A Random Forest–Based Framework for Multiclass Student Outcome Prediction and Probabilistic Dropout Risk Stratification

Shraddha Dattaguru Vedpathak[1], Anushree Sharad Ashtul[2]

[1]P.G. Student, Department of Computer Applications SVERI's College of Engineering Pandharpur, India.

[2]Assistant Professor, Department of Computer Applications, SVERI's College of Engineering Pandharpur, India.

E-mail: [1]shraddhadvedpathak@coep.sveri.ac.in,[2]asashtul@coe.sveri.ac.in

Abstract

Early identification of students at risk of discontinuing their studies requires integrated analysis of academic and student-related characteristics. This research introduces the concept of AI-Based Student Dropout Risk and Intervention System (ADSIRS) that is based on a multi-class classification of the outcome, risk stratification, intervention suggestion, and role-based surveillance in a single architecture. The pre-processed data about students are applied for Machine Learning (ML)-based classification, and the obtained classifications are further mapped to low, medium, and high-risk categories in order to prioritize the intervention process. Testing showed that the accuracy of ADSIRS was 92.0%, and its multiclass ROC-AUC value was 0.94. This system uses a Python-powered web application that incorporates MySQL to link prediction results with role-based dashboards and intervention alerts. This approach creates an effective computing workflow from the processing of student data and outcomes prediction to risk assessment and interventions, which can be used as a method for student retention management.

Keywords: Student Dropout Prediction, Risk Stratification, Intervention Recommendation, Multiclass Outcome Classification, Academic Risk Analytics, Role-Based Monitoring, Student Retention Management.

DOI: https://doi.org/10.36548/jucct.2026.4.008
Extended version of article received from **6th International Conference on Artificial Intelligence, 6G Communications and Network Technologies 2026**
Received: 10.09.2026, received in revised form: 07.10.2026, accepted: 19.10.2026, published: 03.11.2026
© 2026 Inventive Research Organization. This is an open access article under the Creative Commons Attribution-NonCommercial International (CC BY-NC 4.0) License

1. Introduction

Student retention has been a major concern for educational institutions due to the potential impact that discontinuation of education might have on the academic success of the students as well as the institution's performance. Early detection of students at risk of dropout is therefore very essential for academic assistance. Traditional methods of monitoring students usually focus on examination results, attendance, observation by the lecturers, and manual monitoring. The methods might be insufficient to detect complicated trends in the various attributes of the student.

Machine-Learning-based prediction provides an evidence-based methodology for examining the correlation between student traits and academic success. Performance, attendance, engagement, behaviors related to studying and many other student traits could be combined and analyzed to detect patterns that can be linked to various educational outcomes. Therefore, predictive analytics could help in identifying students that should be monitored more closely than the rest whose academic trajectories are stable enough. Nevertheless, a prediction model alone does not represent the full solution to the problem of student retention as risk needs to be categorized and linked to interventions.

To address this requirement, the development of an AI-Based Student Dropout Risk and Intervention System has been proposed. This system combines student data processing, prediction of the result using machine learning, risk categorization, and intervention recommendations into one. The system processes student data, determines the probability of an unfavourable outcome, classifies it in terms of low, medium, or high risk, and then generates intervention actions. The Python-based Web Application that uses MySQL database supports customized dashboards that can help in monitoring student data, predictions and risk factors. This way, the proposed system creates an end-to-end process from data collection and pre-processing to prediction, risk assessment, intervention and monitoring, creating a real foundation for data-driven support of students' retention.

2. Related Works

2.1 Machine-Learning-Based Student Dropout Prediction

The recent researches have indicated the use of machine learning approaches for predicting student dropout and attrition. ML and statistical approaches have been used for early university dropout prediction [1]. The ML approach has also been used for estimating student attrition risk [2]. Prediction of dropout during all the year has also been considered in order to consider the changes in student risk during the academic phases [4]. The various machine learning approaches have been compared for university dropout prediction [5] and ensemble classification approaches for early course dropout prediction have been studied [6].

2.2 Academic and Behavioural Risk Factors

The prediction of student dropouts has included more academic and behavioral data to achieve better identification of risks at an early stage. There were several studies examining the features of students through the academic period [4] and various characteristics of students for the purpose of attrition prediction [1], [2], [5]. Another approach that was studied is the application of digital footprints of students, which included the use of clustering techniques for learning activity data to improve early dropout prediction [9].

2.3 Explainable and Intervention-Oriented Prediction

Interpretable ML and intervention-focused techniques have been developed to predict students' dropouts. An interpretable machine learning algorithm has been designed to facilitate dropout prevention [3]. The application of explainable ML for identifying predictors for individual student's dropout predictions has also been conducted [7]. Furthermore, a local explainability framework has been developed that links freshman dropout prediction with prevention methods [8]. Machine-Learning algorithms have also been used for student risk identification [10].

2.4 Research Gap

Existing studies have already proven the suitability of machine learning algorithms for predicting student dropouts; however, these techniques have not been able to integrate themselves into the process of risk assessment, risk level classification, intervention, and monitoring. The proposed research will fill this gap as the system will consist of these processes.

3. Proposed Methodology

3.1 System Architecture and Overall Workflow

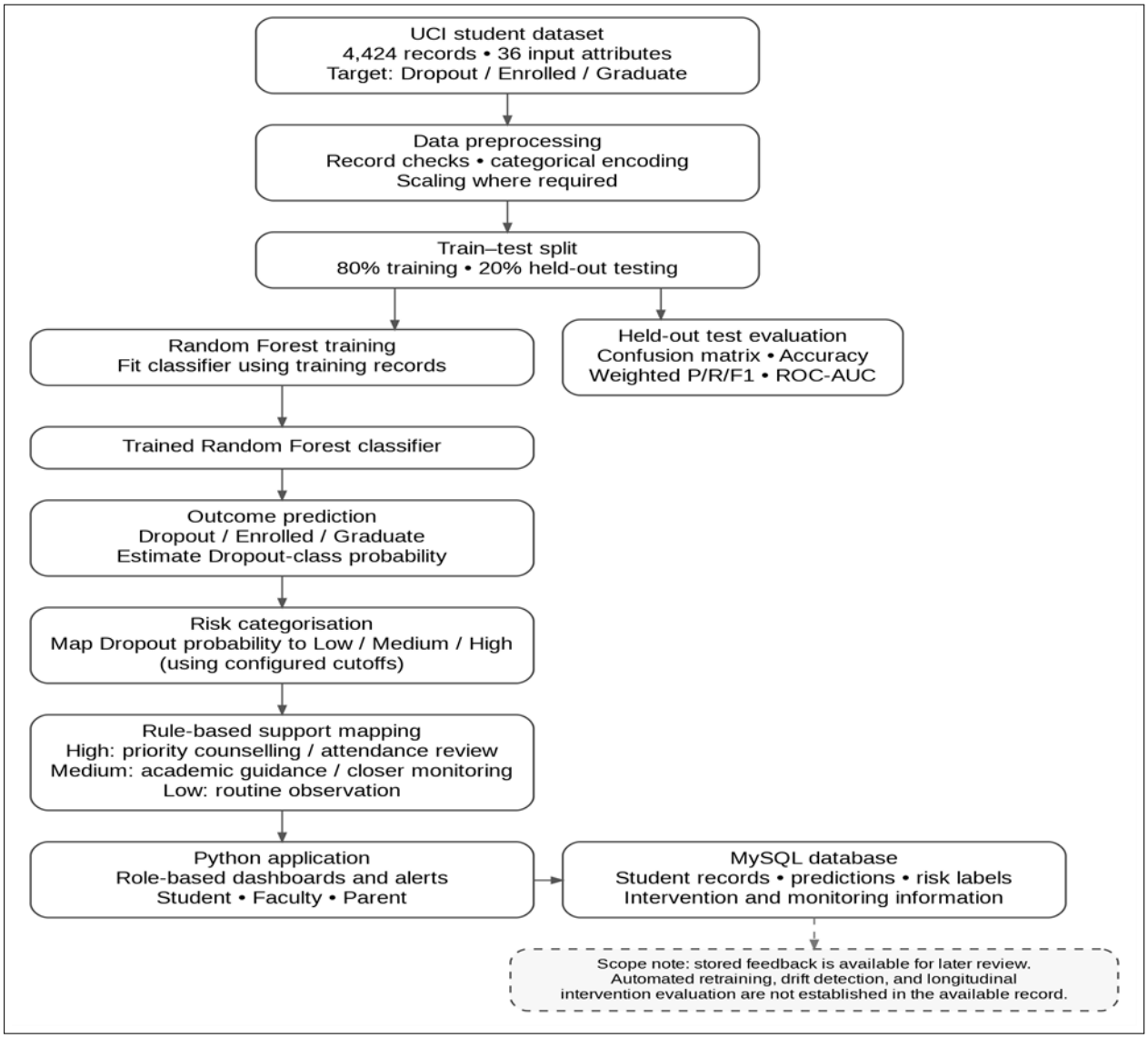


Figure 1. Random forest-based student outcome, risk, and support workflow.

The structured UCI student datasets are pre-processed and split into training and testing subsets where the former is employed in training the Random Forest classifier while the latter is used to derive the stated performance measures for classification. In the prediction phase, the Random Forest classifier provides the outcome in terms of the academic state and the probability of the Dropout class, which is converted to a risk category depending on the configured cut-off values and then mapped to a corresponding support measure. Prediction results, risk factors, and

warnings are shown to the users through the role-based dashboards by the Python application, whereas MySQL database holds the student data, prediction results, classification results, and intervention data. Feedback is stored for future reference; however, there is no evidence of retraining, drift detection, and longitudinal analysis of intervention effects in the implementation document. The overall architecture is presented in Figure 1.

3.2 Functional Modules

Table 1. Functional modules of the proposed system

Module	**Function**
Data Collection	Collects student information from academic records, attendance records, surveys, and institutional sources.
Data Pre-processing and Feature Analysis	Handles missing and inconsistent data, performs categorical encoding and normalization, and identifies relevant prediction attributes.
Machine Learning Model Development	Trains the selected machine-learning model using processed historical student records and observed outcomes.
Dropout Risk Prediction	Estimates individual dropout probability and classifies students into low-, medium-, or high-risk categories.
Intervention and Alert Management	Generates risk-based intervention recommendations and alerts for appropriate academic or support actions.
Monitoring and Feedback	Provides role-based dashboards and reports and retains updated information and intervention outcomes for later review.

The system architecture has been designed to comprise of six interlinked modules addressing the data gathering, pre-processing and feature selection, machine learning model generation, drop-out prediction, intervention and alerting, and monitoring with feedback. The sequence of the modules allows gathering and processing the student data and producing predictions that

further will be used for intervention generation and alerting. The monitoring module ensures the role-based access to the prediction and reporting results while storing up-to-date data and the results of interventions for future use in the model tuning. The functional modules are described in Table 1.

3.3 Dataset Description and Data Collection

Table 2. Input features available in the UCI dataset

Feature group	**Input attributes**
Demographic, enrolment, and socioeconomic [21]	Marital status, Application mode, Application order, Course, Daytime/evening attendance, Previous qualification, Previous qualification grade, Nationality, Mother's qualification, Father's qualification, Mother's occupation, Father's occupation, Admission grade, Displaced, Educational special needs, Debtor, Tuition fees up to date, Gender, Scholarship holder, Age at enrolment, International student status.
First-semester curricular attributes [6]	Credited curricular units, enrolled curricular units, Curricular-unit evaluations, Approved curricular units, Curricular-unit grade, Curricular units without evaluations.
Second-semester curricular attributes [6]	Credited curricular units, enrolled curricular units, Curricular-unit evaluations, Approved curricular units, Curricular-unit grade, Curricular units without evaluations.
Macroeconomic context [3]	Unemployment rate, Inflation rate, Gross domestic product (GDP).

The "Predict Students Dropout and Academic Success" data set collected from the UCI Machine Learning Repository was adopted for prediction of students' outcome [11]. It includes

4,424 data instances and 36 predictors ranging from demographic, socioeconomic, enrolment, semester curriculum and macroeconomic details. The target attribute includes three categories: Dropout, Enrolled, and Graduate. Consequently, the problem is considered to be a multi-class classification problem. An 80:20 training-testing split is cited in the literature for developing and testing the models. The target categories are different from the three categories of Low, Medium and High-Risk for intervention purposes.The dataset features are described in Table 1.

3.4 Data Pre-processing and Feature Selection

The input features were derived from the UCI Student Dropout and Academic Achievement dataset, which is made up of 36 features that include demographics, enrolment, socioeconomic status, academic performance on a per-semester basis, and macroeconomic features. Features under demographic and enrolment include age at enrolment, admission score, application method, course, prior qualification, tuition fee status, scholarship status, and other relevant student features. Academic features include credit units, enrolled units, evaluated units, approved units, grades, and units without evaluation in the first and second semesters. Macroeconomic features include unemployment rate, inflation rate, and GDP.

The categorical features were converted to numeric form for ease of modeling. The numeric features were left as numerics. Scaling was performed wherever needed according to the pre-processing algorithm. Min–max normalization can be expressed as

$$x_i' = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{1}$$

where x_i represents the original value of a feature, x_{min} and x_{max} denote its minimum and maximum values, respectively, and x_i' represents the normalized value. When $x_{\text{max}} = x_{\text{min}}$, the feature requires separate handling to avoid division by zero.

The preprocessed dataset was split into training and testing datasets in the proportion of 80:20. The training dataset was employed to build the random forest classifier whereas the testing dataset was used to validate its prediction accuracy. The dataset splitting can be expressed as

$$D = D_{\text{train}} \cup D_{\text{test}} \, , \; D_{\text{train}} \cap D_{\text{test}} = \emptyset \tag{2}$$

where D denotes the complete dataset, D_{train} represents the training subset, and D_{test} represents the testing subset.

These features were then input into the Random Forest classifier for predicting the three classes of student outcomes, which include Dropout, Enrolled, and Graduate. Through the use of demographics, socioeconomic, enrollment, and academic features, the classifier is able to learn the relationship between student features and educational outcomes.

3.5 Random Forest Classification

Random Forest was used as a classification algorithm to predict the performance of the students from their demographic, enrollment, socio-economic, and academic attributes. The model is built as an ensemble of decision trees created by bootstrapping samples from the training set along with selecting random features at each split point. The architecture of this ensemble allows building non-linear models.

For an input feature vector X, each decision tree produces a predicted class. The final prediction is determined by majority voting across the ensemble:

$$\hat{y} = mode\{h_1(X), h_2(X), \ldots, h_T(X)\} \tag{3}$$

where $h_t(X)$ denotes the prediction of the t-th decision tree, T represents the total number of trees, and $\hat{y}$ denotes the predicted student outcome.

During tree construction, candidate splits can be evaluated using the Gini impurity:

$$G(S) = 1 - \sum_{k=1}^{K} p_k^2 \tag{4}$$

where S represents a node, K denotes the number of outcome classes, and p_k is the proportion of samples belonging to class k. A split is preferred when it reduces the weighted impurity of the resulting child nodes.

3.6 Student Risk Classification

Classification of student risks is done through the probability of dropout that is obtained using the Random Forest classifier. For a student i, the dropout risk score is defined as

$$R_i = P\left(Y_i = \text{ Dropout } \mid X_i\right) \tag{5}$$

where X_i represents the student's input attributes and Y_i denotes the predicted academic outcome. The value of the risk score varies between 0 and 1. The higher the value, the greater the probability of dropout.

Based on two decision thresholds, τ_L and τ_H, students are classified into three risk categories:

$$C_i = \begin{cases} \text{Low Risk}, & R_i < \tau_L \\ \text{Medium Risk}, & \tau_L \leq R_i < \tau_H \\ \text{High Risk}, & R_i \geq \tau_H \end{cases} \tag{6}$$

where C_i denotes the assigned risk category, subject to $0 \leq \tau_L < \tau_H \leq 1$. This classification provides a structured basis for prioritising students for academic support.

3.7 Risk-Based Intervention Recommendation

The recommendations component employs a set of predetermined rules to link the particular risk level assigned to the students into the relevant type of support measure. High-risk students are selected for individual counseling, mentoring, and academic performance review, while medium-risk students are recommended for academic counseling and regular performance monitoring, while low-risk students will have their performance monitored through academic counseling.

However, the effectiveness of these recommendations needs to be measured independently of the prediction results achieved by the classifier. Indicators that can be considered include shifts in attendance, course completion, advancement, and dropout rates before and after the implementation of intervention measures. Comparison of students who have had some intervention and a suitable control group may reveal any effect on academic performance.

3.8 System Implementation and User Dashboards

The proposed system is developed as a Python-based web application linked to MySQL database that contains students' records, predictions results, risk categories, intervention details, and monitoring. The application brings together data processing and machine learning components with the user interface to access student data and risk assessment results by roles through dashboards.

System Home Page: The System Home Page that is presented in Figure 2 serves as the first point of access to the student dropout risk and intervention system.Student Dashboard: As illustrated in Figure 3, the Student Dashboard displays attendance, average marks, subject performance, and notifications so that students can keep track of their academic progress.Faculty Dashboard: The Faculty Dashboard as illustrated in Figure 4 is used for gathering data on students' academic performances and monitoring.The Parent Dashboard is displayed in Figure 5 and contains information such as students' information, attendance, average marks, and risk level.

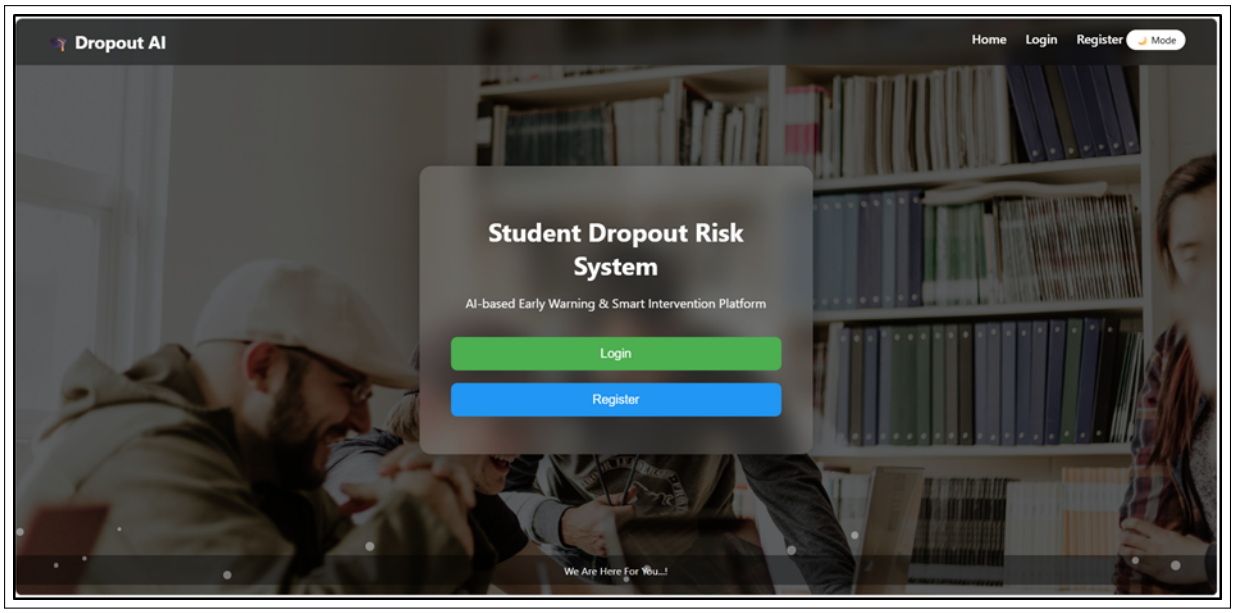


Figure 2. Student home page

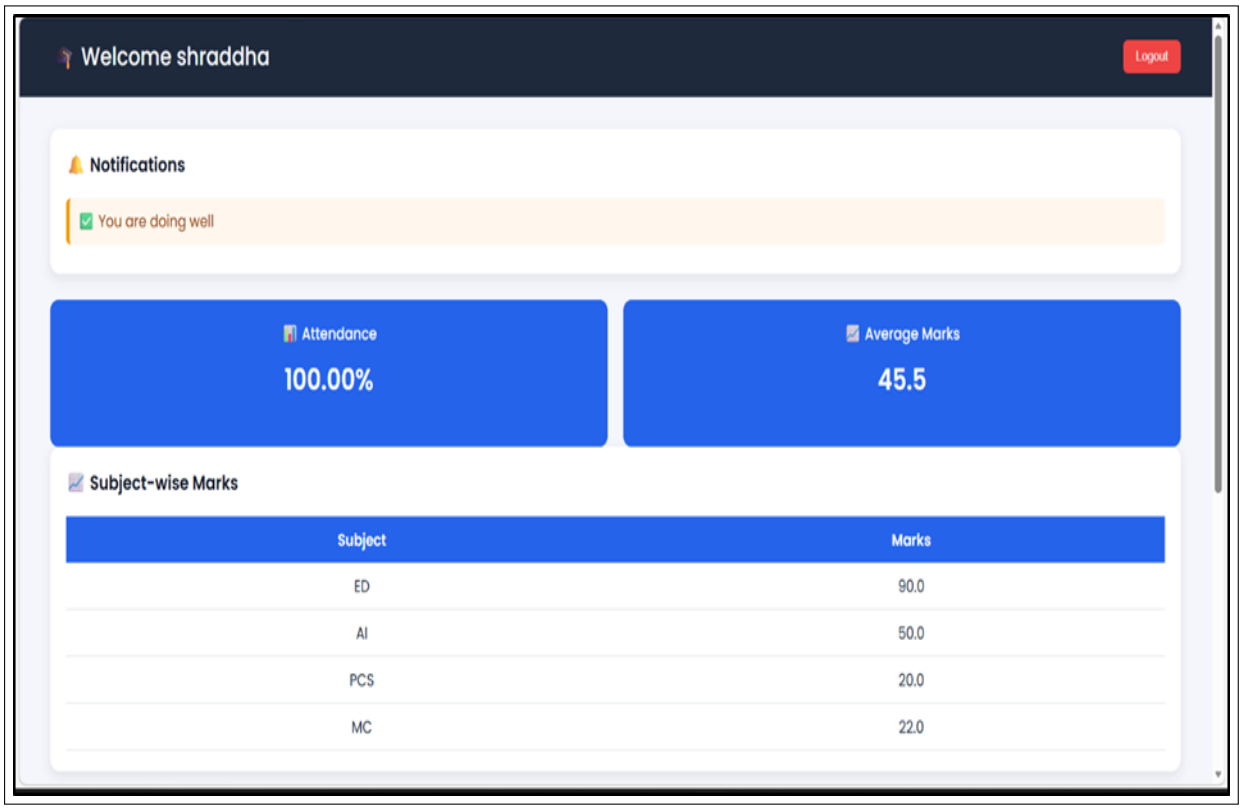


Figure 3. Student dashboard

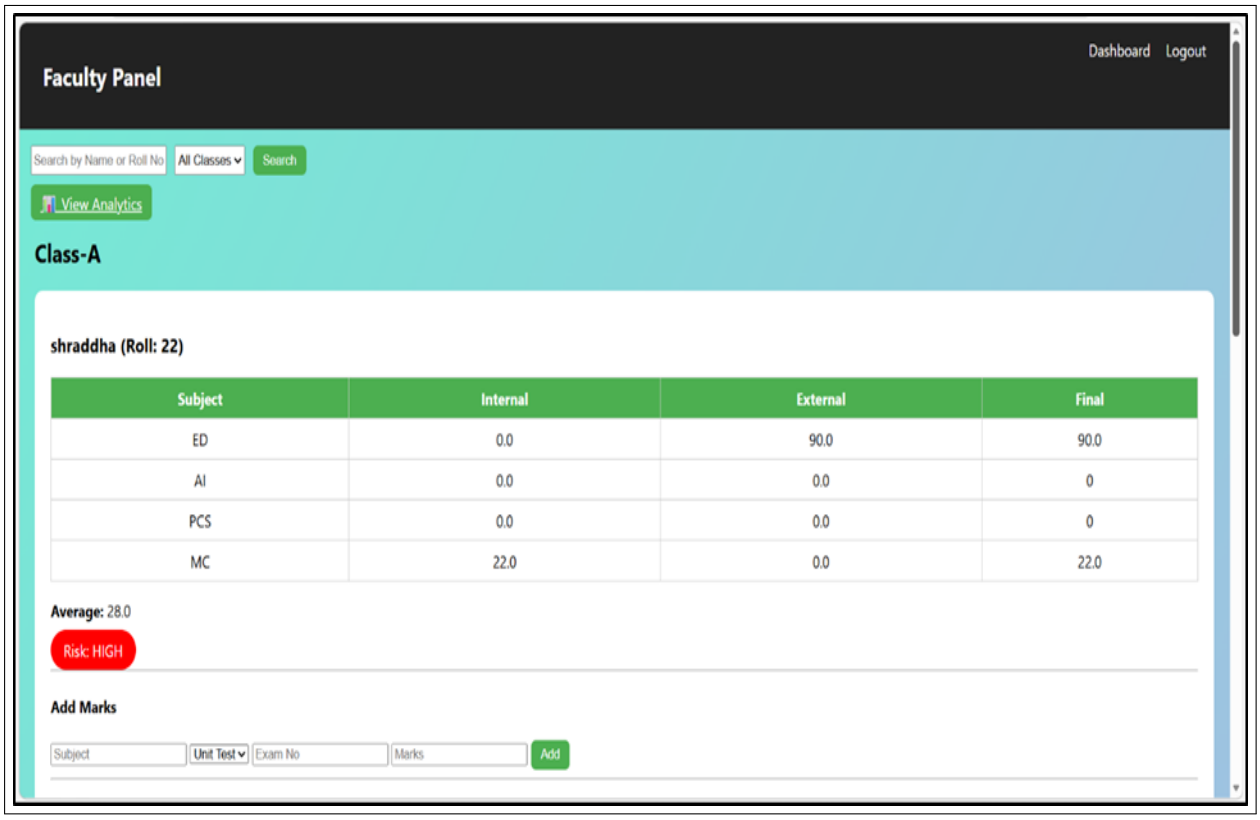


Figure 4. Faculty dashboard

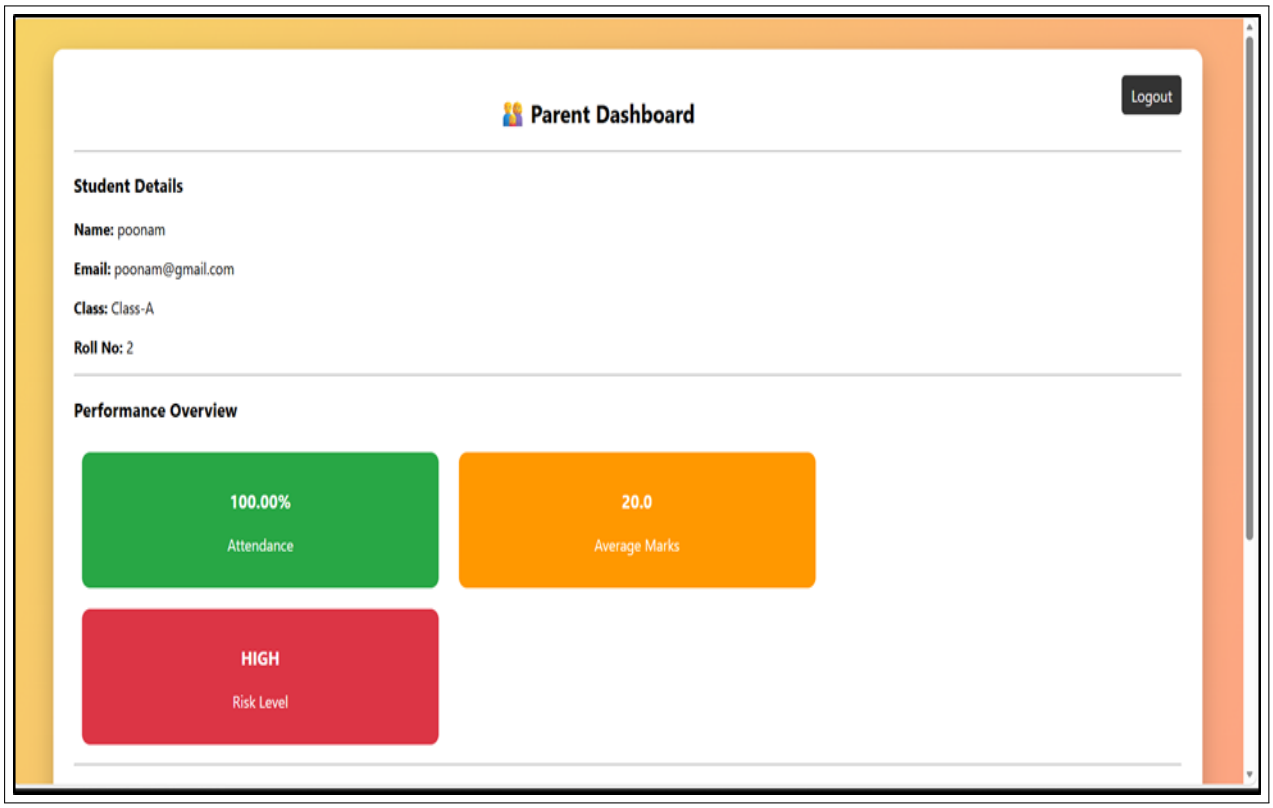


Figure 5. Parent dashboard

3.9 Evaluation and Feedback Mechanism

The performance of the Random Forest classifier was evaluated by means of the held-out test data set to measure how effectively it can classify the academic outcomes of students. The performance measures were measured in terms of accuracy, precision, recall, F1-score, and ROC analysis for all the three classes: Dropout, Enrolled, and Graduate. The confusion matrix was used to examine the distribution of correct and incorrect predictions for each class.

Classification accuracy was calculated as

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Precision and recall were calculated as

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{9}$$

The F1-score was obtained from the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

where TP, TN, FP, $and\ FN$ denote true positives, true negatives, false positives, and false negatives, respectively, for the class under consideration. For the three-class classification task, the reported precision, recall, and F1-score were calculated using weighted averaging across the outcome classes.

ROC analysis was performed using a one-versus-rest strategy to assess the discrimination capability for each outcome class. The area under the ROC curve (AUC) summarises the model's ability to distinguish a given class from the remaining classes. The reported macro-average AUC represents the arithmetic mean of the three class-wise AUC values.

4. Results and Discussion

4.1 Predictive Performance

Table 3. Predictive performance of the proposed model

Metric	**Performance**
Accuracy	92.0%
Weighted Precision	91.8%
Weighted Recall	92.0%
Weighted F1-score	91.8%
ROC-AUC	0.94

The proposed system for predicting student dropout exhibited a uniform classification performance on all the metrics under consideration. According to the figures given in Table 2, the accuracy is 92.0%. The weighted precision, recall, and F1-scores are useful metrics when

it comes to evaluating the prediction performance of the three classes of student outcomes, considering the distribution in the data. It is evident that the measures show similar values, indicating a good prediction performance on all the classes.

4.2 Confusion Matrix Analysis

Figure 6 illustrates class-wise prediction of the suggested model using the confusion matrix. This matrix comprises three result classes which are; Dropout, Enrolled, and Graduate. It can be noted from the matrix that most predictions lie on the principal diagonal, meaning that the model accurately predicts the majority of samples belonging to each of the classes. The highest number of correct predictions occurs in the Graduate class followed by Dropout and Enrolled classes respectively. There is low misclassification since some Dropout and Enrolled samples are predicted to belong to other result classes.

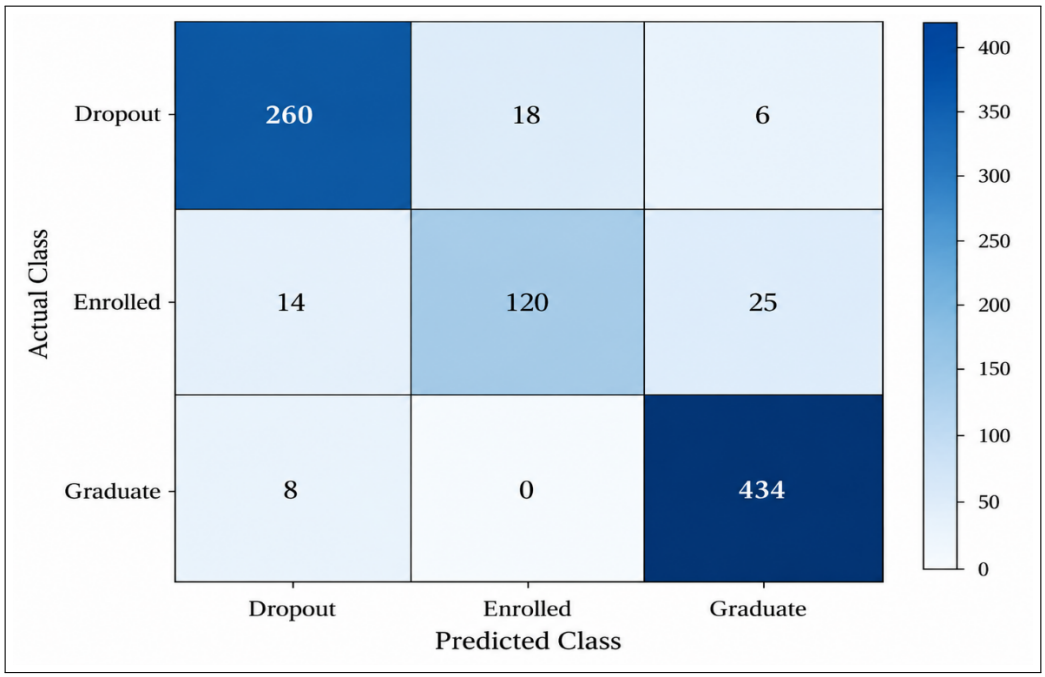


Figure 6. Confusion matrix of the proposed dropout prediction model

4.3 ROC-AUC Analysis

The multiclass ROC plot shown in Figure 7 is used to examine the discriminative power of the model under discussion. It should be noted that the plots of the outcomes Dropout, Enrolled, and Graduate significantly lie above the random classifier line, which indicates successful discrimination of outcome classes at various classification cut-offs. The values of class-wise AUC are as follows: 0.94 for Dropout, 0.91 for Enrolled, and 0.96 for Graduate, where the

highest discriminative power is observed for the Graduate class.

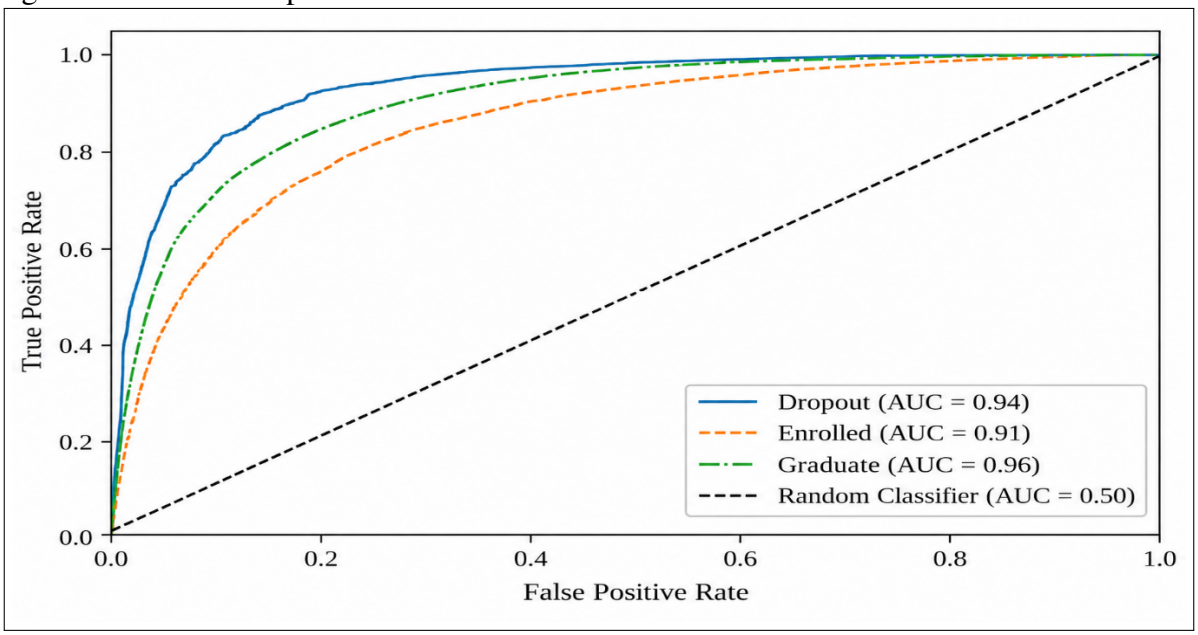


Figure 7. Multiclass ROC curve of the proposed dropout prediction model

4.4 Risk Classification and Intervention Analysis

The three academic results that have been predicted, Dropout, Enrolled, and Graduate, are then used to generate the risk level associated with dropouts for intervention purposes. The probability of the Dropout result is classified into Low, Medium, and High Risk by applying the decision levels defined. The higher risk levels students are put under close academic supervision and counseling, while those at medium risk levels are monitored regarding their academic performance and attendance. Those that fall within the low risk level continue to be under normal academic supervision.

4.5 System and Dashboard Analysis

The proposed system offers role-based interfaces for the purpose of obtaining student performance and risk of dropout information. The Home Page, as illustrated in Figure 2, acts as the primary interface from which users can access the system for login and registration. The Student Dashboard, as illustrated in Figure 3, provides the user with information such as attendance, average marks, performance by each subject, and notifications. This interface allows students to check on their academic standing. The Faculty Dashboard, as illustrated in Figure 4, allows access to student performance and risk information for the purposes of academic monitoring. The Parent Dashboard, as illustrated in Figure 5, offers student

information, attendance, average marks, and associated risk information.

5. Conclusion and Future Work

The proposed AI-Based Student Dropout Risk and Intervention Management System offers a comprehensive approach to detecting outcome patterns among students and offering effective academic interventions in time. In this context, the system represents the combination of student data preprocessing, outcome prediction through the use of machine learning techniques, risk stratification, intervention recommendation, alerting mechanism, and role-based dashboards. The integration of such elements makes it possible to use predicted outcomes as a basis for generating actionable risks for students, teachers, and parents. In addition, this framework ensures the connection between predictive analytics and intervention management. Future work will focus on incorporating of longitudinal data of students, other behavioral markers, dynamic interventions, explainable machine learning models, and real-time tracking. The evaluation of the proposed system with more extensive institutional data will further validate its effectiveness and scalability.

References

[1] Cannistrà, Marta, Chiara Masci, Francesca Ieva, Tommaso Agasisti, and Anna Maria Paganoni. ”Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques.” Studies in Higher Education, 2022, vol. 47, no. 9: 1935-1956.

[2] Barramuño, Mauricio, Claudia Meza-Narváez, and Germán Gálvez-García. ”Prediction of student attrition risk using machine learning.” Journal of Applied Research in Higher Education, 2022, vol. 14, no. 3: 974-986.

[3] Katsuragi, Miki, and Kenji Tanaka. ”Dropout prediction by interpretable machine learning model towards preventing student dropout.” In Transdisciplinarity and the Future of Engineering: Proceedings of the 29th International Society of Transdisciplinary Engineering (ISTE) Global Conference, 2022, 678-683.

[4] Song, Zihan, Sang-Ha Sung, Do-Myung Park, and Byung-Kwon Park. ”All-year dropout prediction modeling and analysis for university students.” Applied Sciences, 2023, vol. 13,

no. 2: 1143.

[5] Cho, Choong Hee, Yang Woo Yu, and Hyeon Gyu Kim. ”A study on dropout prediction for university students using machine learning.” Applied Sciences, 2023, vol. 13, no. 21: 12004.

[6] Pecuchova, Janka, and Martin Drlik. ”Predicting students at risk of early dropping out from course using ensemble classification methods.” Procedia Computer Science, 2023, vol. 225: 3223-3232.

[7] Krueger, Joao Gabriel Correa, Alceu de Souza Britto Jr, and Jean Paul Barddal. ”An explainable machine learning approach for student dropout prediction.” Expert Systems with Applications, 2023, vol. 233: 120933.

[8] Delen, Dursun, Behrooz Davazdahemami, and Elham Rasouli Dezfouli. ”Predicting and mitigating freshmen student attrition: A local-explainable machine learning framework.” Information Systems Frontiers, 2024, vol. 26, no. 2: 641-662.

[9] Pecuchova, Janka, and Martin Drlik. ”Enhancing the early student dropout prediction model through clustering analysis of students' digital traces.” IEEE Access, 2024, vol. 12: 159336-159367.

[10] Kuntintara, Wichukorn, Piya Warabuntaweesuk, and Sountaree Rattapasakorn. ”Student dropout prediction using machine learning.” In 9th International Conference on Business and Industrial Research (ICBIR), IEEE, 2024, 0229-0233.

[11] V. Realinho, M. V. Martins, J. Machado, and L. Baptista, “Predict Students' Dropout and Academic Success,” UCI Machine Learning Repository, 2021.
Available: https://archive.ics.uci.edu/dataset/697/predict%2Bstudents%2Bdropout%2Band%2Bacademic%2Bsuccess