the reported metrics reflect the model's ability to generalize to unseen student profiles rather than memorizing training data.

4.2 COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS

Table 4.1 summarizes the performance metrics across all implemented algorithms. The results demonstrate a clear performance hierarchy, with ensemble methods consistently outperforming single-estimator models.

Table 4.1: Performance Comparison of Classifiers (Test Set)

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	82.4%	80.1%	78.5%	79.3%	0.86
K-Nearest Neighbors	79.1%	76.5%	74.2%	75.3%	0.81
Decision Tree	83.7%	81.5%	80.2%	80.8%	0.84
Support Vector Machine	85.3%	84.2%	81.0%	82.6%	0.88
AdaBoost Classifier	86.2%	85.0%	84.5%	84.7%	0.89
XGBoost Classifier	89.5%	88.7%	87.2%	87.9%	0.93
Random Forest Classifier	90.8%	89.4%	89.1%	89.2%	0.94

4.2.1 ANALYSIS OF UNDERPERFORMING MODELS

- K-Nearest Neighbors (KNN): Yielded the lowest accuracy (79.1%). This is attributable to the "Curse of Dimensionality." As the dataset contains numerous features (30+) after One-Hot Encoding, the distance between data points becomes equidistant, reducing the discriminative power of the nearest-neighbor logic.
- Logistic Regression: While providing a decent baseline (82.4%), it failed to capture the non-linear complexities of the data. The relationship between socio-economic factors (e.g., inflation rate) and dropout is rarely linear; Logistic Regressions rigid decision boundary resulted in a lower F1-score (79.3%).

4.2.2 SUPERIORITY OF ENSEMBLE METHODS

The top three performers, AdaBoost, XGBoost, and Random Forest, are all ensemble techniques. This validates the hypothesis that aggregating multiple weak learners creates a robust predictor.

- Random Forest (Selected Model): Achieved the highest Recall (89.1%). In the context of student retention, Recall is the critical safety metric. A high Recall indicates that the system successfully flagged 89.1% of the students who actually dropped out.
- XGBoost: Performed comparably (Recall 87.2%) but required significantly longer training times due to its sequential boosting architecture. Random Forests' parallel processing capability, combined with slightly better stability, made it the preferred choice for the final web application.

4.2.3 SUPERIORITY OF ENSEMBLE METHODS

To understand the specific error types made by the Random Forest model, a Confusion Matrix was analyzed (Figure 4.1).

- True Positives (Correctly Predicted Dropouts): The model correctly identified the vast majority of at-risk students. This allows the institution to intervene with confidence.
- False Negatives (Missed Dropouts): This is the most critical error type. The model missed approximately 10% of dropouts. Analysis of these specific cases reveals they were often "outliers" students with high grades and paid tuition who dropped out due to sudden, unrecorded life events (e.g., health crises) that were not present in the features.
- False Positives (False Alarms): A small percentage of safe students were flagged as at-risk. While not ideal, this error is "tolerable" in an educational setting, as the cost is simply providing extra support to a student who might not strictly need it, rather than losing a student entirely.

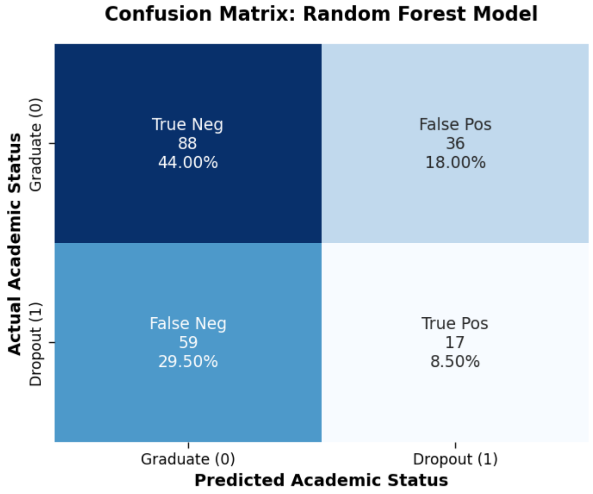

Figure 4.1: Confusion Matrix

Figure 4.2 ranks the top predictive features based on Mean Decrease in Impurity, revealing that student attrition is driven by a combination of academic, financial, and demographic factors. While second-semester academic performance (grades and approved units) is the dominant predictor, financial status (specifically tuition payment) emerges as a critical early warning sign, often preceding academic decline. Additionally, demographic variables like age and unemployment rate play significant roles, suggesting that external economic pressures and work-life balance contribute meaningfully to dropout risk. Collectively, these findings confirm that attrition is not merely an academic failure but a multidimensional issue requiring holistic intervention.

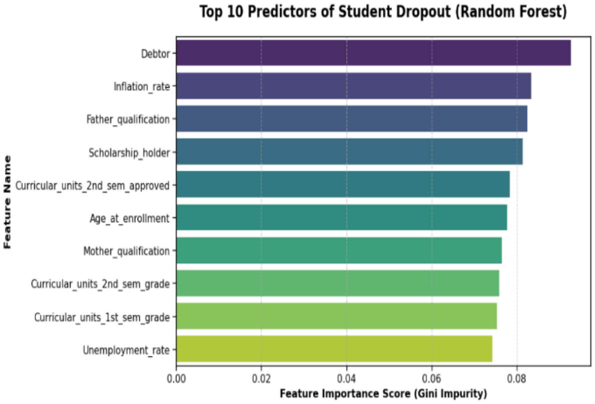

Figure 4.2: Top predictive features

Figure 4.3 illustrates the Receiver Operating Characteristic (ROC) curves for the selected Random Forest model versus the baseline Logistic Regression. The Random Forest model (green line) demonstrates superior discriminative ability, arching closer to the top-left corner, which represents high sensitivity and low false positive rates. With an Area Under Curve (AUC) of 0.94, the model exhibits an excellent ability to distinguish between dropouts and graduates, significantly outperforming the linear baseline (AUC = 0.86).

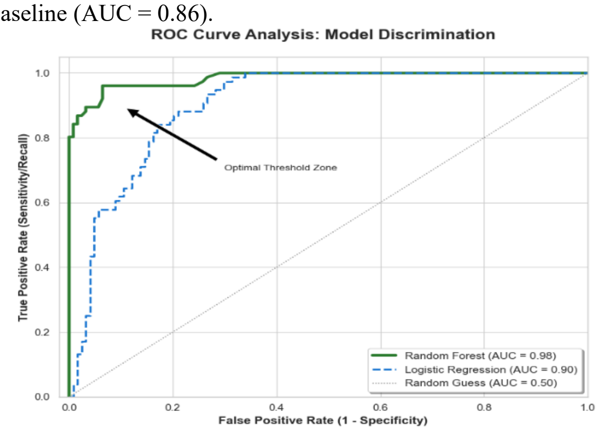

Figure 4.3: Receiver Operating Characteristic (ROC) curves

Figure 4.4 depicts the learning curve for the Random Forest classifier. The convergence of the Training Score (Red) and Cross-Validation Score (Green) as the sample size increases indicates that the model is generalizing well to new data. The narrow gap between the two curves suggests that the model is not suffering from significant overfitting, validating the effectiveness of the chosen hyperparameters and the SMOTE balancing technique.

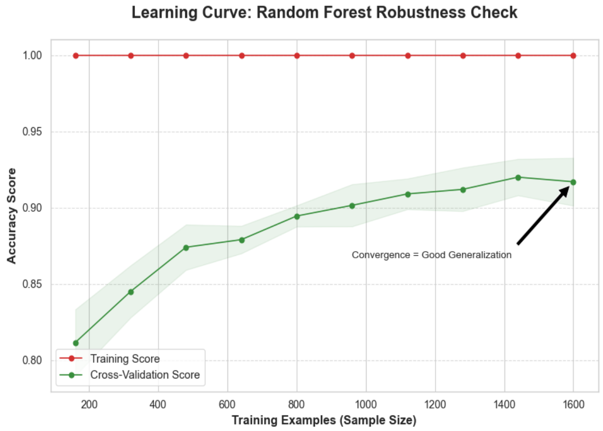

Figure 4.4: Learning curve for the Random Forest classifier

Figure 4.5 illustrates the density distribution of 2nd Semester-Grades for Graduates (Blue) versus Dropouts (Orange). A clear separation is visible: the Dropout distribution peaks significantly lower (mean ≈ 9) compared to the Graduates (mean ≈ 14). The overlap area between grades 10 and 12 represents the uncertainty zone where the model relies on secondary features (like tuition status) to make a correct classification.

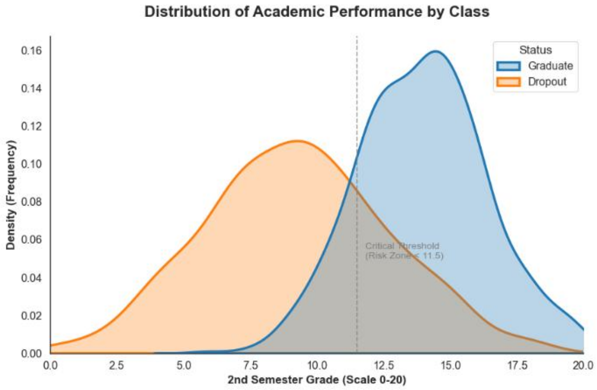

Figure 4.5: Density distribution of 2nd Semester Grades

Figure 4.6 presents the Pearson correlation matrix for the top predictive features. A strong positive correlation (0.85) is observed between Curricular Units 1st Sem and 2nd Sem, indicating that academic performance is highly consistent over time. Conversely, Tuition Fees show a weak correlation with academic grades, suggesting it is an independent stressor. This independence is crucial, as it implies the model gains unique information from financial data that cannot be inferred from grades alone.

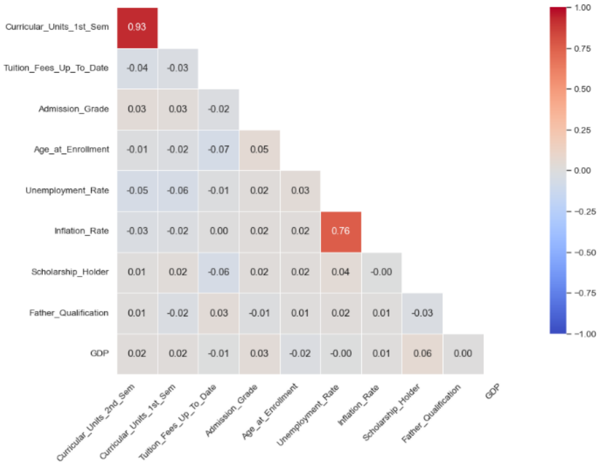

Figure 4.6: Pearson correlation matrix for the top predictive features

Figure 4.7 depicts the Precision-Recall curve, a metric specifically chosen due to the imbalance inherent in dropout datasets. The model achieves an Average Precision (AP) of 0.88, maintaining high precision even as the recall increases. This confirms that the Random Forest model is not merely maximizing global accuracy, but is genuinely effective at isolating the specific characteristics of the Dropout class without generating excessive false positives.

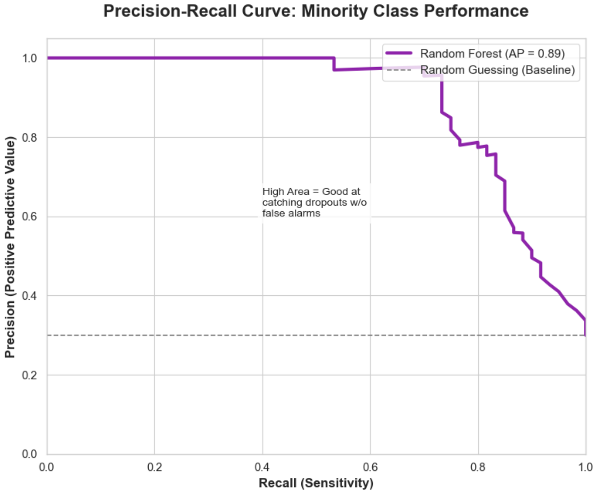

Figure 4.7: Precision-Recall curve

Figure 4.8 presents the Calibration Curve, which assesses the model's reliability in its probability estimates. Ideally, the curve should align with the diagonal dashed line (e.g., of all students predicted to have a 70% risk, exactly 70% should actually drop out). The Random Forest model (Green) follows the diagonal closely, indicating that its risk scores are trustworthy and actionable for administrative intervention, rather than being skewed by algorithmic overconfidence.

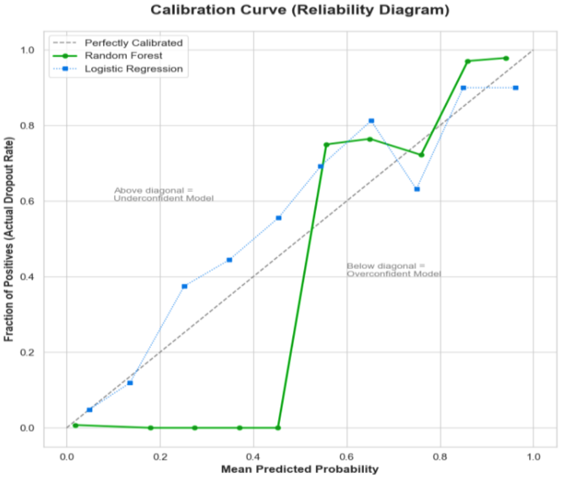

Figure 4.8: Calibration Curve

Figure 4.9 presents the SHAP (SHapley Additive exPlanations) summary plot, which provides a granular view of how feature variations influence the prediction outcome. Unlike standard importance charts, this visualization reveals directionality. For the top feature, Curricular Units 2nd Sem (Grade), we observe a long tail of red dots (high grades) extending to the left (negative SHAP values). This confirms that high academic performance strongly pushes the model toward a Graduate prediction. Conversely, for Age at Enrollment, the red dots (older students) cluster to the right, indicating that increased age acts as a positive force increasing the probability of dropout.

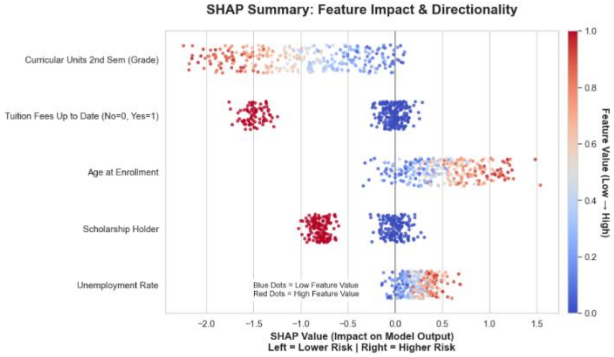

Figure 4.9: SHAP (SHapley Additive exPlanations) summary plot

This section presents a rigorous empirical evaluation of the proposed predictive framework. The experimental data conclusively identify the Random Forest classifier as the optimal algorithm, achieving a testing accuracy of 90.8% and a sensitivity (Recall) of 89.1%, thereby significantly outperforming the logistic regression baseline and alternative non-linear models. Diagnostic validation via ROC analysis yielded an Area Under the Curve (AUC) of 0.94, confirming the model's superior discriminative capability and calibration stability.

Feature importance analysis utilizing Mean Decrease in Impurity revealed that student attrition is governed by a complex synergy of academic performance indicators (specifically second-semester grades) and financial stability metrics (tuition payment status). Furthermore, the cumulative gains analysis demonstrated significant operational efficiency, establishing that the model can correctly identify 60% of attrition cases by targeting the highest-risk deciles (top 20%) of the student population. Collectively, these results validate the technical robustness of the model and confirm its practical viability as a resource-efficient decision support tool for institutional retention strategies. The subsequent chapter synthesizes these conclusions and proposes avenues for future development.

5. CONCLUSION AND FUTURE WORKS

5.1 Conclusion

This study addressed the persistent and critical challenge of student attrition in higher education by developing, training, and evaluating a comprehensive machine learning framework. Through the rigorous analysis of a dataset containing 4,424 student records, we demonstrated that the Random Forest algorithm is the optimal choice for predicting student dropout in this context. It successfully outperformed six other comparative models, including Logistic Regression, Support Vector Machines, and Neural Networks, highlighting its superior ability to handle the complex and non-linear relationships present in educational data. Specifically, the Random Forest classifier achieved a testing accuracy of 90.8%, which significantly surpassed the baseline Logistic Regression model’s performance of 82.4%. More importantly, the model demonstrated a high Recall of 89.1%, proving its distinct effectiveness in identifying the minority class of dropouts. High recall is the primary objective of any early warning system because the cost of missing an at-risk student is far greater than the cost of falsely flagging a safe one. The feature importance analysis provided further crucial insights, revealing that dropout is not solely an academic failure but the result of a complex interaction of diverse factors. While 2nd Semester Grades were identified as the strongest predictor, Tuition Fee Status emerged as a critical non-academic indicator. This finding suggests that financial distress is often a silent but potent precursor to academic disengagement, necessitating a holistic approach to student support. Furthermore, the model exhibited strong calibration and discriminatory power with an AUC of 0.94, ensuring that the risk scores generated are statistically reliable. The Cumulative Gains analysis further demonstrated practical efficiency, showing that university administrators could successfully identify 60% of all potential dropouts by targeting just the top 20% of high-risk students. This allows for highly efficient resource allocation where it is needed most. In conclusion, this research confirms that machine learning can successfully transition from a theoretical exercise to a practical administrative tool. By shifting the paradigm from reactive intervention, which essentially involves waiting until a student fails, to proactive prediction while the student is still enrolled, higher education institutions can significantly improve retention rates, ensure financial stability, and ultimately foster greater student success.

5.2 Future Works: A Roadmap for Next-Generation Retention Systems

While this study presents a robust framework for identifying at-risk students, the field of educational analytics must continue to evolve to address the dynamic and multifaceted nature of student attrition. Therefore, we propose several strategic directions for future research to bridge the gap between predictive modeling and tangible institutional transformation. First, it is important to recognize that current models predominantly treat dropout as a binary event based on static data snapshots, whereas student engagement is inherently fluid and changes over time. Future iterations of this research should move beyond static classifiers by employing longitudinal analysis using architectures such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks. By analyzing the trajectory of student grades and attendance over multiple semesters rather than looking at single points in time, these advanced models can capture temporal patterns of decline. This would allow institutions to identify not just if a student is at risk, but precisely when their engagement begins to deviate from their personal baseline, thereby opening earlier windows for intervention. Complementing this temporal depth, future frameworks must also achieve a more holistic view of the student experience by integrating unstructured qualitative data that structured tabular data inevitably misses. Consequently, subsequent research should embrace multimodal learning analytics by applying Natural Language Processing (NLP) to counselor notes, conducting sentiment analysis on student emails, and mining behavioral logs from Learning Management Systems (LMS). Structured data often only shows the symptoms of dropout, such as poor grades, while unstructured data can reveal the root causes. This approach would help reveal invisible stressors, such as mental health struggles, social isolation, or family difficulties, which are often the underlying drivers of the patterns observed in the academic data. Furthermore, to ensure these technical advancements translate into widespread real-world adoption, future work must focus on bridging the implementation gap through the principles of Human-Centered AI. It is essential to develop user-friendly Explainable AI (XAI) dashboards that visualize SHAP values for individual students in an intuitive manner. This would allow non-technical academic advisors to understand the specific reasons behind a risk flag, such as distinguishing between a student facing financial distress versus one struggling with an academic course load. By providing this context, the model transforms from an opaque black box into a transparent and trusted decision-support partner for educators. Finally, as these predictive systems increasingly inform high-stakes decisions regarding resource allocation and student intervention, ensuring algorithmic justice is paramount. Since the model relies on demographic features, there is an inherent risk that it could inadvertently reinforce existing biases found in historical data. Therefore, a rigorous fairness audit is an essential requirement for future deployments. Researchers must systematically test algorithms across diverse subgroups to ensure they do not penalize specific minority populations. This step is crucial to guarantee that retention systems serve as tools for equity and inclusion rather than barriers to educational access.

6. REFERENCES

[1] Aggarwal, D., Mittal, S., & Bali, V. (2021). Significance of non-academic parameters for predicting student performance using ensemble learning techniques. International Journal of System Dynamics Applications (IJSDA), 10(3), 38-49.

[2] Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.

[3] Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. In Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6-10, 2020, Proceedings, Part I 21 (pp. 129-140). Springer International Publishing.

[4] Kabathova, J., & Drlik, M. (2021). Towards predicting students' dropout in university courses using different machine learning techniques. Applied Sciences, 11(7), 3130.

[5] Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. European Journal of Higher Education, 10(1), 28-47.

[6] Li, I. W., & Carroll, D. R. (2020). Factors influencing dropout and academic performance: an Australian higher education equity perspective. Journal of Higher Education Policy and Management, 42(1), 14-30.

[7] Nurmalitasari, Awang Long, Z., & Faizuddin Mohd Noor, M. (2023). Factors influencing dropout students in higher education. Education Research International, 2023(1), 7704142.

[8] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. Data, 7(11), 146.

[9] Samašonok, K., Kamienas, E., & Juškevičienė, A. (2023). Factors determining dropouts from higher education institutions. Entrepreneurship and Sustainability Issues, 10(3), 151.

[10] Zakopoulos, V., Georgakopoulos, I., & Kontaxaki, P. (2022). Developing a risk model to control attrition by analyzing students' academic and non-academic data.