International Journal of Innovative Research & Growth
E-ISSN: 2455-1848
homepage: http://www.ijirg.com/ijirg/
Volume: 15 Issue: 4, October 2026



Machine Learning-Based Prediction of College Students' Dropout Risk: A Case Study of Maihar District, Madhya Pradesh, India

Shivendra Kumar Dwivedi
Assistant Professor, Department of Computer Science, Sharda Mahavidyalaya, Sarlanagar Maihar-485772, M.P., India

CORRESPONDING AUTHOR

Shivendra Kumar Dwivedi
e-mail: shivendra.mphil@gmail.com

KEYWORDS

Student Dropout, Machine Learning, Educational Data Mining, Random Forest, Predictive Analytics, Logistic Regression, Student Retention.

ARTICLE DETAILS

Received 05 July 2026; revised 25 September 2026; accepted 02 October 2026

DOI: 10.26671/IJIRG.2026.4.15.323

CITATION
Dwivedi, S. K. (2026). Machine Learning-Based Prediction of College Students' Dropout Risk: A Case Study of Maihar District, Madhya Pradesh, India. *Int J Innovat Res Growth*, 15(4), 154090-154099. DOI

This work may be used under the terms of the Creative Commons License.

Abstract

Student dropout in higher education remains a significant challenge, particularly in rural and semi-urban regions where academic, financial, and institutional factors influence student retention. This study investigates the determinants of undergraduate student dropout in Maihar District, Madhya Pradesh, and develops a machine learning-based framework for early dropout prediction. Primary data were collected from six undergraduate colleges (three government and three private) through structured questionnaires and institutional records.

The dataset comprises 315 student instances with variables including regular absence from college, academic performance, family income, parental education, transportation accessibility, digital connectivity, psychological well-being, lack of educational facility, without adult protection and institutional support. After data preprocessing and feature selection, three supervised machine learning algorithms-Logistic Regression, Decision Tree (J48), and Random Forest- were implemented using the WEKA data mining tool. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, ROC area, and confusion matrix. Among the models, Random Forest achieved the highest classification accuracy of 93.02%, followed by J48 (88.57%) and Logistic Regression (84.80%).

The results indicate that regular absence, academic performance, family income, digital connectivity, and weak institutional support are the most influential predictors of student dropout. The proposed framework serves as an effective early warning system, enabling timely interventions to improve student retention and support sustainable higher education development.

1. Introduction

Higher education plays a critical role in socio-economic development by enhancing human capital, by equipping students with the knowledge and skills needed for employment and lifelong learning. Although access to higher education has expanded in Madhya Pradesh through the establishment of colleges in rural and semi-urban areas, student dropout remains a major concern. Leaving college before completing a degree not only affects students' future career prospects but also reduces institutional effectiveness and limits the impact of educational investment.

Student dropout is influenced by multiple interrelated factors, including poor academic performance, irregular attendance, financial constraints, inadequate parental education, limited transportation, insufficient digital connectivity, and weak institutional support. These challenges are particularly evident in rural regions, where many students come from economically disadvantaged families and are often first-generation learners.

However, high student dropout rates remain a persistent structural issue, resulting in an inefficient misallocation of govt. and private educational resources. However, student dropout remains a major barrier to achieving universal educational outcomes.

Despite increasing enrollment in colleges and universities, student dropout remains a major challenge for higher education institutions worldwide. Students who leave their studies before completing their degree programs often face limited employment opportunities and reduced earning potential, while institutions experience lower graduation rates and inefficient use of educational resources. Globally, dropout rates in higher education vary between 20% and 40%, with significantly higher rates in developing countries.

Student dropout is influenced by multiple factors, including regular absence from college, poor academic performance, family income, parental education, transportation accessibility, digital connectivity, psychological well-being, lack of educational facility, without adult protection and institutional support.

Traditional methods of identifying at-risk students are often reactive and may fail to provide timely intervention. Recent advancements in machine learning and educational data mining have created opportunities to predict dropout risk by analyzing large volumes of student data and identifying patterns associated with student attrition.

In Maihar District, Madhya Pradesh, higher education is provided through Government colleges and Private colleges serving students from diverse socio-economic backgrounds. This study investigates the determinants of student dropout and develops a machine learning-based predictive framework for identifying at-risk students in undergraduate colleges of Maihar District, Madhya Pradesh.

1.1 Research Objectives

1. To identify these factors influencing dropout college students' in Maihar district.
2. To develop predictive machine learning model for dropout risk assessment.

2. Literature Review

The literature has used a variety of machine learning algorithms, according to the results of the survey. Most of these algorithms, on the other hand, were created and tested in industrialized countries; as a result, machine learning has been applied in few studies to solve this problem in developing countries. Dropout rates have been investigated from several perspectives. Several machine learning techniques have been used and implemented in several research articles and a few of them have been researched and described in the point below as literature survey-

Ayesha et al. [2] performed study on student learning behavior, for these factors like class quizzes, mid and final exams, assignments are studied. This study will help the tutors to reduce the ratio of dropout and improve the performance level of students.

Bhardwaj and Pal [3] used the decision tree method for classification to evaluate the performance of students. The objective of their study is to discover knowledge that describes students' performance in end semester examination. This study was quite useful for identifying the dropout students at an earlier stage and students who need special attention and allowing the teacher to provide appropriate advising.

Pandey and Pal [4] conducted a study on the student performance by selecting 60 students from a degree college of Dr. R.M.L. Awadh University, Faizabad, India by means of association rule, they found the interestingness of students in opting for class teaching language.

Mercy Paul Selvan et al. [5]. In this paper, in the realm of education, machine learning approaches are very important. Here, they've projected a victimization model that may be able to forecast dropout and provide a response to prediction using decision tree (DT) algorithm. The accuracy observed in this model is 97.67%. Many factors were considered for the cause of dropout, such as family history, money, and social activities, which added to the predictability of the status.

Breiman (2002) [9] introduced random forest as an ensemble learning method that improves predictive accuracy by aggregating multiple decision trees. It is widely used due to its robustness against overfitting and ability to handle nonlinear relationships.

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potential for educational institutes. Data mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the

learning process of students as described by Alaa el-Halees. Student dropout has been extensively studied in educational research. According to Tinto (1993), student retention depends on academic and social integration.

3. Research Methodology

This study adopted a quantitative research methodology to develop a machine learning-based framework for predicting student dropout in undergraduate colleges of Maihar District, Madhya Pradesh. The methodology follows the knowledge discovery in database process, which involves data collection, preprocessing, feature selection, model development and performance evaluation.

Three supervised machine learning algorithms-Logistic Regression, Decision Tree (J48), and Random Forest-were implemented using the WEKA data mining software. A 10-fold cross-validation technique is employed to validate the predictive performance of each model and minimize bias caused by data partitioning.

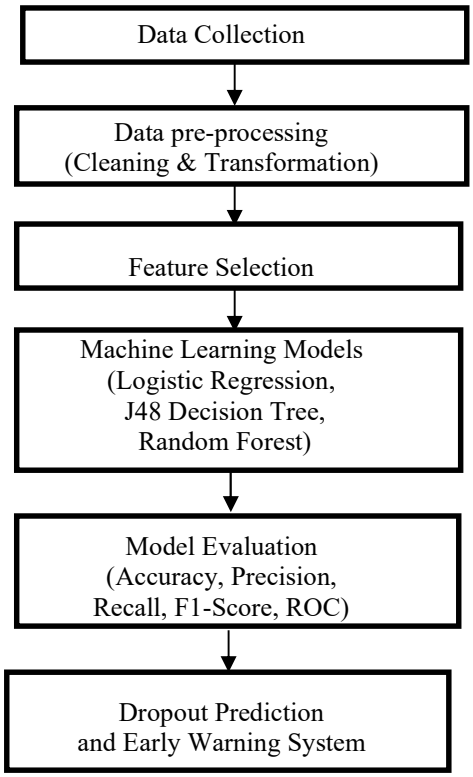


Fig.1 Research Methodology.

The method proposed in this study for predicting the factors affecting the failure of students belongs to the process of knowledge discovery from database and data mining. This methodology provides a systematic approach for identifying students at risk of dropout and supports timely intervention by educational institutions.

3.1 Collection and Dataset

The dataset was collected from 315 students enrolled in six colleges in Maihar district, Madhya Pradesh, India, including three government and three private colleges. It comprises 10 variables representing academic, socioeconomic, accessibility, psychological, and institutional factors. The variables include regular absence from college, academic performance, family income, parental education, transportation accessibility, digital connectivity, psychological well-being, lack of educational facilities, absence of adult protection, and institutional support. These variables collectively represent important factors influencing students' academic participation and educational experiences. Academic performance and attendance reflect educational engagement, while family income and parental education represent socioeconomic conditions. Transportation and digital connectivity indicate resource accessibility, whereas psychological well-being, educational facilities, adult protection, and institutional support represent broader student-support conditions. The dataset provides a suitable foundation for machine learning analysis.

Table 1: Variables related to dropout reason.

Attribute	Description	Possible value
Family Income	Dropout due to lack of income	Yes and no
Academic performance	Poor academic performance	Yes and no
Regular absence from college	Student continues to be absent from college	Yes and no
Parental education	Parent not educated	Yes and no
Transportation Accessibility	Limited availability of transportation	Yes and no

Digital connectivity	Limited access to internet & online services	Yes and no
Psychological well-being	An individual's mental and emotional health	Yes and no
Lack of educational facility	Inadequacy of college, learning resources, trained teachers, and educational infrastructure limits	Yes and no
Without adult protection	Especially children, who lack the care, supervision, and support of responsible adults	Yes and no
Institutional support	Assistance provided by government agencies	Yes and no

This study is basically done on the dataset DCS (Dropout College Students). The dataset contains 5 attributes where the number of instances is 315 to maintain the reliability, consistency and efficiency of this study.

3.2 Data Description

The dataset used in this research is collected from six undergraduate colleges in Maihar District, including 03 government colleges and 03 private colleges. Both primary and secondary sources are used to obtain reliable information.

Primary data is collected through structured questionnaires and personal interactions with students, while secondary data is obtained from institutes. Dataset (DCS) having 6 attributes where the number of instances is 315. These 5 attributes include 4 nominal attributes category (General, OBC, SC, ST), college type (Govt., Private), Dropout reasons (regular absence from college, academic performance, family income, parental education, transportation accessibility, digital connectivity, psychological well-being, lack of educational facility, without adult protection and institutional support), Dropout (Yes, No), and numeric attribute year (from 2023 to 2025).

Table 2: Description of datasets.

Attributes	Types of attributes	Description	Possible Value
Category	Nominal	Represents the category of students	GEN, OBC, SC, ST
College Type	Nominal	Describes the type of college	Govt., Private
Dropout Reasons	Nominal		Regular absence from college, academic performance, family income, parental education, transportation accessibility, digital connectivity, psychological well-being, lack of educational facility, without adult protection and institutional support.
Dropout	Nominal	Dropout or not	Y or N
Year	Numeric	Year of Dropout	2023, 2024, 2025

In this research, three attributes are selected as target attributes to achieve the desired results under different variations. The DCS dataset contains only one numeric attribute, 'Year', while the remaining attributes are nominal. The 'Year' attribute consists of only three values, namely 2023, 2024, and 2025; therefore, statistical analysis of this attribute is not considered necessary.

Furthermore, representing the various parameters is essential for understanding the influence and contribution of the remaining four nominal attributes. Such representation helps analyze their individual characteristics and overall impact on the investigation and the obtained results.

Table 3: Statistical value of dataset.

Dropout Attributes	Types of attributes	Count
Category	Nominal	
GEN		102
OBC		128
SC		49
ST		36
College Type	Nominal	
Govt.		218
Private		97
Year	Numeric	
2023		124
2024		113
2025		78
Dropout	Nominal	
Yes		234
No		81

3.3 Data Mining Tool

In recent years, WEKA is one of the most popular and emerging data mining tools and it is very user-friendly for the naïve users. WEKA (Waikato Environment for Knowledge Analysis) is an open-source tool for data mining under the GNU License overall population permits [11]. It contains an assortment of tools for preprocessing of data, grouping and regression analysis, clustering, classification, association rule mining and representation of data.

3.4 Data Preprocessing

Data pre-processing is an essential stage in the development of reliable machine learning models, particularly when working with educational datasets collected from heterogeneous sources. Raw datasets may contain missing values, inconsistent formats, duplicate records, inaccurate entries, and other data-quality issues that can adversely affect the performance of predictive models.
Therefore, appropriate preprocessing techniques are required to improve data quality, consistency, and suitability for subsequent analysis. In the present study, data preprocessing involves data cleaning, data transformation, and feature selection to prepare the dataset for the development and evaluation of student dropout prediction models.

3.5 Data Cleaning and Transformation

Data cleaning is one of the most important activities within the preprocessing stage. The primary objective of data cleaning is to identify and correct or remove erroneous, incomplete, inconsistent, and irrelevant observations. In the present study, missing or NA values were examined and appropriately addressed to minimize their potential effect on the prediction results. In addition, inaccurate records and potential outliers were identified and removed where necessary. This process helps reduce noise and prevents inappropriate observations from influencing the learning process and degrading the predictive capability of the developed models.
The target variable in this study represents the educational status of students and indicates whether a student dropped out (Yes) or continued education (No). The original dataset was maintained in Microsoft Excel (.xls) format. To facilitate data processing and compatibility with machine learning tools, the Excel dataset was subsequently converted into Comma-Separated Values (.csv) format. Following this conversion, the CSV dataset was transformed into Attribute-Relation File Format (.arff), which is supported by the Waikato Environment for Knowledge Analysis (WEKA) platform. The ARFF format enabled the dataset to be efficiently interpreted and processed during the implementation of the machine learning algorithms.

3.6 Feature Selection

Feature selection is another important stage of data pre-processing that involves identifying the most relevant attributes associated with the target variable while eliminating redundant or less informative variables [8]. Selecting appropriate features can improve model accuracy, reduce computational complexity, minimize overfitting, and enhance the interpretability of the resulting prediction models.
The selection of relevant attributes in this study was performed based on an extensive literature review and preliminary analysis of the dataset. The variables initially considered included regular absence from college, academic performance, family income, parental education, transportation accessibility, digital connectivity, psychological well-being, lack of educational facilities, absence of adult protection, and institutional support.
Following the analysis, regular absence, academic performance, family income, digital connectivity, and institutional support were identified as the most influential variables for predicting student dropout. Consequently, these five attributes were selected as the final input features and were consistently used across all machine learning models investigated in this study. The selected features provide a focused representation of the academic, socioeconomic, technological, and institutional factors associated with student dropout, thereby supporting the development of an effective and interpretable predictive framework.

4. Proposed Machine Learning-Based Dropout Prediction Model:

The proposed framework follows the Knowledge Discovery in Database (KDD) process to identify Students at risk of dropping out before they permanently discontinue their studies. The framework combines data collection, preprocessing, machine learning classification, and institutional intervention into a single predictive system.

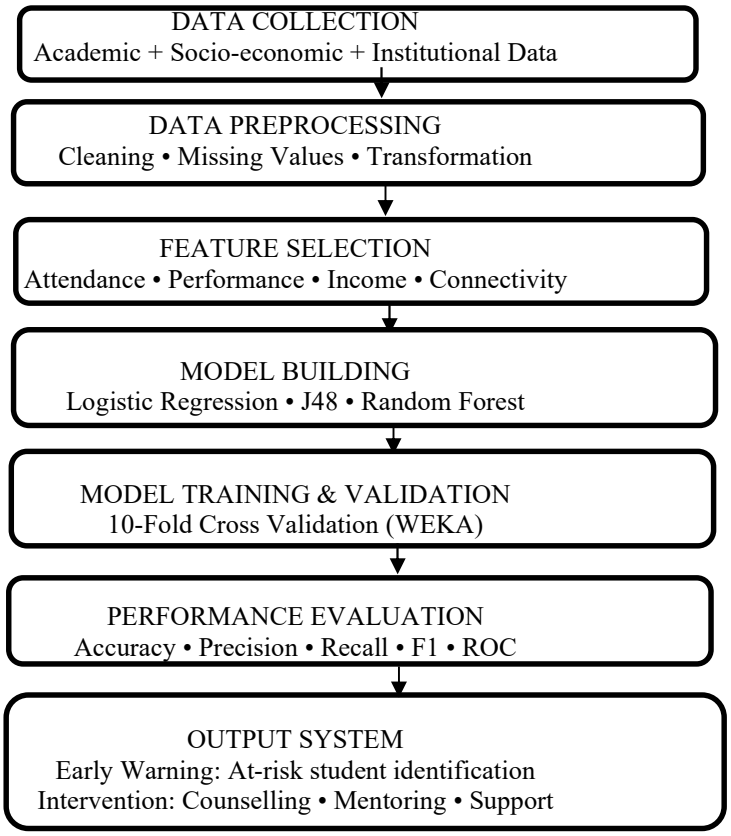


Fig.2 Proposed Machine Learning Model.

The proposed model enables educational institutes to identify high-risk students at an early stage and implement targeted interventions such as academic mentoring, financial assistance, counselling, and attendance monitoring. By integrating predictive analytics with institutional decision-making, the framework supports proactive student retention strategies and improves educational outcomes.

After completing data pre-processing and feature selection, supervised machine learning algorithms were applied using the WEKA data mining environment to predict student dropout risk. The main objective is to classify students into two categories: Dropout (Yes) and Non-Dropout (No). Three widely used classification algorithms -logistic regression, decision tree (J48) and random forest- are selected for model development and comparative analysis. These algorithms are chosen because of their effectiveness in educational data mining and their capability to identify important factors influencing student attrition.

4.1 Logistic Regression

Logistic Regression is a statistical classification method used to predict binary outcomes [10]. In this study, it estimated the probability of student dropout using academic, socio-economic, and institutional factors, including regular college absence, academic performance, family income, parental education, transportation accessibility, digital connectivity, psychological well-being, lack of educational facilities, absence of adult protection, and institutional support. The model was developed in WEKA using 10-fold cross-validation and achieved an overall classification accuracy of 84.80%. This result indicates satisfactory performance in identifying students at risk of dropping out and demonstrates the model's usefulness for analyzing factors associated with student dropout [7].

4.2. Decision Tree

Decision tree is a popular classification technique widely used in educational data mining. In WEKA, the J48 algorithm, an implementation of Quinlan's C4.5 decision tree, constructs a tree structure by recursively splitting the dataset based on attributes with the highest information gain.

In this study, regular absence from college emerged as the most influential predictor and became the root node of the decision tree. [12] Other important attributes included academic performance, family income, digital connectivity, and institutional support. The generated tree produced several understandable classification rules, such as students with regular absence and poor academic performance being highly likely to drop out.

The decision tree generated several interpretable classification rules. For example:

Rule 1: IF Regular Absence = Yes AND Academic Performance = Poor THEN Dropout = Yes
Rule 2: IF Regular Absence = Yes AND Family Income = Low THEN Dropout = Yes
Rule 3: IF Digital Connectivity = No THEN Dropout = Yes
Rule 4: IF Institutional Support = Yes THEN Dropout = No

The J48 model achieved an accuracy of 88.57%, demonstrating better predictive performance than logistic regression. Its transparent decision rules make it particularly useful for educational administrators seeking to identify and support at-risk students.

4.3 Random Forest

Random forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting [1]. Each tree is constructed using bootstrap samples and random subsets of predictor variables, while the final classification is determined through majority voting.

The algorithm effectively captured complex relationships among attendance, academic performance, family income, transportation accessibility, digital connectivity, and institutional support. Its advantages include high accuracy, robustness against noise, handling of large datasets, and improved generalization capability.

Implemented in WEKA, the random forest model achieved the highest classification accuracy of 93.02%. It also produced superior precision, recall, F1-score, and ROC values compared to the other models, indicating strong predictive capability for an educational early warning system.

5. Experimental Results and Performance Evaluation Model

Table 4: Performance Model Evaluation

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC
Logistic Regression	84.80%	0.84	0.83	0.84	0.87
Decision Tree	88.57%	0.88	0.87	0.88	0.91
Random Forest	93.02%	0.92	0.93	0.92	0.97



Fig.3 Performance Comparison chart of Models.

The random forest classifier achieved the highest predictive performance among all evaluated models.

The Logistic Regression model achieved an accuracy of 84.80% with satisfactory values for Precision, Recall, F1-Score, and ROC Area. Although it performed well, its predictive capability was lower than the tree-based algorithms because it assumes linear relationships among predictor variables.

Overall, the WEKA experimental results indicate that Random Forest is the most effective algorithm for developing an educational early warning system to predict student dropout.

The performance of three machine learning models is evaluated using WEKA data mining software. To ensure reliable and unbiased prediction, a 10-fold cross-validation technique was employed. This approach divides the data set into ten equal subsets, using nine subsets for training and one subset for testing in each iteration.

The process is repeated ten times, and the average performance is reported. Model performance was assessed using standard evaluation metrics including Accuracy, Precision, Recall, F1-Score, ROC Area, and Confusion Matrix. The following metrics are used to compare the predictive performance of the classification algorithms

Accuracy: percentage of correctly classified students' records.

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN)$$

Precision: Proportion of predicted dropout students who are actually dropouts.

$$\text{Precision}=TP/(TP+FP)$$

Recall: Ability of the model to correctly identify actual dropout students

Recall=TP/(TP+FN)

F1-Score: Harmonic mean of precision and recall

F1 = 2 × (Precision × Recall) / (Precision + Recall)

ROC Area: Measures the model's ability to distinguish between dropout and non-dropout students

6. Experimental Results

6.1 Confusion Matrix Analysis

A confusion matrix is a performance evaluation tool used in data science to measure the accuracy of a classification model. It compares the model's predicted results with the actual values. The matrix contains four outcomes: [6] True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). By analyzing these values, data scientists calculate metrics such as accuracy, precision, recall, and F1-score to evaluate model performance and identify classification errors for improvement. To further evaluate the performance of the best-performing model, the confusion matrix of the random forest classifier is analyzed.

Random Forest

Table 5: Confusion Matrix of Random Forest.

Actual predict	Predict dropout	No dropout
Dropout	**218**	**16**
No dropout	**6**	**75**

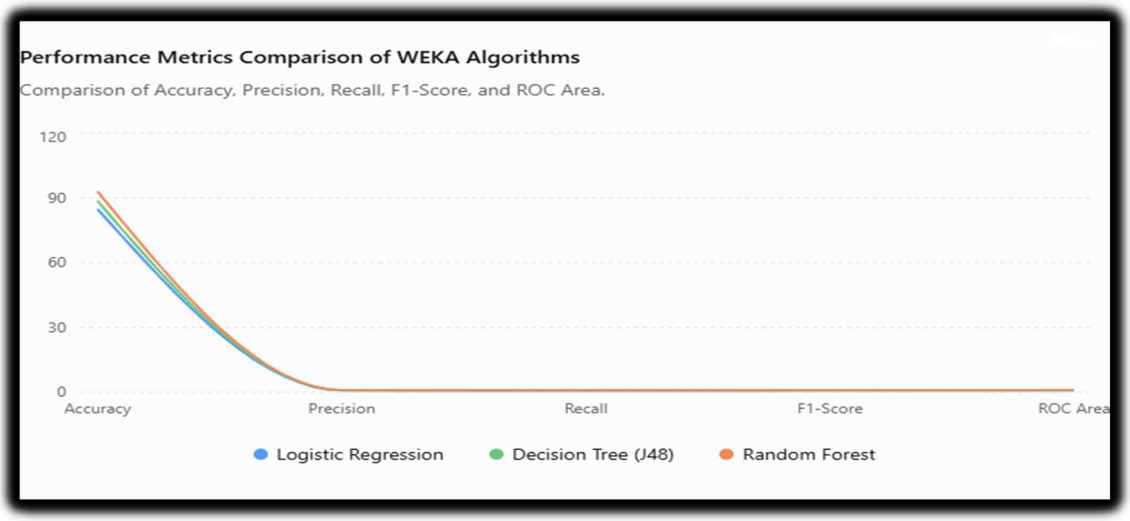


Fig.3 Graph of performance metrics of WEKA.

6.2 Interpretation

The confusion matrix presented in Table 5 demonstrates the classification performance of the Random Forest model in predicting student dropout status. The model correctly classified 218 students as dropout, indicating a high ability to identify students who actually belonged to the dropout category. Similarly, 75 students who did not drop out were correctly classified as no dropout. These values represent the true positive and true negative predictions, respectively.

The model incorrectly classified 16 students as dropout when they were actually in the no-dropout category, representing false-positive predictions. Conversely, 6 students who actually dropped out were predicted as no dropout, constituting false-negative predictions. Thus, the model produced 293 correct classifications out of 315 observations, while only 22 cases were misclassified.

Overall, the confusion matrix indicates that the Random Forest model performed effectively in distinguishing between dropout and no-dropout students. The relatively small number of false negatives is particularly relevant because it indicates that the model missed only a limited number of students who actually dropped out. Based on these results, the model demonstrates strong predictive capability and appears particularly effective at identifying students at risk of dropout. However, additional performance measures, such as accuracy, precision, recall, F1-score, and specificity, should also be considered to provide a comprehensive assessment of model performance.

The green line representing Random Forest remains consistently above the blue and orange lines, indicating the best overall performance in Accuracy, Precision, Recall, F1-Score, and ROC Area. The Decision Tree model shows intermediate performance, while Logistic Regression records the lowest values among the three models. These findings confirm that Random Forest is the most reliable and accurate classifier for student dropout

prediction in the WEKA environment. The random forest model achieved this objective effectively, making it suitable for implementation as an institutional early warning system.
The findings of this study indicate that student dropout is influenced by a combination of academic, socio-economic, and institutional factors rather than a single variable. Among all predictors, regular attendance emerged as the strongest determinant of student retention. Students with frequent absenteeism are significantly more likely to discontinue their studies before graduation. Poor academic performance further increased the probability of dropout, indicating the need for continuous academic monitoring and timely support.

7. Conclusion

This study investigated student dropout among undergraduate colleges in Maihar District, Madhya Pradesh, using machine learning techniques implemented in WEKA. A dataset of 315 student records collected from six colleges between 2023 and 2025 was analyzed. Data pre-processing, feature selection, and classification techniques were employed to identify significant predictors of dropout.
The Decision Tree analysis revealed that regular absence, academic performance, family income, digital connectivity, and institutional support are the most influential determinants of student attrition. The generated decision rules provide a practical framework for identifying at-risk students. While the J48 Decision Tree achieved an accuracy of 88.57%, Random Forest produced the highest predictive performance with an accuracy of 93.02%.
The study demonstrates that machine learning can serve as an effective early warning system for educational institutions. The findings can assist policymakers and administrators in designing targeted interventions to improve student retention and promote sustainable higher education development in rural and semi-urban regions.

8. Future work

Future research can extend this study by using larger data sets collected from multiple districts of Madhya Pradesh to improve model generalization. Additional variables such as gender, semester-wise, extracurricular participation, family occupation, health status, and psychological factors may further enhance prediction accuracy.
Advanced techniques such as Extreme Gradient Boosting, Support Vector Machine, Artificial Neural Network, and Long Short-Term Memory network can also be explored for improved predictive performance.

Acknowledgement

The author would like to express their sincere gratitude to Dr. Preetam Singh Gour, Chief Editor, IJIRG, India, for his valuable suggestions and encouragement, which contributed to the development and refinement of this research work.
The author sincerely acknowledges the support and cooperation of all individuals who directly or indirectly contributed to the successful completion of this study entitled “Machine Learning-Based Prediction of College Students’ Dropout Risk: A Case Study of Maihar District, Madhya Pradesh, India.”

Conflict of Interest Statement

The author declares that there is no conflict of interest, financial or otherwise, regarding the research, authorship, analysis or publication of this manuscript.

Funding Statement

The author declares that no specific financial support or external funding was received from any public, commercial, or non-profit organization for conducting this research.

References

[1] Bean, J. P., Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, 55(4), 485–540. doi: 10.3102/00346543055004485.
[2] Romero, C., Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601–618. doi: 10.1109/TSMCC.2010.2053532.
[3] Bhise, R. B. (2013). Importance of data mining in higher education system. *OSR Journal of Humanities and Social Science*, 6(6), 18–21. doi: 10.9790/0837-0661821.
[4] Kotsiantis, S. B. (2009). Educational data mining: A case study for predicting dropout-prone students. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 1(2), 101–111. doi: 10.1504/IJKESDP.2009.022718.
[5] Selvan, M. P., Navadurga, N., Prasanna, N. L. (2019). An efficient model for predicting student dropout using data mining and machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering*, 8(9S2), 750–752. doi: 10.35940/IJITEE.I1155.0789S219.
[6] Altamimi, A. B. (2023). Big data in education: Students at risk as a case study. *Engineering, Technology & Applied Science Research*, 13(5), 11705–11714. doi: 10.48084/etasr.6190.

[7] Witten, I. H., Frank, E., Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, pp. 629.
[8] Alban, M. S., Mauricio, D. (2018). Factors to predict dropout at the universities: A case of study in Ecuador. in *IEEE Global Engineering Education Conference (EDUCON)*, 1238–1242, doi: 10.1109/EDUCON.2018.8363371.
[9] Pandey, U. K., Pal, S. (2011). Data mining: A prediction of performer or underperformer using classification. *Journal of Computer Science and Technology*, 4(2), 686-690.
[10] Dasi, H., Kanakala, S. (2022). Student dropout prediction using machine learning techniques. *International Journal of Intelligent Systems and Applications in Engineering*, 10(4), 408–414.
[11] Ben-Zadok, G., Hershkovitz, A., Mintz, R., Nachmias, R. (2007). Examining online learning processes based on log files analysis: A case study. in *Research, Reflections and Innovations in Integrating ICT in Education*, 55–59.
[12] Rokach, L., Maimon, O. Z. (2008). *Data Mining with Decision Trees: Theory and Applications*, 69. Singapore: World Scientific, pp. 244.