

Student Academic Risk Classification Using Machine Learning: A Comparative Analysis of Predictive Models

Natesh Mahadev¹, V. L. Sowmya², Anitha Premkumar³, N. Thangarasu⁴

¹Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, Karnataka, India

²Department of Artificial Intelligence and Machine Learning, BMS Institute of Technology & Management, Bengaluru, 560119, India

³Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India

⁴Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore-21, India

Article Info

Article history:

Received July 04, 2026

Revised August 25, 2026

Accepted September 14, 2026

Keywords:

Student Academic Risk
Machine Learning
Educational Data Mining
Student Performance Prediction
Risk Classification
Academic Analytics

ABSTRACT

There are systems in place to identify at-risk students in educational facilities, however early identification allows for more effective support to be allocated. This research presents a framework to identify students at risk in a tertiary institution using contextual, behavioral and academic data. The framework uses a supervised machine learning approach to classify students as Safe, At-Risk or High-Risk. Eight different classification models were trained and evaluated using five-fold cross-validation. The classification performance of the models was analyzed using accuracy, precision, recall and the F1-score. Discrimination analysis was conducted to evaluate the difference in classification. Of the eight classification models, the support vector classifier (SVC) and Logistic Regression model outperformed the others with the SVC achieving 93.63% accuracy. The classification models demonstrated a preference to overfit on the training data. A feature importance analysis was performed to identify the most influential features to the classification models. Some of the more influential features to the classification of risk included students' previous academic performance and attendance. Based on the results, support systems to identify at-risk students academically can be designed and implemented.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author: Anitha Premkumar (e-mail: p.anitha@manipal.edu)

1. INTRODUCTION

The multiple dimensions of student academic performance include: engagement and participation in the learning process, the student's previous academic performance and behavior, how the student studies, and the environment in which the student engages in the learning process. Identifying students at risk of underachieving academically enables educational institutions to prepare and implement timely and effective interventions to support the students. Traditional methods of identifying at-risk students mainly depend on teachers' observations and students' academic performance and records, and they often lead to the identification of students at risk at a late stage. It is for this reason that several researchers have focused on predicting students' academic performance using data science and other allied fields.

Educational Analytics combines multiple dimensions of students' performance in ways that Academic Performance, Behavioral, and Social-Emotional dimensions of students had not been previously integrated. Prediction of students' performance and achievement in the field of Educational Data Mining has gained increased attention, despite challenges of applying and adapting it to different educational situations. It has also focused on refining methods of classification and risk prediction. Studies have documented the use of varying datasets and different techniques in academic performance and achievement prediction. These studies have shown the need for further and more methodical studies in the field.

The current study explores student characteristics related to different components of the learning process and environment. Factors such as attendance, study hours, and completion of assignments reflect the

level of participation and engagement in the learning process. Past grades and failures describe past learning experiences. Other variables included in this study describe the family situation and social environment of a student. These variables describe the student's context and environment and could be associated with learning opportunities and academic achievements.

Academic risk is defined as the probability of not achieving academic goals and standards. For this study, academic risk is considered at three levels: safe, at-risk, and high-risk. Traditionally, the term 'academic risk' is used to describe students who failed or dropped out of a course. Other students with less severe academic concerns are often unrecognized. In the field of early-warning research, the use of machine learning in recognizing at-risk students has gained popularity, and the focus has shifted to the timely implementation of remedial measures [1]. Given the variability in the student data, the statistics and data analysis should take this into account. Both numerical and categorical variables were reported in the student data. To use the variables in classification, certain data transformations are required. In the current study, categorical variables were converted to numerical values. Numerical variables were analyzed after replacing missing values and standardizing the values.

The study investigates several learning strategies using Logistic Regression, Gaussian Naive Bayes, SVC, KNN, Decision Tree, Extra Trees, Random Forest, and Bagging. Different classification behaviors are analyzed by various modeling principles, such as classification by a line, classification by probability, classification by similarity, and classification by a tree. This provides an opportunity to analyze classification behavior without distinguishing the principles of data processing. This kind of analysis is important, since no ML algorithm has been shown to be clearly dominant in the field of educational data prediction [2]. To assess models, several methods have been used, such as classification by a line, by probability, and by a tree. Moreover, to evaluate classification behavior, an analysis of classification by a tree was done using a one-against-the-rest approach. To gain a deep understanding of classification behavior, other methods such as a five-fold cross-validation and analysis of accuracy, precision, recall, and F1-score were performed.

Because classification results should be explained by a human, the influence of each feature on the classification result was analyzed. Interpreting the results allows teachers to better understand the reasoning for classification and provide suggestions for improving student learning [3]. For this research, feature importance was analyzed to improve understanding of the classification of academic risk. While there are studies that explore the use of ML to predict academic performance and dropout using different terminologies and frameworks, there are very few studies that use a unified framework for the classification of academic risk to include behavioral and contextual factors to identify Safe, At-Risk, and High-Risk students. The current study proposes and implements a comprehensive machine learning framework to classify levels of academic risk and to identify characteristics of students based on that classification.

The problem statement can be addressed as follows:

- Classify students as Safe, At-Risk, or High-Risk.
- Determine the variables that led to the classification of students as At-Risk.
- Evaluate classification performance using different classification algorithms.
- Assess the classification and distinction performance of different classification algorithms.
- Determine the most notable factors that led to the classification of students as at-risk.

2. RELATED WORK

There are several applications of ML and EDM to detect at-risk students, including students who experience academic underachievement and students who leave the education system. These studies have assessed and integrated behavioral, social, and demographic factors and employed different classification techniques. Prediction of student dropout and changes in academic status are closely related and have been studied extensively. In one study, the researchers formulated a system to predict student dropout using the integration of different engineering and scientific techniques in ML and responded to class imbalance using SMOTE. Other studies have focused on the integration of different techniques in classification and employed different approaches to address class imbalance and dimensionality. These studies have laid the foundation for the use of data science and integration of different techniques to address class imbalance and other challenges that hinder the identification of at-risk students [4].

Evaluation of traditional classification techniques has been studied as well. For example, a study evaluated the performance of Decision Trees and Random Forests in predicting students' academic performance [5]. The authors suggested that evaluation procedures should be standardized when comparing the performance of different classification techniques. Data balance has been an important consideration for predicting educational outcomes. The effect of SMOTE data augmentation has been studied by [6] for a range of educational datasets, and the authors reported that resampling techniques greatly impacted the classification results. Similarly, [7] performed classification for predicting secondary school graduation using

Logistic Regression, Random Forest, XGBoost, and SVM and reported that implementation details and choices of preprocessing techniques influenced the results significantly. The above-mentioned studies suggest that a combination of appropriate data processing and model selection is critical for developing high-quality educational prediction systems.

One of the recent advances in the area of educational predictive systems is presented by [8], where the authors proposed an LSTM autoencoder-based framework for educational decision-making. Used a range of classification and regression models for postgraduate students' academic outcome prediction and concluded that the integration of various types of student data (e.g., academic, demographic, and pre-enrollment data) enabled them to capture various types of student performance and behavior. These studies demonstrate the application of predictive learning to capture different educational outcomes using advanced and complex architectures [9].

Other research focuses on student-related issues of a broader nature and on other methodological issues [10]. One study used different ML methods to create models that assess the depression risk of university students using variables related to academics, finance, and social relationships. This study [11] illustrates that, with ML, student variables beyond academics can be analyzed. A systematic review focused on prediction studies of various clusters and discussed the challenges of using ML in different institutions. Another study reported that, in the context of ML and educational research, the use of different algorithms for the same task can lead to inconsistent results and, therefore, different conclusions, and this is particularly valid when the performance of the models is analyzed without sufficient justification.

The same type of ML research focused on prediction of academic performance has been extended to other fields [12]. For instance, one study used different ML methods for classification and comparison in the field of finance, and reported that, although the data may be related to different fields, the use of different ML methods can help address the same problem in different ways. Other studies [13] have also extended research focused on predicting student dropout to other areas beyond academic performance, and reported the use of ML methods for predictions and integrative, interpretable, and exploratory analyses.

Research on ML algorithms indicates that no one algorithm covers all territories in the field of education. Research focused on the prediction of academic performance (PAP) is comprehensive, large-scale, and addresses the differences in datasets, algorithms, variables, and metrics [14]. Other studies in the area of PAP address the balance and outliers in the samples. Research on predicting academic achievement and dropout focuses on varying algorithms, which cover different frames of the educational domain [15]. Likewise, research in predicting academic performance and dropout concentrates on the variability of the algorithms and addresses the scarcity and imbalance in the samples, using multiple classifiers and the majority voting scheme [16].

Predicting student multilevel academic risk is not extensively studied, and literature focusing on other academic outcome risks, including underachievement, is even more scarce. Most studies focus on predicting academic outcome risks, student under- and overachievement, failures, and student dropout. The reasons for student underachievement may vary at different levels and may include the student's academic, behavioral, and social contextual conditions. Most of the works address the variability of student outcome risks by focusing on student achievement, behavior, underachievement, and dropout. This is done by exploiting the differentiability and plasticity of ML algorithms, and typically there is little to no concern for the interpretability of the models.

This study aims to build upon previous research by creating a more nuanced academic risk classification framework. The framework will consist of three classes: safe, at risk, and high risk. The study will gather and analyze data based on traditional academic indicators as well as behavioral and contextual data. The study will use standard machine learning (ML) classifiers after adjusting and selecting the most relevant features. The study aims to gain a comprehensive understanding of each classifier by using cross-validation and alternative assessments of model behavior. To extract the features most relevant to risk classification, the study will assess and evaluate the features in relation to risk classification. Overall, the goal of this study is to comprehensively assess and analyze risk in the academic context as opposed to predicting or explaining a single outcome (i.e., dropout).

3. METHOD

This study presents a machine learning model for labeling students as At-Risk, High-Risk, or Safe. The model development process consisted of several steps, including data preparation, transformation, model building, and testing. The complete process is illustrated in Figure 1.

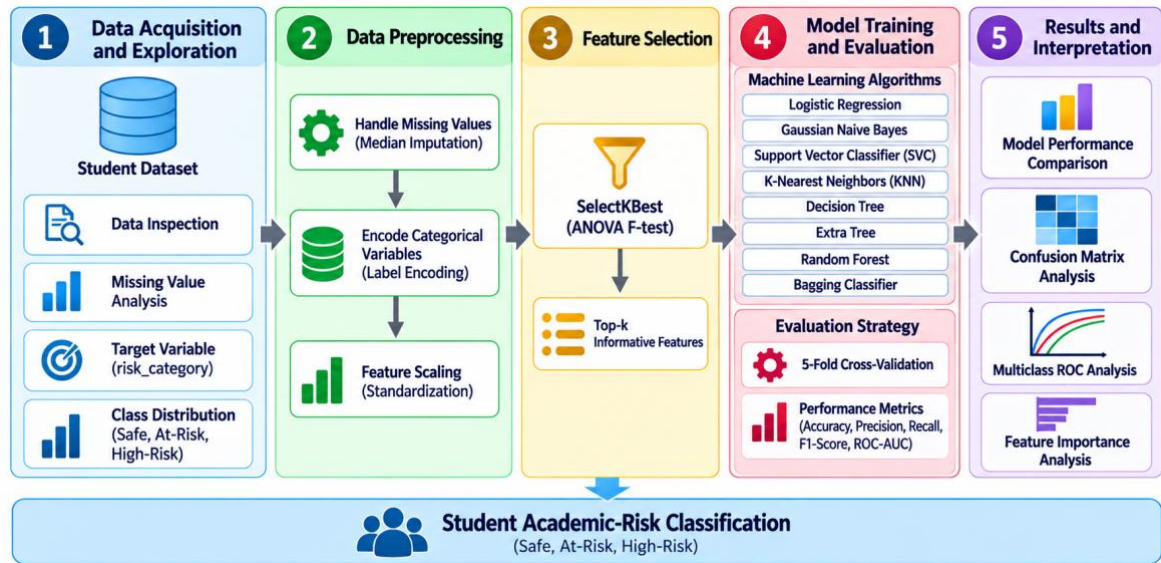


Figure 1. Proposed Machine Learning Framework for Student Academic-Risk Classification

3.1. Dataset and Data Preprocessing

The dataset used in this study has 100,000 records. It contains 10 predictors and 1 response variable. The predictors are attendance, study hours, past failures, assignment completion, parental education, family income, extracurricular activities, internet access, previous grade, and final grade. The response variable is risk category, which consists of At-Risk, High-Risk, and Safe. The data were initially screened for errors, and it was discovered that some students were missing values for attendance and study hours. Median imputation was performed for the predictors containing numerical values. For a numerical variable X_j , if the value was missing, it was replaced with the median of all the values of X_j using Equation (1).

$$X_{ij}^* = \begin{cases} X_{ij}, & X_{ij} \text{ is available} \\ \text{median}(X_j), & X_{ij} \text{ is missing.} \end{cases} \quad (1)$$

Equation (1) shows that if the value is not available, it will be replaced with the median of all the available values. After the value was imputed, no observations were found to be missing. The categorical variables were represented numerically based on the frequency of each value using Equation (2).

$$f(c) = \sum_{i=1}^N I(x_i = c) \quad (2)$$

Equation (2) shows that the value of a categorical variable is converted to the number of observations having that value. The value of the target variable was converted to numerical values using LabelEncoder, where the value of At-Risk is assigned 0, High-Risk is assigned 1, and Safe is assigned 2. The X included the ten input variables, and y was the risk class factor.

3.2. Data Partitioning, Scaling, and Feature Selection

The course data were divided into training and testing sets using an 80:20 proportion, with random state 42. The training set mean and standard deviation were used to perform feature standardization for the training set and were applied to both the training and testing sets, as shown in equation (3).

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (3)$$

Where x_{ij} is the feature value, μ_j and σ_j are the mean and standard deviation of the training set, respectively. Standardizing features in this manner benefits distance and margin classifiers, like KNN and SVM, which rely on those principles to make classifications. Feature selection was performed using the ANOVA F-test and the K Best feature selection technique. The test retained eight features from the training set. The ANOVA test assesses the feature value in relation to the average value of the feature across the categories. Features that have larger F-values indicate greater differentiability across the categories.

$$F_j = \frac{\sum_{k=1}^K n_k (\bar{x}_{jk} - \bar{x}_j)^2 / (K-1)}{\sum_{k=1}^K \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{jk})^2 / (N-K)} \quad (4)$$

where K is the number of classes, n_k is the number of observations in class k , $\bar{x}_{jk}$ is the mean of feature j in class k , and $\bar{x}_j$ is the overall mean. Equation (4) measures the extent to which a feature differs across the risk categories; features with stronger class-separating capability receive larger F -statistics.

3.3. Best-Performing Classification Model

A set of eight classification models was developed as part of the experiment. Because the main interest of this study is on the best-performing classifier, the other models are not discussed in detail. The Support Vector Classifier (SVC) is examined in detail, therefore, based on the results of this experiment. Given the large sample size, an SVC with a linear kernel was developed. Such a model finds a separating hyperplane so that the distance from the hyperplane to the closest observation in the training sample is maximized. The decision function, therefore, takes the form.

$$f(x) = w^T x + b \quad (5)$$

Where w and b are the weight and bias, respectively, and x is the sample vector. A hyperplane is defined by the decision function as shown in equation (5). The objective of the model is to minimize the following function as represented in equation (6).

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (6)$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (7)$$

The constraint is satisfied, where y is the class label. Equations (6) and (7) define the margin for the classification function. Equation (6) is the penalty function for classification error. The classification function finds a decision boundary so that the training sample is maximally classified. The function allows for slack in the classification, in which case the variable ξ_i represents the degree of slack in the classification function. The margin is larger when the function allows for more classification error. In the case of SVC, the margin is defined by the support vectors.

3.4. Cross-Validation and Performance Evaluation

The SVC was configured to permit the estimation of classification probability. For the case of multi-class classification, the SVC determines class-discriminant functions and assigns sample classification based on the function. For N observations, classification accuracy was defined as represented in equation (8).

$$Accuracy = \frac{1}{N} \sum_{i=1}^N I(y_i = \hat{y}_i) \quad (8)$$

where y_i and $\hat{y}_i$ denote predicted and actual classes, respectively. Equation (8) calculates the overall classification accuracy with respect to the student population under consideration. For the case of multiclass classification, weighted precision was computed as shown in equation (9) and weighted recall as shown in equation (10).

$$Precision_w = \sum_{k=1}^K \frac{n_k}{N} Precision_k \quad (9)$$

$$Recall_w = \sum_{k=1}^K \frac{n_k}{N} Recall_k \quad (10)$$

In the above equations, n_k represents the total number of samples belonging to class k . Equations (9) and (10) provide the overall classification performance of the model considering all the classes and respective samples. The weighted F1 score of a class is computed as shown in equation (11).

$$F1_w = \sum_{k=1}^K \frac{n_k}{N} \left(\frac{2 Precision_k Recall_k}{Precision_k + Recall_k} \right) \quad (11)$$

Classification performance of the model is further analyzed using class-specific classification errors, based on 5-fold cross-validation. Further, to evaluate the discrimination of the model across the three risk factors, we have relied on a one-versus-all strategy and reported the macro average value of the resulting ROC curves. Lastly, to validate our classification system, we relied on feature importance to identify strong predictors of student academic risk.

4. RESULTS AND DISCUSSION

The experimental analysis is focused on the evaluation of the student academic risk classification framework developed during the course of this research. The evaluation is based on the results of the models across various validation folds. The features selected, which encompassed the student's academic and behavioral performance as well as contextual features of the learning environment, were also considered. Classification differences between the considered learning approaches were analyzed. Various classification and regression techniques were employed to address the research questions. Classification performance of the developed models was analyzed using confusion matrices and ROC curves. The training and test generalization was also analyzed. The final element of the analysis was to explore the reasons for the academic risk classification. These analyses provided a foundation for the interpretation of the framework and helped identify the blind spots in classification.

4.1. Confusion Matrix Analysis

The figures in 2 illustrate the classification predictions of the three classes: At-Risk, High-Risk, and Safe. Each of the matrices has actual class values for a row, and predicted class values for a column. A position on the main diagonal of the matrix indicates that classification was correct. All off-diagonal elements represent classification errors. It is evident from the matrices that the models were able to classify At-Risk and Safe class values with minimal error. The errors occurred mainly in the classification of the High-Risk class values. It is evident from the matrices that the SVC and Logistic Regression models illustrated the best results, as they had the higher concentration of values on the main diagonal. Relatively higher off-diagonal values were observed for K-Nearest Neighbors and Extra Tree models. Figure 2 illustrates the prediction behavior of the models for the different risk class values and the errors that occur in the classification.

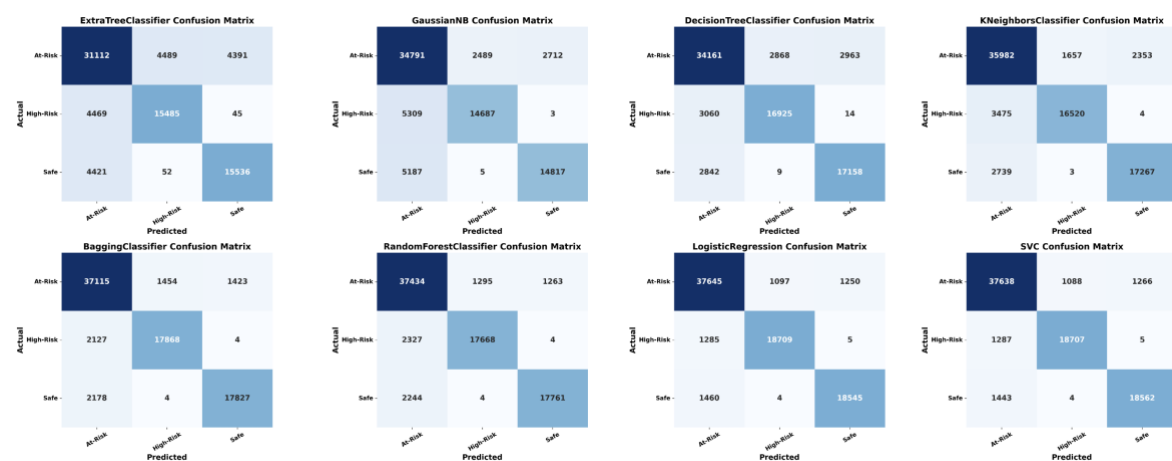


Figure 2. Evaluated ML Classifiers for Student Academic Risk Classification: Confusion Matrices

4.2. Performance Comparison

Figure 3 shows the assessment of the eight classifiers based on accuracy, precision, recall, and F1 score. The Support Vector Classifier (SVC) showed the best performance out of all the classifiers, with an accuracy of 93.63%, precision of 93.64%, recall of 93.63%, and F1 score of 93.63%. Logistic Regression performed second best and had a score of 93.62% for accuracy, recall, and F1 score, and a precision of 93.63%. Random Forest and Bagging performed at approximately 91% for each of the four performance metrics, and K-Nearest Neighbors (K-NN) performed at 87.21% accuracy and recall, with 87.35% and 87.2% precision and F1 score. The Decision Tree Classifier performed at 85.3% for all four metrics. The Naive Bayes Classifier performed at 80.37% for each of the four metrics. Extra Tree performed at 77.67% for each of the four metrics. The classifiers were evaluated based on predictive performance.

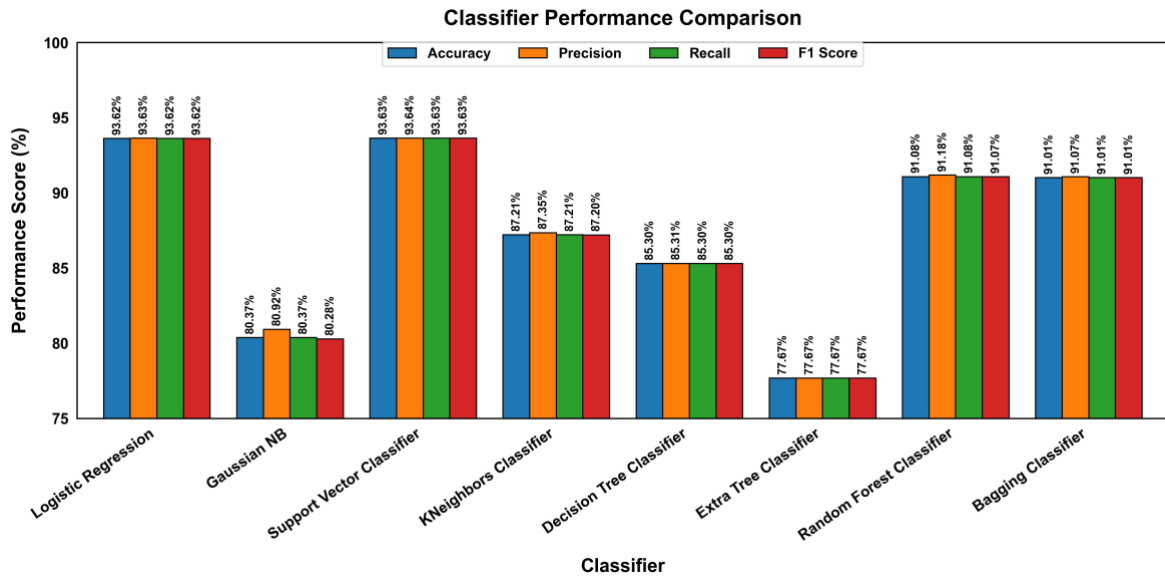


Figure 3. Comparative Performance of Machine Learning Classifiers

4.3. Multiclass ROC-AUC Analysis

Macro-average ROC curves and the corresponding AUCs are shown in Figure 4. The best discrimination of the classifiers is shown by the SVC and Logistic Regression classifiers, with an AUC of 0.993. Random Forest and Bagging classifiers have an AUC of 0.987, followed by K-Nearest Neighbors at 0.965. The rest of the classifiers have an AUC less than 0.9. Gaussian Naive Bayes has the lowest, at 0.938. The rest of the classifiers have an AUC in the range of 0.89 to 0.837. Discrimination of the classifiers is done by comparing the ROC curves, and based on that, it is shown that SVC and Logistic Regression classifiers, along with Random Forest and Bagging classifiers, are the best, since their ROC curves touch the upper-left region of the graph and are the farthest from the other classifiers.

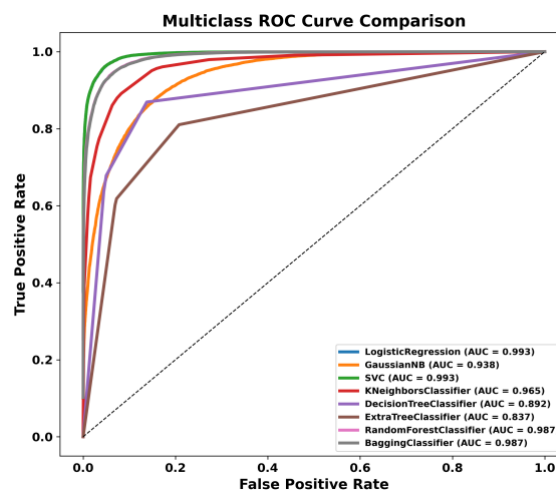


Figure 4. Multiclass ROC Curves of the Evaluated Machine Learning Classifiers

4.4. Multiclass Precision-Recall Analysis

Precision and recall scores are determined as follows: for a given class (i.e., low, medium, or high risk), the area under the corresponding precision-recall curve is calculated and referred to as average precision (AP). Figure 5 shows the multiclass precision and recall curves as well as average precisions for the classification models. Based on the average precision, the SVC and Logistic Regression models performed the best (i.e., highest precision and recall scores), followed by the Random Forest and Bagging models. The K-Nearest Neighbors model again performed the worst. As with the ROC analysis, the SVC and Logistic Regression models performed best, followed by the Random Forest and Bagging models. Figure 5 also shows the precision-recall curves for the tree models (i.e., the Decision Tree, Extra Trees, and Gaussian Naive Bayes models). It should be noted that the model performs worse in terms of precision for higher recall.

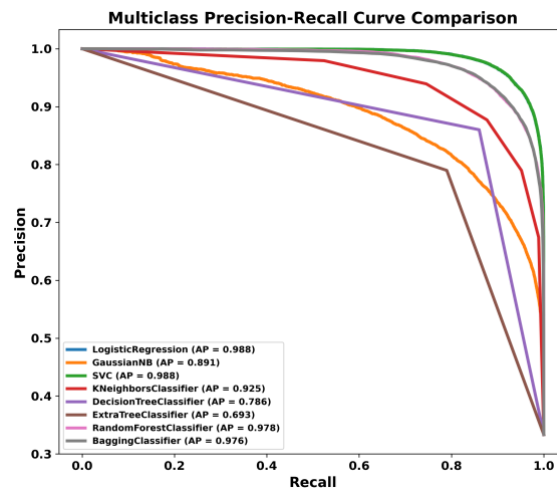


Figure 5. Multiclass Precision–Recall Curves of the Evaluated Machine Learning Classifiers

4.5. Importance of Features

Figure 6 illustrates the importance of the various features in the SVC model. The feature importance is determined by the relative change in the classification error when the feature is permuted. The features ‘previous academic performance’ and ‘study behavior’ are of particular importance. Less important are features such as final grade, extracurricular activities, parental education and income, and students’ internet use.

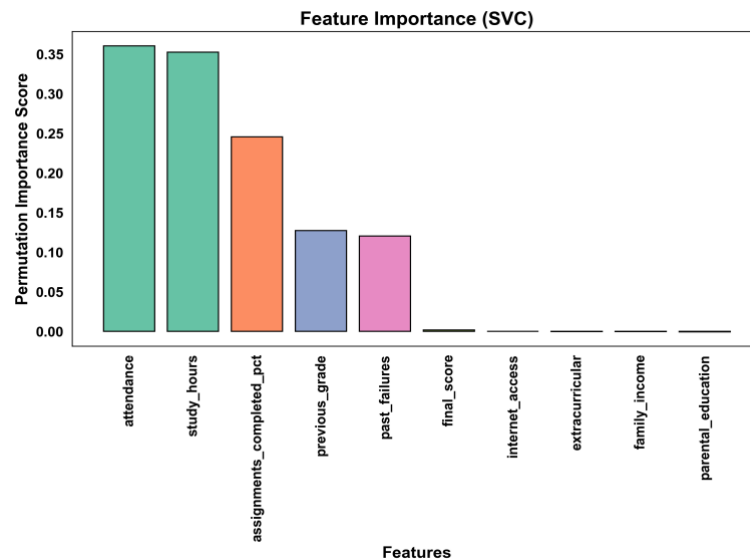


Figure 6. Permutation Feature Importance of the Support Vector Classifier

4.6. Comparison with Other Studies

Table 1 provided in this section places our research within other studies that have utilized ML to forecast student academic risk and performance. Because the other studies used different datasets and defined their target variables and evaluations differently, the case accuracies reported should be interpreted liberally. This study explains academic risk as a three-class problem (at risk, high risk, and safe) and uses different ML classifiers to build a risk classification model. In this regard, the SVC showed the best performance, with an accuracy of 93.63%.

Table 1. Comparative Analysis of the Present Study with Existing Machine Learning Studies on Student Academic Risk and Performance Prediction

Study	Dataset	Task	Main Model	Reported Accuracy
Chui et al. [17]	University students, 32,593 records	At-risk student prediction	RTV-SVM	92.2–93.8%
KSA study [16]	Saudi university students	Academic performance and dropout prediction	Majority Voting	92.7%
Mindanao State University study [13]	Higher-education students, 482 records	Student retention prediction	XGBoost	90.66%
College Statistics study [18]	College statistics students	Identification of at-risk students	Random Forest	91%
Present Study	Student Academic Risk Dataset, 100,000 records	Three-class risk classification	SVC	93.63%

5. CONCLUSION

The goal of this study was to develop a machine learning framework to classify students as At-Risk, High-Risk, or Safe based on various behavioral and psychological factors, as well as academic data. Eight classifiers were trained and tested using multiple metrics within the same parameter limits. Out of all the models assessed, Support Vector Classifier (SVC) showed the strongest performance, with an accuracy of 93.63%, and corresponding values of precision, recall, and F1 score of 93.64, 93.63, and 93.63, respectively. The assessed models showed variations in their predictive behaviors and in the ways they generalized to unseen data. According to variable importance analyses, predictors of academic risk included, but were not limited to, attendance, study hours, completion of assignments, past grades, and past academic failures. In summary, results revealed that the use of ML techniques to classify academic risk is viable and supports the case for using data to find learners who may benefit from proactive support and intervening before it may be too late. For this research, we will be able to assess the generalizability of our ML framework for academic risk by using extended, real-world, and longitudinal student data from numerous institutions. To further enhance the ML academic risk framework and allow for its broader use, we will integrate Explainable AI and ensure ML models used embody standards of fairness and equality.

DATA AVAILABILITY STATEMENT

The original data presented in the study are openly available in Kaggle "Student Academic Risk Prediction Dataset", at doi: <https://www.kaggle.com/datasets/arpitsharmaml/student-academic-risk-prediction-dataset?resource=download>

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

- [1] B. B. Alkan, S. Kuzucuk, N. Alkan, and A. Sinan, "Using machine learning to predict student outcomes for early intervention and formative assessment," *Sci Rep*, vol. 15, no. 1, p. 39797, Nov. 2025, doi: [10.1038/s41598-025-23409-w](https://doi.org/10.1038/s41598-025-23409-w).
- [2] J. Pan, Z. Zhao, and D. Han, "Academic Performance Prediction Using Machine Learning Approaches: A Survey," *IEEE Trans Learn Technol*, vol. 18, pp. 351–368, 2025, doi: [10.1109/TLT.2025.3554174](https://doi.org/10.1109/TLT.2025.3554174).
- [3] R. Guevara-Reyes, I. Ortiz-Garcés, R. Andrade, F. Cox-Riquetti, and W. Villegas-Ch, "Machine learning models for academic performance prediction: interpretability and application in educational decision-making," *Front Educ*, vol. 10, Aug. 2025, doi: [10.3389/educ.2025.1632315](https://doi.org/10.3389/educ.2025.1632315).
- [4] E. Guamán-Hidalgo and L. Enciso, "Early Prediction of Student Dropout Through Machine Learning in Educational Data Mining: A Comparative Analysis of Algorithms and Data Balancing Techniques," *Appl Sci*, vol. 16, no. 18, p. 8915, Sep. 2026, doi: [10.3390/app16188915](https://doi.org/10.3390/app16188915).
- [5] A. R. Ladia and S. S. Caluya, "Assessing the Influence of Video Games on Academic Success Through Machine Learning Techniques," in *2026 International Conference on Electronics, Communications and Computer Science (ICECCS)*, IEEE, Jul. 2026, pp. 1–5. doi: [10.1109/ICECCS70432.2026.11651272](https://doi.org/10.1109/ICECCS70432.2026.11651272).
- [6] S. El Kafhali and Z. S. Hafdi, "Enhancing Early Academic Outcome Prediction in Small Educational Datasets Through Data Augmentation Techniques," *Data*, vol. 11, no. 7, p. 161, Jul. 2026, doi: [10.3390/data11070161](https://doi.org/10.3390/data11070161).
- [7] Y. S. Mulyawan and M. I. C. Rachmatullah, "Comparative Analysis of Machine Learning Models for Student Performance Prediction with SHAP- Based Explainability Approach," in *2026 International Conference on*

- Smart Computing, IoT, and Machine Learning (SIML)*, IEEE, Jun. 2026, pp. 1–6. doi: [10.1109/SIML69834.2026.11621106](https://doi.org/10.1109/SIML69834.2026.11621106).
- [8] A. H. Altalhi and M. Ragab, "Modelling of enhanced predictive analytics and decision-making approach using recurrent autoencoder with attribute subset reduction for students' academic performance," *Ain Shams Eng J*, vol. 17, no. 3, p. 103999, Mar. 2026, doi: [10.1016/j.asej.2026.103999](https://doi.org/10.1016/j.asej.2026.103999).
- [9] A. F. A. Sayed, M. A. Arafa, N. A. El-Nimr, K. A. S. Banawan, and M. S. Abdou, "Comparison of machine learning classification and regression models for prediction of academic performance among postgraduate public health students," *Sci Rep*, vol. 15, no. 1, p. 44056, Dec. 2025, doi: [10.1038/s41598-025-31023-z](https://doi.org/10.1038/s41598-025-31023-z).
- [10] M. A. U. Alif, M. Ahmed, M. Y. Ali, and R. Shamsi, "A Comparative Analysis of Machine Learning Algorithms for Predicting Depression in Bangladeshi University Students," in *2025 IEEE 4th International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*, IEEE, Nov. 2025, pp. 760–765. doi: [10.1109/RAAICON69033.2025.11502398](https://doi.org/10.1109/RAAICON69033.2025.11502398).
- [11] Y. Lu, S. Yeom, J. Maktoubian, M. M. Rahman, and S.-H. Kim, "Improve Student Risk Prediction with Clustering Techniques: A Systematic Review in Education Data Mining," *Educ Sci*, vol. 15, no. 12, p. 1695, Dec. 2025, doi: [10.3390/educsci15121695](https://doi.org/10.3390/educsci15121695).
- [12] A. Tolkynebekova, D. Koishiyeva, A. Bissembayev, D. Mukhammejanova, A. Mukasheva, and J. W. Kang, "Comparative Analysis of the Predictive Risk Assessment Modeling Technique Using Artificial Intelligence," *J Electr Eng Technol*, vol. 20, no. 6, pp. 4509–4526, Sep. 2025, doi: [10.1007/s42835-025-02269-4](https://doi.org/10.1007/s42835-025-02269-4).
- [13] R. D. Deleña *et al.*, "Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors," *Syst Soft Comput*, vol. 7, p. 200352, Dec. 2025, doi: [10.1016/j.sasc.2025.200352](https://doi.org/10.1016/j.sasc.2025.200352).
- [14] A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discov Artif Intell*, vol. 4, no. 1, p. 2, Jan. 2024, doi: [10.1007/s44163-023-00079-z](https://doi.org/10.1007/s44163-023-00079-z).
- [15] M. M. P. Cruz and R. G. Lumauag, "Comparative Analysis of Machine Learning Algorithms for Predicting Student Academic Performance in Higher Education," in *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, IEEE, Dec. 2024, pp. 888–896. doi: [10.1109/ICUIS64676.2024.10866086](https://doi.org/10.1109/ICUIS64676.2024.10866086).
- [16] M. Ben Said, Y. Hadj Kacem, A. Algarni, and A. Masmoudi, "Early prediction of Student academic performance based on Machine Learning algorithms: A case study of bachelor's degree students in KSA," *Educ Inf Technol*, vol. 29, no. 11, pp. 13247–13270, Aug. 2024, doi: [10.1007/s10639-023-12370-8](https://doi.org/10.1007/s10639-023-12370-8).
- [17] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, "Predicting at-risk university students in a virtual learning environment via a machine learning algorithm," *Comput Human Behav*, vol. 107, p. 105584, Jun. 2020, doi: [10.1016/j.chb.2018.06.032](https://doi.org/10.1016/j.chb.2018.06.032).
- [18] I. Wickramasinghe, "Enhancing Student Success: Early Prediction of Struggles in College Statistics," *Scatterplot*, vol. 2, no. 1, Dec. 2025, doi: [10.1080/29932955.2025.2507457](https://doi.org/10.1080/29932955.2025.2507457).

BIOGRAPHIES OF AUTHORS



Natesh Mahadev Completed his PhD in computer science and engineering in Visveswaraya technological university (VTU), MTech from Visveswaraya Technological University (VTU), B.E (CSE) from VTU, currently he is working as professor in the department of Computer science and engineering in Vidya vardhaka college of engineering, Mysore, Karnataka, India. His research interests include Digital image processing, Blockchain, Machine learning, Artificial Intelligence, Data science. He has published a various paper in reputed National and international journals. He can be contacted at email: natesh.m@vvce.ac.in



Sowmya V. L. is an Assistant Professor in the Department of Artificial Intelligence and Machine Learning at BMS Institute of Technology and Management, Bengaluru, India. She received her Ph.D. in Computer Science and Engineering from Visveswaraya Technological University, Karnataka, India. Her doctoral research focused on radiogenomics-based approaches in neuro-oncology, integrating radiomics, genomics, and machine learning for computational biomarker discovery and disease characterization. Her research interests include artificial intelligence in healthcare, medical image analysis, radiomics, genomics, bioinformatics, multimodal data integration, and explainable machine learning. She is actively involved in interdisciplinary research aimed at developing computational methodologies for medical diagnosis, disease characterization, and precision medicine. She can be contacted at email: soumyavl@bmsit.in



Anitha Premkumar is currently serving as an Assistant Professor (Senior Scale) at the Manipal Institute of Technology Bengaluru. She earned her Ph.D. from the Vellore Institute of Technology, Vellore, Tamil Nadu, India. She has an extensive publication record, with numerous research papers presented at international journals and conferences. Her primary research interests include cloud computing, blockchain, and machine learning. She can be contacted at email: p.anitha@manipal.edu



N. Thangarasu is currently working as an assistant professor in the department of Computer Science, at Karpagam Academy of Higher Education, Coimbatore. I am greatly fascinated with the advanced computing technology and research programs is Cluster Computing, Cryptography and Network Security, Cloud Computing, Artificial Intelligent System, Information Security in large Database and Data Mining as well as the strong teaching experience. My doctoral dissertation also focuses on advanced security systems with cloud computing, and I have published more than 80 publications in reputed journals, which I find would be a great addition to the success of your teaching and research department. He can be contacted at email: drthangarasu.n@kahedu.edu.in