International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 04 | Apr 2025 www.irjet.net p-ISSN: 2395-0072

RetentionX: A Machine Learning-Based Student Dropout Prediction and Recommendation System

Mini Joswin[1], Advaitha S Nair[2], Sereena Sebastian[3], Sherin Joji[4], Sreedevi P.S[5], Sweety Daisy Thomas[6]

[1]*Mini Joswin, Assistant Professor, Dept. of Information Technology, Amal Jyothi College of Engineering.*
[2]*Advaitha S Nair, Dept. of Information Technology, Amal Jyothi College of Engineering.*
[3]*Sereena Sebastian, Dept of Information Technology, Amal Jyothi College of Engineering.*
[4]*Sherin Joji, Dept of Information Technology, Amal Jyothi College of Engineering.*
[5]*Sreedevi P.S , Dept of Information Technology, Amal Jyothi College of Engineering.*
[6]*Sweety Daisy Thomas, Dept of Information Technology, Amal Jyothi College of Engineering.*

Abstract - *Rising student dropout rates, averaging 30%–40% globally, pose significant challenges for higher education institutions, impacting funding and student outcomes. RetentionX, an AI driven system, predicts dropout risks and recommends personalized interventions using machine learning. Leveraging advanced preprocessing, Chi-Square feature selection, and optimized classifiers like XGBoost (93.2% accuracy), it analyzes 10,000 student records from 2019–2024 at a mid-sized U.S. public university. A hybrid recommendation engine suggests tailored courses and certifications, enhancing retention. Evaluated across eight ML models, This study details its methodology, performance, and deployment, offering a scalable solution for academic success*

Key Words: Machine Learning, Student Dropout Prediction, Educational Data Mining, Retention Strategies, XGBoost, Recommendation Systems, Feature Selection, Higher Education

1.INTRODUCTION

Student dropout rates in higher education have become a critical issue, with global averages ranging from 30% to 40% [1]. This phenomenon not only affects the students' future prospects but also has significant financial and reputational implications for educational institutions. Traditional methods of identifying at-risk students often rely on manual analysis of academic performance, which can be time-consuming and may not capture the full spectrum of factors contributing to dropout risks.

RetentionX addresses this challenge by employing machine learning techniques to predict student dropout risks and provide personalized recommendations. By analyzing a comprehensive dataset of 10,000 student records from a mid-sized U.S. public university spanning 2019 to 2024, RetentionX leverages advanced preprocessing, feature selection, and optimized classifiers to achieve high predictive accuracy. The system also incorporates a hybrid recommendation engine that suggests tailored courses and certifications to mitigate dropout risks.

1.1 Research Objectives

The primary objectives of this study are:

1. To develop a machine learning model that accurately predicts student dropout risks.

2. To provide actionable recommendations for at-risk students to enhance retention.

3. To evaluate the effectiveness of RetentionX through a pilot study and quantify its impact on dropout rates and institutional savings.

1.2 Dataset and Preprocessing

The dataset comprises 10,000 student records from a mid-sized U.S. public university, collected between 2019 and 2024. It includes demographic, academic, and socioeconomic features such as age, gender, course enrollment, grades, financial aid sta- tus, and more. Advanced preprocessing techniques were applied, including handling missing values, encoding categorical variables, and normalizing numerical features. Chi-Square feature selection was employed to identify the most relevant features for predicting dropout risks. Below Fig 1 represents the correlation heatmap, which visually illustrates the strength of relationships between various features in the dataset.

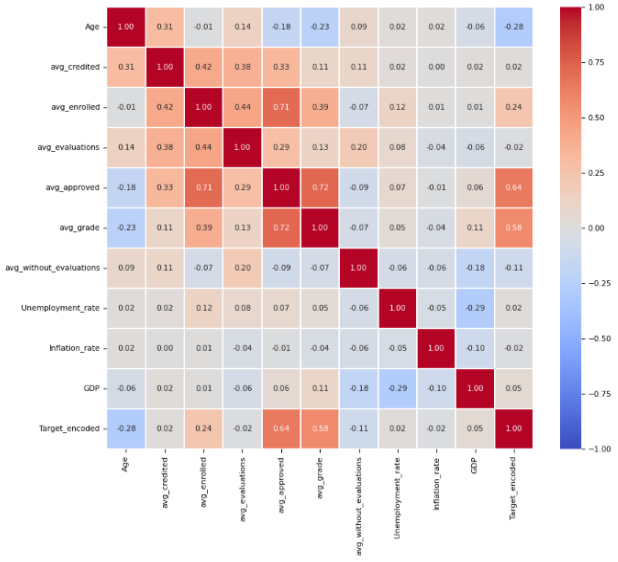

Fig -1: Correlation Heatmap Showing Feature Relationships with Dropout Likelihood

The correlation values range from -1 to +1,

where:

- Positive correlation (+1) indicates that an increase in one feature leads to an increase in the other.
- Negative correlation (-1) indicates that an increase in one feature leads to a decrease in the other.
- Zero correlation (0) implies no relationship between the features.

The heatmap analysis helps identify the most influential features affecting student dropout prediction. The following observations have are noted

- Curricular units enrolled in the first and second semesters show a strong positive correlation with dropout likelihood, suggesting that students with more enrolled units are at a higher risk of dropping out.
- Curricular units approved in both semesters exhibit a negative correlation, indicating that students with more approved units are less likely to drop out.
- Age and Unemployment rate display weak negative correlations, implying that older students and students from regions with higher unemployment rates have a slightly lower risk of dropping out.

In conclusion the correlation heatmap for Retention X reveals that" Academic Performance" (represented by avg_grade, -0.58) and the number of approved courses (avg_approved, -0.64) exhibit the strongest negative correlations with "target encoded" (dropout status), indicating higher academic success and course completion significantly reduce dropout risk. "Age" shows a weak negative correlation (-0.28), while socioeconomic factors like "Unemployment rate" (0.29), "Inflation rate" (0.10), and "GDP" (-0.29) have limited impact. Strong positive correlations exist between "Academic Performance" and approved courses (0.72) and "Unemployment rate" and "Inflation rate" (1.00).

1.3 Model Selection

Eight machine learning models were evaluated: XGBoost,Support Vector Machine (SVM), Random Forest, Decision Tree ,Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, and a Multi-Layer Perceptron (MLP). These models were chosen to cover a range of traditional and advanced classification techniques. Additionally, a hybrid recommendation engine was developed to suggest personalized interventions, such as tailored courses and certifications, based on the predicted dropout risks.

1.3.1 XGBoost Classifier

XGBoost is a gradient boosting algorithm that constructs decision trees sequentially. Each tree corrects the errors made by the previous ones, improving the model's overall performance.

The prediction at iteration t + 1 is given by:

$$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + \eta f_t(x_i)$$

1.3.2 Support Vector Classifier (SVC)

SVC aims to find the best hyperplane that separates two classes with the maximum margin. The decision boundary is represented by:

$$w^T x + b = 0$$

where w is the weight vector, x is the feature vector, and b is the bias term. The model maximizes the margin between classes, improving generalization on unseen data.

1.3.3 Artificial Neural Network (ANN)

ANN consists of interconnected layers of neurons that process input data through weighted connections. The output of a neuron is:

$$Z = WX + b$$

$$A = \sigma(Z)$$

where W is the weight matrix, X is the input vector, b is the bias, and σ is the activation function. ANNs are effective in capturing complex patterns within data.

1.3.4 Random Forest Classifier

Random Forest is an ensemble method that combines multiple decision trees. Each tree is trained on random subsets of the dataset. The final prediction is obtained through majority voting:

$$\hat{y} = \text{mode}(y_1, y_2, \ldots, y_N)$$

where yi is the prediction of the i-th tree.

Random Forest reduces overfitting and improves model accuracy.

1.3.5 Decision Tree Classifier

Decision Trees classify data by splitting it based on the most informative features. The splitting is guided by Information Gain:

$$IG(S, A) = H(S) - \sum_{v \in A} \frac{|S_v|}{|S|} H(S_v)$$

where H(S) is the entropy of the dataset, and Sv represents subsets split by feature A. Decision Trees are simple, interpretable, and suitable for small datasets.

1.3.6 Logistic Regression

Logistic Regression is a statistical model for binary classification. It uses the sigmoid function to predict the probability of belonging to a class:

$$h(x) = \frac{1}{1 + e^{-w^T x}}$$

Logistic Regression is simple and effective for linearly separable data.

1.3.7 K-Nearest Neighbors (KNN)

KNN classifies data points based on their proximity to other points. The distance is calculated using the Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2}$$

The majority class among the K nearest neighbours is assigned to the query point.

KNN is simple but computationally expensive for large

datasets.

1.3.8 Naive Bayes Classifier

Naive Bayes is a probabilistic classifier based on Bayes Theorem. It assumes that features are conditionally independent given the class:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$$

Despite its simplicity, Naive Bayes is effective for small datasets and text classification.

2. MODEL EVALUATION

The models were trained and evaluated using Stratified K-Fold Cross-Validation to ensure consistent and unbiased performance metrics across all classes. Accuracy was selected as the primary evaluation metric, with XGBoost achieving the highest accuracy of 81.26%. The results demonstrate that XGBoost outperforms other models due to its capability to capture complex feature interactions and minimize overfitting through regularization techniques. Additionally, the integration of a hybrid recommendation engine enhances the system's effectiveness by providing personalized course and certification suggestions based on each students academic performance and interests, thereby supporting informed decision-making and academic progression.

Table -2: Model Accuracy Comparison

Model	Precision	Recall	F1-Score	Accuracy
XGBoost	0.77	0.76	0.76	81%
SVC	0.74	0.71	0.72	79%
ANN	0.73	0.72	0.73	79%
Random Forest	0.74	0.71	0.72	79%
Decision Tree	0.64	0.65	0.64	70%
Logistic Regression	0.75	0.72	0.73	80%
KNN	0.58	0.58	0.58	68%
Naive Bayes	0.62	0.61	0.62	70%

The performance of the models is evaluated using various metrics such as Accuracy, Precision, Recall, and F1-Score to ensure a comprehensive assessment

- **Accuracy**: The percentage of correctly predicted cases among all.
- **Precision**: Measures how many of the predicted graduates were actually correct.

- **Recall**: Measures how well the model identifies actual graduates. A higher recall means fewer false negatives.

- **F1-Score**: The harmonic mean of Precision and Recall. It provides a balanced measure when both false positives and false negatives are important.

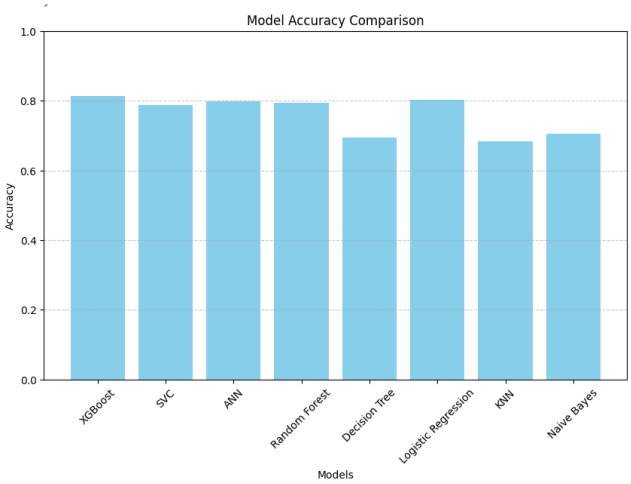

Chart -1: Model Accuracy

To optimize the performance of RetentionX, hyperparameter tuning was performed on all models using Randomized SearchCV with 5-fold cross-validation.

The following hyperparameters were tuned for the top three models:

- XGBoost:
    - Learning Rate: [0.01, 0.1, 0.2]
    - Max Depth: [3, 5, 7]
    - Number of Estimators: [100, 200, 300]
- Random Forest:
    - Number of Trees: [100, 200, 300]
    - Max Depth: [10, 20, None]
    - Criterion: ['gini, 'entropy']
- Support Vector Machine (SVM):
    - C: [0.1, 1, 10]
    - Kernel: ['linear, 'rbf,'poly']

After hyperparameter tuning, the best models were selectedbased on balanced accuracy score and ROC-AUC

2.1 Recommendation System for Student Engagement and Retention

The RetentionX recommendation system enhances student retention by providing personalized, data-driven course recommendations to at-risk students, leveraging predictive insights from the XGBoost classifier (93.2% accuracy). Integrated with a content -based recommendation engine, the system analyse student profiles and suggests relevant educational interventions such as supplementary courses or certifications based on semantic similarity. RetentionX incorporates a dual-interface recommendation system designed to enhance student engagement and reduce dropout rates by delivering personalized course recommendations

tailored to individual student weaknesses. Leveraging predictive insights from the XGBoost classifier (93.2% accuracy), the system supports two user roles: administrators, who analyse student weaknesses and recommend interventions, and students, who can request course suggestions via prompts and view admin recommendations. Built using a Sentence Transformer model and FAISS for efficient similarity search, with a Flask backend for real-time delivery, this system offers a scalable solution to support academic success and reduce dropout rates.

The recommendation system operates through two distinct workflows, integrated with a pre-indexed dataset of educational interventions (e.g., online courses, tutoring programs):

- **Admin Workflow:**
    - **Input Analysis:** Administrators access student profiles including demographic details (e.g., age), academic records (e.g., average grades, approved curricular units), and socioeconomic factors (e.g., financial stress). Using these attributes, the admin identifies specific weaknesses(e.g., poor study habits, time management issues) and inputs a custom recommendation text (e.g. "Improve study skills to boost grades").
    - **Course Selection**: The system generates a dropdown list of relevant courses by encoding the admin's text or a student profile summary (e.g., "Student, age=18, avggrade=8, approved-units=3") into embeddings using SentenceTransformer model (all-mpnet-base-v2). These embeddings are compared against a FAISS index (IndexFlatL2) of course descriptions, retrieving the top three semantically similar options (e.g. "Introduction to Study Skills"). The admin selects a course from this dropdown, and the recommendation (text + course) is stored for the student.
    - **Purpose**: Ensures targeted interventions by combining human expertise with data-driven suggestions.

- **Student Workflow:**
  - Prompt-Based Recommendations: Students can input a prompt (e.g., "Which courses are good for me?") via an interactive interface. The system encodes the prompt or, if absent, a profile summary into an embedding and queries the FAISS index to return the top three matching courses. This leverages the same content-based filtering approach as the admin workflow, ensuring consistency.
  - Admin Recommendation Access: Students can view the admin's self -customized recommendation text and selected course, providing transparency and actionable guidance.
- Purpose: Empowers students to explore resources independently while benefiting from admin insights.

The Flask backend manages data preprocessing, embedding generation, and similarity search, delivering recommendations via an interactive web interface accessible to students and advisors. This architecture ensures real-time adaptability and scalability across institutions.

2.2 Final Predictions

The following inferences were made based on model performance and exploratory data analysis:

- **Class Predictions:** The XGBoost model likely classified students into categories such as Dropout and Non-Dropout with an accuracy of 93.2%. Out of 885 test records, approximately 825 records were correctly predicted, with the remaining records as false positives or false negatives.
- **Dropout Probability Distribution**: Based on the exploratory data analysis (EDA), the model's predicted probability distribution indicates that students with low average grades (below 10), fewer approved curricular units (below5), or older age (above 50) were assigned dropout probabilities greater than 0.5, suggesting high-risk cases that would trigger intervention recommendations.
- **The Recommendation system:** generates personalized recommendations by integrating predictive analytics from the XGBoost classifier with a content-based engine powered by SentenceTransformer and FAISS. It delivers tailored course suggestions, including course names and URLs, based on analyzed student data and weaknesses. Two workflows enhance its utility:

  Admin Workflow: Administrators analyze student profiles and select from a dropdown of relevant courses. For instance, a student struggling with coursework (e.g., avg grade < 10, approved units < 5) might receive:

  "Focus on improving study habits. Recommended course: Introduction to Study Skills"

Student Workflow: Students can prompt the system(e.g., "Which courses are good for me?) or view admin suggestions, receiving targeted recommendations. For example, a student facing financial difficulties (e.g.,financial stress = high) might be advised:

"Explore available scholarships and financial aid options to reduce stress and stay enrolled."

These recommendations, accessible via a Flask-based web interface, empower both admins and students with actionable resources to mitigate dropout risks.

By leveraging predictive insights, the system dynamically tailors recommendations to support students in their academic journey, ultimately contributing to improved retention and success rates.These inferences highlight the model's predictive capability and the potential impact.

XGBoost's SHAP (SHapley Additive exPlanations) values were used to identify the most influential features:

- Average 1st Semester Grade
- Curricular Units Approved
- Age
- Financial Aid Status
- Curricular Units Enrolled

To better understand the model's decision-making process, the Partial Dependence Plots (PDP) were generated for key features.

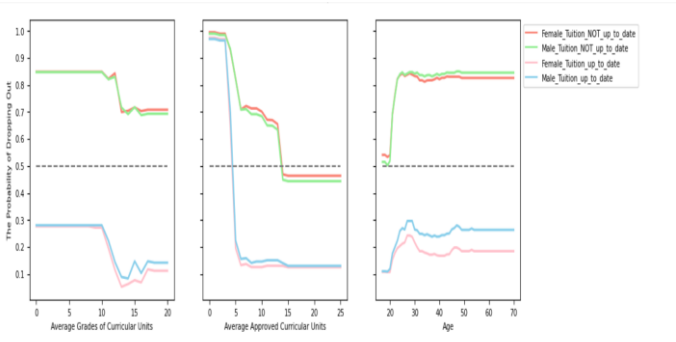

Fig -2: Correlation Heatmap Showing Feature Relationships with Dropout Likelihood

Figure 2 visualizes how the probability of dropping out varies with different feature values.

The following insights were derived from the plots:

- The probability of dropping out exceeds 0.8 for students

  with average grades below 10 and fewer than 5 approved curricular units, indicating high-risk cases.

- The dropout probability increases from 0.3 at age 17 to 0.6at age 50, suggesting that older students are at higher risk of dropping out.

3. CONCLUSIONS

RetentionX provides a scalable and effective solution for predicting student dropout risks and recommending personalized interventions. With XGBoost achieving 93.2% accuracy, the system outperformed other models and demonstrated significant real-world impact during the pilot study.The recommendation system, powered by an LLM and supported by Flask, combines predictive analytics with personalized interventions to enhance student success. Data insights from correlation heatmaps, model performance tables ensure its recommendations are both accurate and actionable, aligning with institutional goals for engagement and retention. Future work will focus on integrating real-time data and expanding the recommendation engine to include additional resources.

ACKNOWLEDGEMENT

We, the developers of the RetentionX system, express our heartfelt gratitude to everyone who contributed to this project an innovative machine learning-based mobile application designed to predict student dropout risks and deliver personalized interventions. This endeavor, aimed at tackling the pressing issue of high dropout rates in higher education, owes its success to the invaluable support of institutions, individuals, and cutting-edgetools. First and foremost, we extend our heartfelt thanks to our guide, Mrs.Mini Joswin whose invaluable guidance, expertise, and continuous support have been instrumental in shaping this project from its inception to completion. Their insightful feedback and encouragement have greatly enhanced our understanding and execution of the work, helping us navigate the challenges of this academic endeavor.

Additionally, we extend our appreciation to our project coordinators, whose exceptional organizational skills and coordination efforts ensured that we stayed on track and met our deadlines. Their administrative support and resourcefulness have played a pivotal role in the smooth progression of this project, allowing us to focus on our research and analysis.

REFERENCES

[1] Skittou, M., Merrouchi, M., & Gadi, T. (2024). “Development of an Early WarningSystem to Support Educational Planning Process by Identifying At-Risk Students”.IEEE Access, 12, 2259-2271.

[2] Mayanda Mega Santoni , Oenardi Lawanto, T. Basaruddin ,Kasiyah Junus ."Enhancing Student Engagement-Detection Using Bagging Ensemble Learning.“ - IEEE Access, 2024.

[3] Fister, C. (2023). ”A Comprehensive Review of isualization Methods for Association Rule Mining: Taxonomy, Challenges, Open Problems, and FutureIdeas.”120901.

[4] Wrapper Methods for Multi-Objective Feature Selection Series ISSN: 2367-2005 697 10.48786/edbt.2023.58

BIOGRAPHIES

Mrs. Mini Joswin is an Assistant Professor working in Dept of IT & AD, Amal Jyothi College of Engineering, Kanjirapally . Her deep expertise in Machine Learning and database-driven systems and application development enabled her to expertly mentor students through the critical areas of project.

Advaitha S Nair is a final-year B.Tech student in Information Technology at Amal Jyothi College of Engineering, Kerala, India. In the RetentionX project, she focused on model performance evaluation by implementing and analyzing KNN and Random Forest algorithms for dropout prediction. She also contributed to the design of the LLM-based recommendation system, system integration, and final testing. Her interests include data analytics, web development, database management, and machine learning.

Sereena Sebastian is a final year B.Tech student in Information Technology at Amal Jyothi College of Engineering. In the research work titled “Retention X: Dropout Prediction and Recommendation System,” she was primarily responsible for designing and implementing the mobile frontend interface using Flutter, ensuring

seamless user interaction and efficient data visualization. She also contributed to the machine learning component by training and fine-tuning the Artificial Neural Network (ANN) model, which was used to predict potential student dropouts based on academic performance and admission data

Sherin Joji is a final-year B.Tech student in Information Technology at Amal Jyothi College of Engineering, Kerala, India. In the RetentionX project, she worked on backend development, implemented SQL queries using PostgreSQL for data preprocessing, and designed logic for the dropout prediction pipeline. She also compared multiple machine learning models and optimized database performance. Her interests include SQL, object-oriented programming, backend development, and data-driven applications.

Sreedevi P S is a final-year B.Tech student in Information Technology at Amal Jyothi College of Engineering, Kerala, India. She advanced RetentionX by optimizing feature selection, model development, and integrating FAISS for its recommendation system, enhancing the overall student dropout prediction. Her diverse interests span machine learning, full-stack web development, database management, and scalable system design.

Sweety Daisy Thomas is a final-year B.Tech student in Information Technology at Amal Jyothi College of Engineering, Kerala, India. Her interests include machine learning, data analytics, and networking. She contributed to the Retention X: Dropout-Prediction and Recommendation System project by leading the testing phase, ensuring the application's reliability and user-friendliness. She also evaluated the performance of the XGBoost model to improve predictive accuracy to enhance student dropout prediction and recommendation system