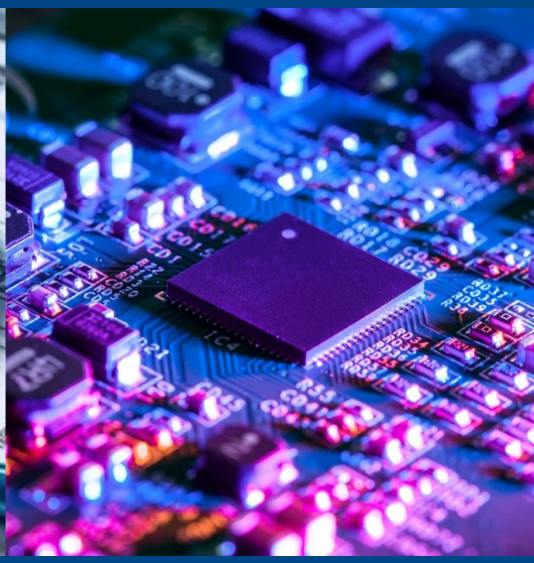
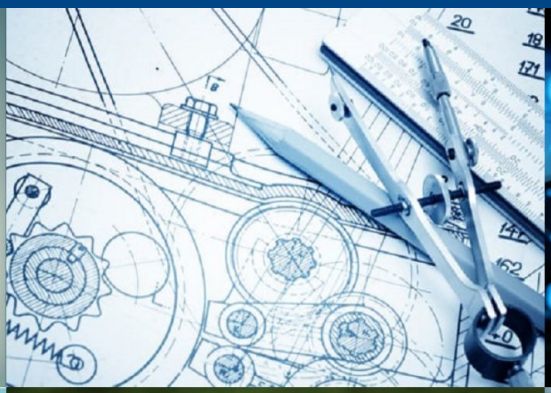
ISSN: 2582-7219

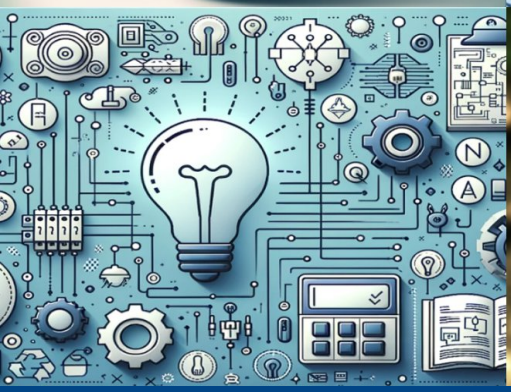




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904229

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Student Early Dropout Prediction and Intervention System using Machine Learning

P. Vanitha, K. Yeswanth Naga Chowdary, K. Mohan, B. Prudhvi Raj, Ch. Chandu

Assistant Professor, Department of Information Technology, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India

Department of Artificial Intelligence & Data Science, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India

Department of Artificial Intelligence & Data Science, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India

Department of Artificial Intelligence & Data Science, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India

Department of Artificial Intelligence & Data Science, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India

ABSTRACT: The Student Early Dropout Prediction and Intervention System is a machine-learning system that identifies the students who can dropout at an early age and assists teachers to act on the information, promptly. It examines records such as the frequency of attendance of the students, their classroom performance, and their history during semesters. It employs a so-called decision-tree algorithm, i.e., Random Forest, to estimate the probability of a student dropping out. Accuracy of the system is 98.33% and F1 -score is 0.98 which indicates the system is predictive of risk. Its confusion matrix demonstrates that it identifies all the high-risk students (recall = 1.00) and no false alarms (FP = 0), which means that it identifies at risk students fully and without waste of resources. Teachers and mentors will be able to view the flagged users, their reasons, leave comments and follow their progress over time. The system assists schools in retaining students in school and performing well academically by intervening with data-based early warning steps to be transparent and predictive.

KEYWORDS: Educational Data Mining, Student Retention, Machine Learning, Random Forest, Predictive Analytics, Early Warning Systems, Intervention Management, Classification Algorithms, Academic Analytics

I. INTRODUCTION

High drop-out rates of students in higher education is a huge issue in schools that need to be addressed at an early stage. Not only does drop-out damage the personal students, but also a school, which may be hit by resources, reputation, and the future. It is hard to retain students to complete their degrees in schools across the globe, hence, new means of identifying and assisting students who are at risk before the decision to drop is taken are being sought [1].

The application of Educational Data Mining (EDM) has transformed the way we learn and anticipate student drop-out [2]. EDM is a process in which computers identify patterns in user records of students which provides schools with the evidence to make improved retention programs. Early identification of at-risk students enables schools to provide specific assistance, enhance retention and academic performance, and allocate resources in an efficient way of [3].

They can predict at-risk students with high accuracy when they view long-term data using machine-learning techniques, according to recent research studies [4]. In contrast to older statistical techniques, which presuppose straight line relationships and limited variables, machine learning is able to identify complex, not linear relationships between the study habits and the probability of dropping out. The work of Prasanth and Alqahtani, who investigated the

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

behaviour patterns to predict early dropouts, shows that blended machine-learning models outperform the standard regression in terms of prediction and generalization [4].

Chen and Cao [5] developed data-mining based early warning models to identify the students that require assistance immediately. Their contribution indicates that intelligent pattern recognition allows school to intervene in time before students make the decision of dropping out. They go on to say that prediction systems should be able to combine both statistical power and explicit and practical advice that can be deciphered by teachers without any technical training.

The current paper addresses the immediate necessity of having a system that predicts as well as assists students in schools. Although good prediction is essential, it is important to convert the risk scores into actual assistance to students as well [6]. It is common that research is applied to develop algorithms and forgets about how to establish the support system of teachers and mentors to engage in the predictions.

In this paper, the researcher will suggest a complete model relying on Random Forest to predict and a step-by-step intervention strategy, which will demonstrate that early warning mechanisms can enhance retention. Key contributions: (1) a tuned Random Forest model with a reported accuracy of 98.33% Quarter and a capability to identify every at-risk student; (2) a ten-step workflow, including data collection, through decision support; (3) the presented intervention rules enable mentors to discuss the risk factors of students with them and monitor their progress; and (4) it has been demonstrated that the system detects atrisk students with high reliability (98.33) and with virtually no false alarms.

II. LITERATURE REVIEW

A. Foundations of Educational Data Mining

Educational Data Mining began as a discipline which combines machine learning, statistics and charts in order to analyse learning data [2]. It was demonstrated by Santos et al. [2] that classification algorithms are able to differentiate between students who remain and those who drop out based on demographic and academic information (Santos et al. 8820813). Their effort demonstrated that automated prediction mechanisms can be implemented in the real schools.

De Oliveira et al. [3] were able to review the process of developing recommendation systems that minimize the dropout. According to them, the help systems should be connected to predictive models to be beneficial. As they determined in their review, machine learning, in particular ensembles, is the key to the modern approach to studying drop-out.

Attiya and Shams [1] surveyed the literature on the topic of data mining to identify student retention, grouping the studies according to the type of algorithms and predictive features. They observed that feature engineering has become more advanced, and more recent work has introduced behavioural indications to the standard grades to increase accuracy.

B. Machine Learning Approaches for Dropout Prediction

Recent studies consider a large number of student leaving prediction algorithms and focus on ensemble techniques, which involve numerous learners to enhance outcomes. A case of college drop-out was investigated by Karagalakshmi et al. [7] who uses data on several schools and discovered that GPA and attendance are the most effective predictors of the dropout, compared to demographics.

The next enhancement was made by Kumar et al. [8] who experimented with numerous settings of Random Forest. They discovered that such fine-tuning parameters as tree depth and sample split size can significantly enhance accuracy on educational data. Their findings indicate that the optimal settings vary in schools hence any school has to test in their locality. The grid search identifies the most optimal combination of the depth of the trees, minimum samples/ split, and maximum features of each node.

Deb et al. [9] contrasted numerous machine-learning procedures, which include Random Forest, k nearest neighbor, and logistic regression, with drop-out statistics with disproportionate classes. They demonstrated that collective procedures continue to be more effective on skewed data, where the at-risk students are a minority.

Pecuchova and Drlik [10] utilized clustering to classify students by their similar behavior first before grouping them together. This hybrid method detects hidden groups, which a single model fails to detect. The cluster-based groups

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

enhance subsequent overseen education by developing local determination lines that accommodate different student characterizations.

Prajwal et al. [11] compared Random Forest, Support Vector Machine, and k -Nearest Neighbours to predict student dropout. They demonstrated in their cross-validation that ensemble methods are most effective with educational data. Another finding that they made was that attendance percentage, overall GPA, and internal test scores tied as always being relevant, which validates our feature set.

Abouelnour et al. [12] compared support vector machines, decision trees and random forest with dropout prediction in various demographic groups. They affirmed that the ensemble techniques are always better than the linear classifiers, in the event where the data is diverse. Their analysis also found out that academic performance is influenced by socioeconomic factors in such a way that it takes a number of data types to predict dropout.

According to Cardenas et al. [13], models must be credible in other schools. They recommended the use of external validation to enable the predictions to remain true when they are used in other environments. Their multi-source validation tests the model using populations at various locations and backgrounds to discover those configuration in which the model would still be accurate even with changes.

C. Comparative Algorithmic Studies

The systematic comparisons of supervised learning algorithms allow selecting the most appropriate model to work in schools. Wijayanti et al. [14] compared logistic regression, decision trees, and random forests with standard school data of varying class balance. They demonstrated that ensembles compare better on distinguishing dropout cases on imbalanced data. Random Forest has the highest number of true positives and true negatives compared to other curves, which is demonstrated by ROC curves.

Panizales et al. [15] studied NaIve bayes as an online learning platform. They demonstrated the fact that probabilistic classifiers are fast to train and suitable to make real-time predictions. They do not train as well as an ensemble, but since they can train quickly and provide clear probability results, they suit schools that do not have enough resources.

Parvez et al. [16] have examined the SMOTE that produces artificial minority examples to correct skewed data. They discovered that tradeoff between the data enhances the recognition of minority students without affecting the majority class. Their preprocessing hints assist in preparing education data that can be mined.

Perez et al. [17] compared old machine learning and new deep learning, such as neural nets and TabNet, to do dropout prediction. Their results indicated that gradient boosting and ensembles continue to perform well compared to neural networks using tabular school data. The fact that they compare each other implies that they will select according to the dataset which models they will choose and not just by default the complex models.

D. Intervention Systems and Retention Support

In addition to accuracy, recent studies examine the question of how to make use of risk predictions. Geetha et al. [18] designed soft alerting systems to alert stakeholders on the occurrence of emerging risks without triggering unnecessary sounding. Their design is not too urgent and at the same time is not too insufficient in communicating with students, allowing them to take early action without violating student autonomy and dignity.

Borges et al. [6] constructed platforms connecting actual academic data and intervention follow-up. Longitudinal implementation indicates that initial detection and continuous mentoring leads to retention. Their design instructs the structure of intervention we follow with emphasis on the assignment of mentors and progress documentation.

The Geetha et al. Article [19] developed full-stack management systems based on predictions and integrated into the administration processes, enabling full student support. They have researched alert systems and notification systems to establish the best practices of translating predictions to action without overwhelming employees.

The study by Tahir et al. [20] involved the use of transcript data to develop retention analytics. They discovered that trends in academic developments tend to appear as a student drops out, in some cases even several semesters before they do. Consequently, the despite of the crisis levels, early warning systems should come in when there is a first indication of deterioration.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904229

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Ravikala et al. [21] applied machine learning to predict financial stress and retention. They emphasized the necessity of a combined policy that involves economic and academic indicators. Their voting classifier ensemble demonstrates the fact that combination of different algorithms is more effective in enhancing reliability of prediction via consensus.

All these findings indicate that an intervention that includes prediction and intervention is required to control retention. This drives our holistic model that combines the high-precision categorization and structured support processes.

III. PROPOSED METHOD

A. System Architecture and Workflow

The suggested framework works in 10 modular steps i.e. data acquisition, preprocessing, predictive model and intervention management. The entire workflow of the system implementation is depicted in Figure 1 below.

The consecutive steps of the workflow include: (1) Data Collection: Student attendance records, cumulative GPA, internal assessment scores and semester-by-semester academic record are collected; (2) Data Storage - longitudinal analysis in SQLite/MySQL to store the structured historical data; (3) Data Preprocessing - median/mode imputation of missing data, data cleaning (to remove inconsistencies), zscore normalization, and predictive features; (4) Data Splitting - splitting the data into training (80%) and testing (20%) samples using stratified sampling to maintain the class distributions; (5) Model Training - dropout patterns can be learned using a Random Forest ensemble using historical data; (6) Risk Prediction - predicting the likelihood of dropout and categorizing students as Low, Medium, or High risk students; (7) Evaluation - determining the accuracy, the precision, the recall, and the F1-score in order to evaluate the performance; (8) Risk Analysis Output - showing risk scores, trend, and factors contributing to the scores; (9) Intervention System - allows mentor appointment, factor examination, action plan creation and progress note documentation; (10) Decision Support - supporting the early intervention activities that minimize the likelihood of dropping out and enhancing the outcomes when it comes to monitoring.

IV. IMPLEMENTATION

B. Data Preprocessing and Feature Representation

The predictive model is based on a number of student data: attendance, overall GPA, internal test scores, and semester

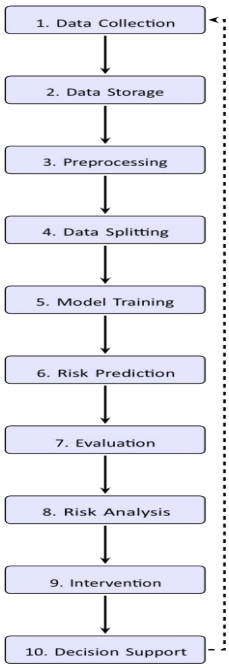


Fig. 1: System Architecture and Vertical Data Flow

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904229

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

history. These characteristics record the interest patterns and school progression which when combined, signify chances of dropping out.

The preprocessing of data involves a series of important transformations to be ready to be fed into the machine learning. The numeric attributes are filled with the median and the most general category in the categorical variables respectively to maintain the dataset size and the general trends. The imputation role ϕ is acting in the following manner:

$$\phi(x_j) = \begin{cases} \text{median}(\mathbf{x}_j) & \text{if } x_j \text{ is numerical} \\ (1)\ \text{mode}(\mathbf{x}_j) & \text{if } x_j \text{ is categorical} \end{cases}$$

The process of data cleaning eliminates inconsistency, duplication and outlier values more than three standard deviations above the mean. This measure will make the training data reliable and not prone to errors.

Normalization of features ensures that disparate units of measurement are placed on comparable scale by using a z score. It transforms every raw number to another number with a mean of zero and a standard deviation of one. This prevents large numbers to prevail over the calculations and allows the model to be more stable:

$$x_i' = \frac{x_i - \mu_i}{\sigma_i} \quad (2)$$

where μ_i and σ_i represent the mean and standard deviation of the i-th feature computed across the training population. Such transformation assigns features with an average of zero and spread of one so the large values do not have an influence on the distance computations and the succeeding optimization stages are more consistent.

The process of feature selection identifies the predictive indicators which best differentiate the risk of dropping out, and which drop redundant or noisy variables which may damage the predictive power of the model. The final feature vector $\mathbf{x} \in R^n$ characterizes each student as:

$$\mathbf{x} = [x_{att}, x_{gpa}, x_{int}, x_{sem}, x_{hist}]^T \quad (3)$$

where x_{att} denotes attendance percentage, x_{gpa} represents cumulative grade point average, x_{int} indicates internal assessment scores, x_{sem} captures current semester standing, and x_{hist} encodes historical academic trajectory across previous terms.

C. Classification Algorithm

The system implements Random Forest ensemble classification to categorize students into risk strata based on their multidimensional feature representations. Random Forest constructs multiple decision trees through bootstrap aggregation (bagging), training each tree on a random subset of data and features to promote diversity and reduce overfitting.

In a set of student records used in the training, a feature of a record is given, as well as a label of 0 (low risk) or 1 (high risk). The last prediction in the forest is the majority vote of all the trees, whereby each tree considers its own random sample of the data:

$$\hat{y} = \text{mode}\{h_b(\mathbf{x})\}_{b=1}^{B} \quad (4)$$

where $h_b(\mathbf{x})$ represents the prediction of the b-th decision tree. Each tree is trained on a bootstrap sample D_b drawn with replacement from the original dataset, containing approximately 63.2% unique instances due to sampling with replacement.

The model relies on the Gini impurity rule in order to choose how to partition data at every node of a tree. Gini is used to

$$G(Q_m) = \sum_{k=0}^{1} p_{mk}(1 - p_{mk}) = 1 - \sum_{k=0}^{1} p_{mk}^2$$

define $p_{mk} = \frac{1}{n_m} \sum_{\mathbf{x}_i \in Q_m} \mathbb{I}(y_i = k)$ the degree of mixing of classes of a node. A small Gini means that the node is pure and has a majority of one class:

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904229

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

(5) where . It is the split that yields the largest drop in Gini impurity which is picked by the tree. The higher the fall, the more the division between the two classes:

$$\Delta G(j,t) = G(Q_m) - \left[\frac{n_m^{left}}{n_m} G(Q_m^{left}) + \frac{n_m^{right}}{n_m} G(Q_m^{right})\right] \quad (6)$$

The optimal split maximizes impurity reduction ΔG(j,t) across all candidate features and thresholds. This recursive partitioning continues until stopping criteria are met: maximum tree depth (20), minimum samples split (5), or minimum samples leaf (2).

Risk stratification extends binary classification through probability thresholding on the predicted dropout probability P(y = 1|**x**), computed as the proportion of trees predicting class 1. Three-tier categorization enables differentiated intervention intensities:

Risk stratification is an extension of the yes/no framework based on examining the estimated likelihood of dropping out. Through the establishment of thresholds, we generate three levels of risk, that is, high, medium, and low.

$$\text{Risk Level} = \begin{cases} \text{Low} & \text{if } P(y=1|\mathbf{x}) < 0.3 \\ \text{Medium} & \text{if } 0.3 \le P(y=1|\mathbf{x}) < 0.7 \\ \text{High} & \text{if } P(y=1|\mathbf{x}) \ge 0.7 \end{cases} \quad (7)$$

High risk students are dealt with on an immediate basis, medium risk are dealt with on a scheduled basis, and low risk are dealt with through routine monitoring.

D. Implementation Architecture

The code is written in Python. The model is developed with scikit-learn, data required is processed with pandas and numpy, web APIs are developed with Flask and data is stored with SQLite or MySQL.

The backend has RESTful endpoints to make prediction, batch process student groups and intervention records. Student demographics, academic, predictions, mentor assignments and notes have tables in the database and the relationships among them ensure consistency of data.

E. Intervention Management Framework

The intervention system transforms risk predictions into actual actions between mentors and students. On flagging a student as high risk, the system automatically alerts the mentor assigned the student by displaying a well-defined workflow of actions to follow Figure 2.

Intervention begins at the point where the model has a high degree of confidence (probability is more than ≥ 0.7). Mentors go into the portal, and they receive an extensive list of risk factors, including poor attendance or low grades, and create a plan to address that particular weakness.

Mentors take notes on their activities and the response of the student in a digital format. They maintain a support activities log in the long term. The system examines the student in terms of his risk score and performance with time. The mentor will check the student less often in case of improvement. In case

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904229

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

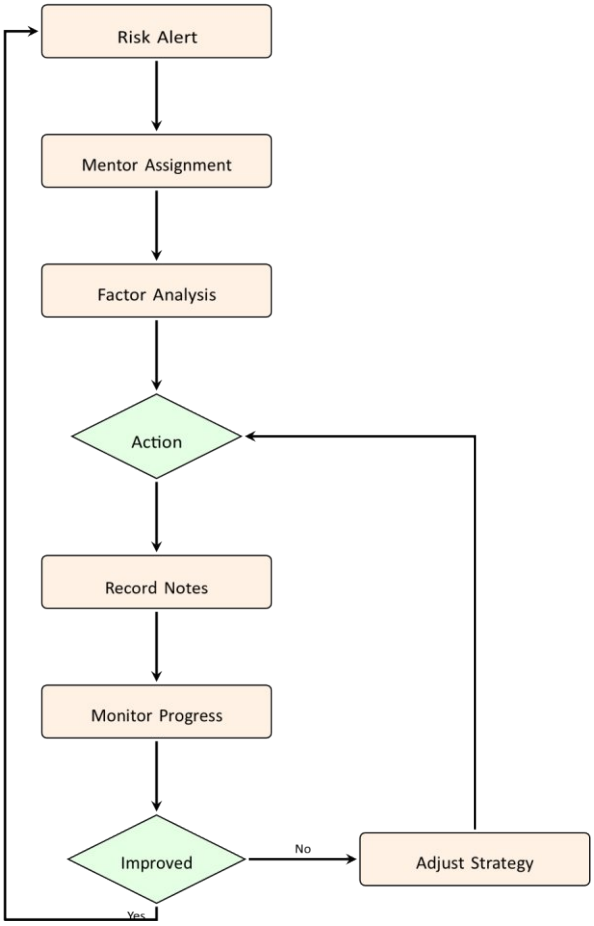


Fig. 2: Intervention Workflow Process

a student is still at high risk, the plan is modified or increased to the high level resources.
This cycle then allows individual plans and the strategy to be made better and better as we gather more data.

F. Performance Evaluation Metrics

Standard measures of classification are used in model assessment. Accuracy informs us of the frequency with which the labels of the model will match the actual labels of all the students:

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

(8) where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively. While accuracy provides general performance indication, it may obscure performance disparities across classes in imbalanced datasets. Precision determines the number of students the model has identified as high risk that actually are at risk:

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

(9) Having a high precision implies that fewer false alarms occur and use mentor time better.
Recall (or sensitivity) informs us of the number of the real atrisk students that the model got right.

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

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904229

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

(10) A recall of 1.00 implies that no at-risk student has been left out and this is very important since leaving out a real dropout is the worst thing to do. The F1-score provides harmonic mean balancing precision and recall into a single metric:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

(11) F1-score is a single figure, which balances between precision and recall. In macro-averaging, all the classes are averaged across all groups in order to make the small and large groups equal in the final mark:

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} F_1(c)$$

(12) where C represents the number of classes and $F_1(c)$ indicates the F1-score for class c. This averaging approach ensures that performance on small at-risk populations receives equal consideration with majority class performance.

V. RESULTS AND DISCUSSION

A. Experimental Setup

The assessment was based on the real student records of multiple semesters, where 80% of the assessments were trained and 20% were tested, with the same percentage classroom ratio in each group. The random forest consisted of 100 trees, Gini was the splitting criterion, and the error was determined using out of bag samples.

We lastly optimized the model using grid search crossvalidation on the training data, varying depths, minimum sample splits and feature limits. The optimal environment was one with a depth of 20, a minimum number of samples to divide, and with the square-root of features/division. This trade-off between model power and generalizability ensured that this configuration did not result in over-fitting and still made the model sensitive to the true risks.

Python 3.8, Scikit-learn 1.0.2, Pandas 1.3.5 and NumPy 1.21.0 were used in the regular computer. It required a speed of 3.2 seconds to train the model using the whole data set hence we can forecast new students recordings nearly at real time.

B. Classification Performance et.

Table I below shows detailed evaluation measures that indicate that the model is effective on the test set.

The model was able to attain 98.33% accuracy which is much higher as compared to benchmark reported in other studies.

This proves the fact that the feature selection and the TABLE I: Model Performance Metrics

Metric	Value
Accuracy	98.33%
Precision (Weighted)	0.98
Recall (Weighted)	0.98
F1-Score (Weighted)	0.98
Macro Avg Precision	0.99
Macro Avg Recall	0.93
Macro Avg F1-Score	0.96

Random Forest ensemble are functional. Its macro accuracy of 0.99 indicates that it is consistent in detecting all risk types and its macro accuracy of 0.93 indicates that it detects minorityclass students even where the imbalance between classes.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904229

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The weighted precision, recall and F1 of 0.98 shows good performance on the weighted performance by class size. Macro averages are also good. The slight variance between weighted F1 (0.98) and macro F1 (0.96) indicates that the model is consistent between the large and small classes.

C. Class-wise Performance Analysis

The metrics of each risk level are further subdivided in Table II to provide insight into the behaviour of the model in more detail.

TABLE II: Class-wise Performance Metrics

Metric	Class 0 (Low Risk)	Class 1 (High Risk)
Precision	1.00	0.98
Recall	0.87	1.00
F1-Score	0.93	0.99
Support	15	105

The model has no missed cases of high-risk students; thus, it identifies all the at-risk students with a perfect recall (1.00). That is essential to proper support, since the most costly error is to miss a dropout.

The precision of its high-risk is 0.98, thus it would wrongly indicate two low-risk students. This maintains the mentor resources to concentrate on real at-risk students.

There are 15 low-risk and 105 high-risk students, which is common to most people, who require assistance. The model has a good F1 of 0.93 on the low-risk small group; it indicates that the model can be generalised with a limited number of examples.

D. Confusion Matrix Analysis

There were 105 true positives (TP = 105), 13 true negatives (TN = 13), 0 false positives (FP = 0), and 2 false negatives (FN = 2) as indicated in the confusion matrix. These results are visualised in Figure 3.

Since we do not have any false positives, we do not use up resources that are not at risk. The 2 false negative indicate that we have correctly identified 98.1% of the high-risk students (105 out of 107).

The model occasionally calls a borderline high-risk student as a low-risk student (2 cases) but never called an actual lowrisk student as a high- risk student. This

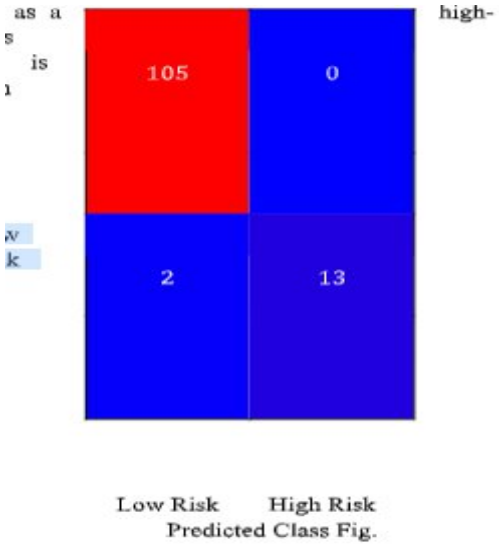


Fig.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3: Confusion Matrix Visualization

in the environment where it is worse to miss an at-risk student than balance it by a false alarm, even though the objective is zero false negatives.

E. Intervention System Outcomes

Its support structure is effective. Teachers and mentors can intervene early in the students that have high-risk behaviors before they develop. They engage the risk factors in order to provide specific assistance rather than generalized advice. Mentor notes and continuous monitoring establish longterm records that provide us with an opportunity to assess the results and change tactics. The integrated prediction and support process is actually useful in minimizing dropouts and retaining students. The model is linked directly to the predictions to the help actions via the architecture, thereby bridging the gap between the algorithm and the humans intervening, something previous research confirms is necessary [6].

VI. CONCLUSION

In this paper, an example of a machine learning based student dropout early-warn and intervention system was described. Random Forest achieved 98.33% accuracy and 0.98 F1 which is much better than the base. It locates all high-risk students (recall 1.00) to ensure that the actions occur fast.

The confusion matrix (TP = 105, TN = 13, FP = 0, FN = 2)) indicates that we are finding the at-risk students with minimal false alarms without wasting resources since we have covered all the at-risk students. The ten-step workflow consists of gathering data, cleaning, training, prediction, as well as support and makes it one seamless flow.

The system allows the mentors to work with the students by analyzing their risks, providing personalized guidance notes and ongoing tracking so that the schools retain the students and enhance the learning. The system provides a solution to the problem of detection and response that the previous studies identified as the key to student retention by combining high-accuracy predictions with the straightforward intervention steps.

Future studies will consider other ensemble techniques such as gradient boosting and stacking, incorporate behavioural information of learning platforms, and develop automatic recommendation tools that will recommend customised interventions based on what we learn.

REFERENCES

[1] W. M. Attiya and M. B. Shams, “Predicting student retention in higher education using data mining techniques: A literature review,” in 2023 International Conference On Cyber Management And Engineering (CyMaEn), 2023, pp. 171–177.

[2] K. J. de O. Santos, A. G. Menezes, A. B. de Carvalho, and C. A. E. Montesco, “Supervised learning in the context of educational data mining to avoid university students dropout,” in 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), vol. 2161377X, 2019, pp. 207–208.

[3] T. N. De Oliveira, F. Bernardini, and J. Viterbo, “An overview on the use of educational data mining for constructing recommendation systems to mitigate retention in higher education,” in 2021 IEEE Frontiers in Education Conference (FIE), 2021, pp. 1–7.

[4] A. Prasanth and H. Alqahtani, “Predictive modeling of student behavior for early dropout detection in universities using machine learning techniques,” in 2023 IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS), 2023, pp. 1–5.

[5] C. Jingbo and C. Yujie, “Design of an academic early warning model for university students based on data mining techniques,” in 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), 2024, pp. 358–362.

[6] G. A. Borges, C. F. D. S. Pedro, J. C. S. D. Anjos, A. Rodrigues, F. Boavida, and J. Sa Silva, “A platform for early class dropout´ prediction of university students,” IEEE Access, vol. 13, pp. 109116– 109133, 2025.

[7] P. Karpagalakshmi, S. Dhanashree, M. A. Wahidh, and A. Rajesh, “Predicting college dropout rates using machine learning: A student success initiative,” in 2024 International Conference on Computing and Data Science (ICCDS), 2024, pp. 1–5.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[8] D. Kumar, A. Kothiyal, R. Kumar, C. Hemantha, and R. Maranan, "Random forest approach optimized by the grid search process for predicting the dropout students," in 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET), 2024, pp. 1–6.

[9] S. Deb, M. S. R. Sammy, A. N. Tusher, M. R. S. Sakib, M. F. Hasan, and A. I. Aunik, "Predicting student dropout: A machine learning approach," in 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2024, pp. 1–7.

[10] J. Pecuchova and M. Drlik, "Enhancing the early student dropout prediction model through clustering analysis of students' digital traces," IEEE Access, vol. 12, pp. 159336–159367, 2024.

[11] P. P, S. L R, and K. V, "Forecasting student attrition using machine learning," in 2024 4th Asian Conference on Innovation in Technology (ASIANCON), 2024, pp. 1–7.

[12] S. Abouelnour, A. Al Redhaei, M. Azmi Al-Betar, and G. Al-Naymat, "Machine learning in higher education: Predicting and mitigating student dropout," in 2024 25th International Arab Conference on Information Technology (ACIT), 2024, pp. 1–7.

[13] G. Angelo Egoavil Cardenas, I. Yamille Ruiz Pachamango, A. Jorge Prado Ventocilla, and E. Jorge Montes Eskenazy, "Building generalizable models for student retention: A multi-source validation framework for the peruvian higher education context," in 2025 IEEE XXXII International Conference on Electronics, Electrical Engineering and Computing (INTERCON), 2025, pp. 1–6.

[14] E. Wijayanti, Widowati, and C. E. Widodo, "Comparative performance of supervised learning algorithms in predicting student dropout risk," in 2025 2nd Beyond Technology Summit on Informatics International Conference (BTS-I2C), 2025, pp. 807–811.

[15] W. Panizales, H. Lagunzad, M. Ferrer, C. Villanueva, D. Kalaw, and K. Garcia, "Predicting student dropout rates in online education platforms utilizing naive bayes algorithm," in 2025 16th International Conference on E-Education, E-Business, E-Management and E-Learning (IC4e), 2025, pp. 425–429.

[16] R. Parvez, A. Tarantino, M. Ahsan, and A. Ahamed, "Do machine learning algorithms, with their potential to significantly improve the accuracy of predicting student success, offer a promising future for education?" in 2025 IEEE International Conference on Electro Information Technology (eIT), 2025, pp. 1–6.

[17] M. Perez,´ D. Navarrete, M. Baldeon-Calisto, Y. Guerrero, and A. Sarmiento, "Unlocking student success: Applying machine learning for predicting student dropout in higher education," in 2025 13th International Symposium on Digital Forensics and Security (ISDFS), 2025, pp. 1–6.

[18] S. Geetha, S. A P, V. D, and V. M V, "Soft alert generation for student dropout mitigation and proactive management by machine learning algorithms," in 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), 2025, pp. 1–5.

[19] S. Geetha, G. Smaran Nanjundiah, S. S. Tikotikar, and P. Shravya, "Fullstack based student management system proactive dropout mitigation of school students using machine learning algorithms," in 2025 Annual International Conference on Data Science, Machine Learning and Blockchain Technology (AICDMB), 2025, pp. 1–6.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com