Dropout Prediction with Supervised Learning

Asst. Prof. Shivani Awasthi[1], Rahul Biradar[2], Ritik Singh[3]

[1]*Asst. Professor, Lokmanya Tilak College of Engineering, Navi Mumbai, Maharashtra* [2]*Dept. of Computer, Lokmanya Tilak College of Engineering, Navi Mumbai, Maharashtra* [3]*Dept. of Computer, Lokmanya Tilak College of Engineering, Navi Mumbai, Maharashtra*

Abstract— Student dropout is a critical issue in the education sector, impacting institutional efficiency and student success. This project, Dropout Prediction with Supervised Learning, leverages machine learning models—Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbours (KNN), and Naïve Bayes (NB)—to predict student dropouts based on historical academic, demographic, and behavioural data.

The study involves data preprocessing, feature selection, and model evaluation to identify key factors influencing dropout rates. Supervised learning techniques are employed to classify students into "at-risk" and "not at-risk" categories. The performance of each model is assessed using accuracy, precision, recall, and F1-score metrics to determine the most effective predictor.

The findings aim to provide educational institutions with actionable insights, enabling early intervention strategies such as academic counselling and financial aid support. By implementing predictive analytics, institutions can enhance student retention and improve overall educational outcomes.

Keywords – Dropout Prediction, Supervised Learning, Machine Learning Models, Student Retention, Predictive Analytics, Classification Algorithms

I. INTRODUCTION

Predicting student behavior is critical for improving curriculum design and tailoring academic support interventions. By identifying patterns and trends in student performance, institutions can create personalized strategies to enhance learning outcomes and overall academic success. One key area of concern is the identification of at-risk students who are likely to drop out. Dropouts face significant societal challenges, including increased likelihood of antisocial behaviors and difficulties in the labor market, which underscores the importance of early intervention.

Machine learning techniques, such as predictive modeling, have emerged as powerful tools for enhancing student retention. These methods enable institutions to analyze large datasets, identify weaknesses in student performance, and provide timely support to those in need. This study builds on existing research by incorporating both supervised and unsupervised classification algorithms to predict student dropouts effectively.

To address class imbalances often found in dropout prediction datasets, resampling techniques like Synthetic Minority Oversampling Technique (SMOTE) are applied. This ensures that predictive models are trained on balanced data, improving their accuracy and reliability. Furthermore, the research investigates the effectiveness of various machine learning algorithms while identifying key factors that influence student success.

By leveraging advanced machine learning approaches, this study aims to provide actionable insights that can help educational institutions reduce dropout rates and improve student outcomes.

II. LITERATURE SURVEY

The prediction of student dropout has been a significant area of research, focusing on understanding the factors influencing dropout rates and developing machine learning models to address the issue. Below is a summary of relevant studies:

Nurhana RoslanJastini et al. (2024) analyzed student dropout in Malaysian private higher education institutions using clustering, association rules, and classification methods. Their study emphasized integrating diverse data sources and real-time data to enhance prediction accuracy [1].

Alice Villar and Carolina Robledo Velini de Andrade explored supervised machine learning algorithms for predicting student dropout and academic success. They compared Decision Trees, Random Forests, Support Vector Machines (SVM), and Naive Bayes models. The study highlighted the importance of comprehensive datasets and real-time intervention systems to improve model accuracy [2].

Nassim Mouchantaf and Maroun Chamoun investigated dropout prediction with minimal information using Decision Trees, Logistic Regression, and Neural Networks. Their research identified the need for robust models capable of effectively handling sparse or limited data inputs [3].

Shiful Islam Shohag and Masum Bakaul applied Logistic Regression, Neural Networks, and gradient-boosting techniques to detect student dropouts at universities. They pointed out the necessity of enhancing model generalizability across diverse datasets while addressing biases [4].

III. PROPOSEDSYSTEM

Student dropout is a significant challenge faced by educational institutions globally, with far-reaching social and economic consequences. To address this issue, the proposed system integrates machine learning algorithms to analyze diverse student data and predict dropout rates and academic success. By leveraging predictive modeling techniques, the system aims to identify at-risk students early in their academic journey, enabling institutions to implement timely interventions that enhance retention rates and improve educational outcomes.

The system is designed to utilize a comprehensive dataset comprising administrative records, behavioral data, and socio-economic factors. Administrative records include demographic information, enrollment history, attendance records, and academic performance metrics such as grades and course credits. Behavioral data captures engagement patterns, such as participation in extracurricular activities and interactions with e-learning platforms, while socio-economic factors provide insights into financial aid status and family background. These diverse data sources are integrated into the system to ensure a holistic analysis of student behaviour.

At the core of the system is a predictive modeling engine that employs several supervised machine learning algorithms, including Random Forests, Decision Trees, Logistic Regression, and Support Vector Machines (SVM). Each algorithm is trained on historical student data using labeled outcomes (e.g., dropout or graduate) to forecast future trends. To address class imbalances commonly found in dropout prediction datasets, resampling techniques like Synthetic Minority Oversampling Technique (SMOTE) are applied. This ensures that minority classes, such as dropouts, are adequately represented during model training, improving prediction accuracy.

The system incorporates an early warning mechanism that flags students at risk of dropping out based on predictive modeling results. By analyzing patterns such as low attendance rates, poor academic performance, or lack of engagement in extracurricular activities, the system identifies students who require immediate support. Educational institutions can use these insights to implement targeted interventions such as mentoring programs, financial aid assistance, or personalized tutoring.

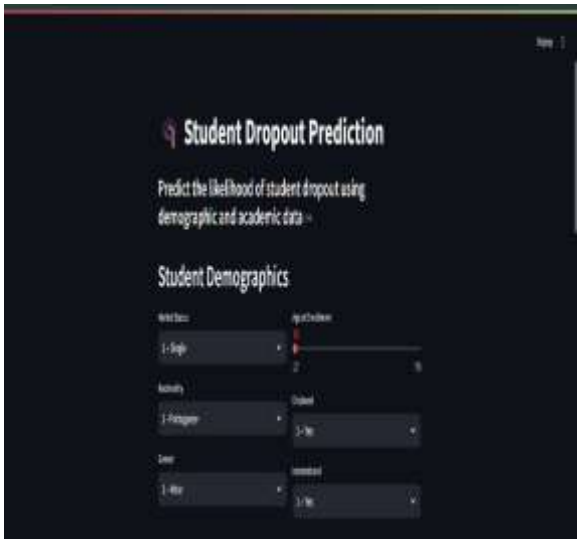
To ensure continuous improvement in prediction accuracy, the system integrates feedback mechanisms that refine models over time. Real-time updates allow the system to incorporate new student data as it becomes available (e.g., semester grades or attendance records), while outcome tracking validates predictions against actual results. This

feedback loop enables the system to adjust model parameters periodically, ensuring adaptability to changing trends in student behaviour.

The proposed system also evaluates the effectiveness of different machine learning algorithms using standard performance metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC). Comparative analysis helps identify the best-performing models for deployment in live environments. By providing actionable insights derived from robust data analysis, the system empowers educators to make informed decisions regarding curriculum design and resource allocation.

In conclusion, this proposed system offers a comprehensive solution for predicting student dropout rates and academic success. By integrating diverse datasets, employing advanced machine learning techniques, addressing class imbalances through resampling methods like SMOTE, and incorporating feedback mechanisms for continuous refinement, the system aims to reduce dropout rates and foster a supportive academic environment. This approach not only enhances student retention but also contributes to improved educational outcomes and institutional efficiency.

IV. PROPOSED SYSTEM INTERFACES



The interface is organized into structured sections for data input:

1. Student Demographics

This section collects essential demographic information required for prediction:

Marital Status: A dropdown menu allowing users to select options such as "Single."

Age at Enrollment: A slider ranging from 17 to 70 years, with the current value set at 18.

Nationality: A dropdown menu with predefined options, such as "Portuguese."

Displaced: A dropdown menu indicating whether the student is displaced (e.g., "Yes").

International: A dropdown menu specifying whether the student is an international student (e.g., "Yes").

Gender: A dropdown menu with options like "Male."

The interface serves as a front-end tool for implementing machine learning models that analyze demographic and academic data to predict student dropout probabilities. It enables users to input relevant features, which are processed by underlying supervised learning algorithms to generate predictions.

This interface exemplifies a user-friendly design tailored for research and practical applications in educational institutions, aiding in identifying at-risk students and implementing preventive measures.

The interface includes several sections for inputting relevant student data:

1. Family Background

Mother's Qualification: A dropdown menu allows selection of the mother's highest qualification, with an example entry "1 - Secondary Education - 12th Year of Sc...".

Mother's Occupation: A dropdown menu to specify the mother's occupation, with "0 - Student" as a selected example.

Father's Qualification: Similar to the mother's qualification, this dropdown menu allows selection of the father's highest qualification.

Father's Occupation: A dropdown menu to specify the father's occupation.

2. Academic Background

Previous Qualification: A dropdown menu to specify the student's previous qualification, with an example entry "1 - Secondary education."

Previous Qualification Grade: A slider to input the grade obtained in the previous qualification, currently set at approximately 150.00 out of 200.00.

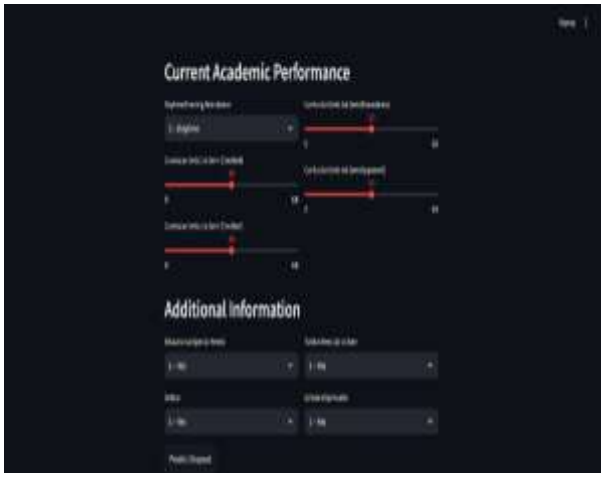
Application Mode: A dropdown menu to specify the mode of application, with an example entry "1 - 1st phase - general contingent."

Application Order: A numerical input field for specifying the application order, set to "0".

Admission Grade: A slider to input the admission grade, also set around 150.00 out of 200.00.

Course: A dropdown menu to select the course, with "33 - Biofuel Production Technologies" as the selected option.

It collects detailed family and academic background information to be used as features in a predictive model. The use of sliders for grade input allows for precise data entry, while dropdown menus ensure standardized input for categorical variables.



The interface is divided into two main sections: "Current Academic Performance" and "Additional Information."

1. Current Academic Performance

Daytime/Evening Attendance: A dropdown menu allowing users to specify whether the student attends classes during the day or evening, with "1 - Daytime" selected.

Curricular Units 1st Sem (Evaluations): A slider to input the number of curricular units evaluated in the first semester, currently set to 30 (with a range from 0 to 60).

Curricular Units 1st Sem (Credited): A slider to input the number of curricular units credited in the first semester, currently set to 30 (with a range from 0 to 60).

Curricular Units 1st Sem (Approved): A slider to input the number of curricular units approved in the first semester, currently set to 30 (with a range from 0 to 60).

Curricular Units 1st Sem (Enrolled): A slider to input the number of curricular units the student is enrolled in for the first semester, currently set to 30 (with a range from 0 to 60).

2. Additional Information

Educational Special Needs: A dropdown menu to indicate whether the student has any special educational needs, with "1 - Yes" selected.

Tuition Fees Up to Date: A dropdown menu to indicate whether the student's tuition fees are up to date, with "1 - Yes" selected.

Debtor: A dropdown menu to indicate whether the student is a debtor, with "1 - Yes" selected.

Scholarship Holder: A dropdown menu to indicate whether the student is a scholarship holder, with "1 - Yes" selected.

The interface gathers detailed academic and financial information about students. The use of dropdown menus and sliders allows for standardized and precise data entry, which is crucial for the accuracy of the predictive model.

V. FUTURE SCOPE

The future of dropout prediction research lies in advancing the generalizability of predictive models across diverse educational settings. Applying these models to datasets from various institutions and regions can help identify universal patterns while accounting for region-specific factors influencing dropout rates. Additionally, hybrid models combining supervised learning with unsupervised techniques can uncover hidden patterns in unlabeled data, further enriching predictions.

Incorporating advanced methods such as deep learning—especially recurrent neural networks (RNNs) or long short-term memory networks (LSTMs)—can improve predictions by capturing complex, time-dependent patterns in student behavior. Real-time academic data and behavioral insights, such as engagement levels and participation in school activities, can provide a more comprehensive understanding of dropout factors. Finally, integrating these predictive systems into practical educational tools and dashboards can enable timely interventions, improving retention rates and fostering better academic outcomes for at-risk students.

VI. CONCLUSION

The early prediction of student dropout is crucial for academic institutions aiming to provide timely interventions and improve students' success rates. This study employs a variety of machine learning techniques, including Decision Tree, Logistic Regression, Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest, to predict academic dropout among students. By training and testing these models on relevant datasets, the research aims to identify patterns and factors that

contribute to student attrition. The implementation of the proposed prediction approach enables course advisors, educational organizations, and universities to assess students' performance proactively and implement effective interventions tailored to individual needs.

In this investigation, the Random Forest model demonstrated superior performance compared to other algorithms in accurately predicting student dropouts. Its ability to handle large datasets and its robustness against overfitting make it particularly effective for this application. The findings suggest that utilizing machine learning techniques can significantly enhance the predictive capabilities of educational institutions, allowing them to allocate resources more effectively and support at-risk students before they disengage from their studies.

To further increase the accuracy of the proposed model, it is recommended that future research incorporate additional datasets, potentially sourced from academic Big Data repositories. This expansion would allow for a more comprehensive analysis of dropout predictors and could lead to improved model performance through the inclusion of diverse variables that influence student retention. By continuously refining the predictive models with new data, educational institutions can stay responsive to changing trends in student behavior and enhance their strategies for fostering academic success. Overall, this study underscores the importance of leveraging machine learning in educational settings to mitigate dropout rates and promote student achievement effectively.

VII. RESULTS

	Algorithm	Accuracy	Precision	Recall	F1
3	Random Forest	0.911801	0.917158	0.911801	0.908003
4	SVM	0.906832	0.913660	0.906832	0.902301
1	Logistic Regression	0.900621	0.901714	0.900621	0.897460
2	Decision Tree	0.891925	0.895645	0.891925	0.887073
5	Naive Bayes	0.884472	0.885503	0.884472	0.880113
0	K-Nearest Neighbors	0.880745	0.891486	0.880745	0.872636

The performance of six supervised machine learning algorithms for student dropout prediction was evaluated using multiple metrics including accuracy, precision, recall, and F1 score. The comparative results are presented in Table 1.

Random Forest demonstrated the highest overall performance with an accuracy of 0.911801, precision of 0.917158, recall of 0.911801, and F1 score of 0.908003. This indicates that Random Forest correctly identified approximately 91.18% of all instances while maintaining a balance between precision and recall.

Support Vector Machine (SVM) achieved the second-best performance with accuracy, precision, recall, and F1 score of 0.906832, 0.913660, 0.906832, and 0.902301 respectively. The high precision value suggests that SVM was particularly effective at minimizing false positives in the dropout prediction.

Logistic Regression performed adequately with an accuracy of 0.900621, precision of 0.901714, recall of 0.900621, and F1 score of 0.897460. While slightly lower than Random Forest and SVM, these results still indicate strong predictive capability.

Decision Tree achieved an accuracy of 0.891925, precision of 0.895645, recall of 0.891925, and F1 score of 0.887073.

Though lower than the top three models, Decision Tree still provided reasonable predictive power with the benefit of interpretability.

The remaining algorithms—Naive Bayes and K-Nearest Neighbors—showed the lowest performance among the tested methods. Naive Bayes achieved accuracy, precision, recall, and F1 scores of 0.884472, 0.885503, 0.884472, and 0.880113 respectively. K-Nearest Neighbors had the lowest overall performance with an accuracy of 0.880745, precision of 0.891486, recall of 0.880745, and F1 score of 0.872636.

VIII. Performance Comparison Analysis

1. The performance differential between the highest-performing algorithm (Random Forest) and the lowest-performing algorithm (K-Nearest Neighbours) was approximately 3.1 percentage points in accuracy and 3.5 percentage points in F1 score. This relatively narrow margin suggests that all algorithms performed reasonably well for the dropout prediction task, though Random Forest provides a clear advantage.
2. It is noteworthy that for all algorithms except K-Nearest Neighbours, the precision values were consistently higher than their respective recall values, indicating that these models were slightly better at avoiding false positives than false negatives in dropout prediction.
3. The consistent performance across multiple evaluation metrics reinforces the reliability of these results and suggests that ensemble methods like Random Forest may be particularly well-suited for dropout prediction tasks in educational contexts.

REFERENCES

1. RoslanJastini, N., JamilIzwan, M., Mohd Shaharanee, N., & Juma Sultan Alawi, **S.** (2024). Prediction of Student Dropout in Malaysian's Private Higher Education Institute using Data Mining Application. This paper explored the application of clustering, association rules, and classification methods to predict student dropout. The study highlighted the need for integrating more diverse data sources and real-time data.
2. Villar, A., & Robledo Velini de Andrade, C. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. This study compared various supervised machine learning algorithms, such as Decision Trees, Random Forest, Support Vector Machines, and Naive Bayes. The authors emphasized the need for more comprehensive datasets and real-time intervention systems to improve predictive model accuracy in educational contexts.
3. Mouchantaf, N., & Chamoun, M. (2023). Predicting Student Dropout with Minimal Information. This research investigated the use of Decision Trees, Logistic Regression, and Neural Networks for predicting student dropout when only minimal information is available. The study identified the need for enhancing model robustness and accuracy in scenarios with extremely limited or sparse data inputs.
4. Shohag, S.I., & Bakaul, M. (2022). A Machine Learning Approach to Detect Student Dropout at University. This paper presented a machine learning approach using Logistic Regression, Neural Networks, or Gradient Boosting to detect student dropout at a university. Further investigation is needed to enhance model generalizability across different datasets and address potential biases.
5. Babalola, Y. T., Adigun, J. A., & Akindele, M. O. (2023). Analysis of factors influencing students' dropout rate in tertiary institutions: A case study of selected universities in Nigeria. *Education and Development Economics*, *7*(1), 45–58.

Biographies

Author Asst. Prof. Shivani Awasthi

Co Author Rahul Biradar

Co Author Ritik Singh