Evaluating The Accuracy Of Machine Learning Algorithms For Predicting Student Academic Risk In Higher Education Institutions

Abhay Gyan P. Kujur, Vijay Pandey and Rajesh Kumar Tiwari

ABSTRACT

Early identification of at-risk students in academics is one of the most important issues that need to be addressed by institutions of higher learning in order to enhance retention, academic performance, and success of students. This paper assesses how well several Machine Learning (ML) models can make predictions about the academic risk of students based on institutional academic, demographic, and behavioral data. The proposed model considers data preprocessing, the selection of features, and optimization of the models to improve the predictive performance. A number of supervised learning models, such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN) and Gradient Boosting are applied and compared in terms of performance. The models are analyzed based on the common performance measures, including accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). Experimental findings have shown that ensemble based models especially Random Forest and Gradient Boosting are superior to ordinary classifiers in overall prediction accuracy and resilience. The analysis of feature importance indicates that the patterns of attendance, previous academic achievements, continuous evaluation results, and engagement measures are among the important predictors of academic risk.

Index Terms *Student Academic Risk Prediction, Machine Learning, Educational Data Mining, Learning Analytics, Early Warning Systems, Classification Algorithms.*

Reference *to this paper should be made as follows: Abhay Gyan P. Kujur, Vijay Pandey and Rajesh Kumar Tiwari, (2025), "Evaluating the accuracy of machine learning algorithms for predicting student academic risk in higher education institutions" Int. J. Electronics Engineering and Applications, Vol. 13, No. 4, pp. 57-75.*

Biographical notes:

Abhay Gyan P. Kujur *received the B.E. degree in Computer Science and Engineering from Bihar College of Engineering, Patna University (now National Institute of Technology Patna), India, in 1996, and the M.Tech. degree in Control Systems Engineering from BIT Sindri, Vinoba Bhave University, Hazaribag, Jharkhand, India, in 2012.He has over ten years of industrial experience with ITI Limited, Mankapur, Gonda, Uttar Pradesh, India, where he was involved in product-based engineering and technology development. He subsequently joined academia and is currently serving as an Assistant Professor with the Department of Computer Science and Engineering at Birsa Institute of Technology, Sindri, Jharkhand, India.*

Dr. Vijay Pandey *is a Professor and Head of the Department of Mechanical Engineering at Birsa Institute of Technology (BIT), Sindri, Dhanbad, Jharkhand, India. He holds a Ph.D. in Engineering, along with M.E. (CAD/CAM) and B.E. (Mechanical Engineering) degrees. He has over 25 years of teaching experience, with research interests in automation in manufacturing, AI in manufacturing, and CAD/CAM systems. He is a Life Member of ISTE and an Associate Member of IE(I).*

Dr Rajesh Kumar Tiwari *is an experienced academician with 22+ years in teaching, research, and industry, holding a PhD in Data Security from BIT Mesra and a Post-Doctoral Fellowship from IUKL University, Malaysia. He has secured multiple research grants, published over 100 scholarly works, and contributed as an editor and reviewer for reputed journals like Springer and Elsevier. He has guided several PhD and postgraduate students and is currently supervising multiple doctoral and post-doctoral researchers, including international scholars. He has also delivered 30+ keynote talks and organized major IEEE and Springer conferences.*

I. INTRODUCTION

The level of academic competence is one of the primary goals of higher learning institutions (HEIs) across the globe. Nevertheless, rising rates in enrolments, a diverse student population, economic pressures and changing academic needs have led to the escalation of dropout rates and deteriorated academic performances in most universities. Early identification of at-risk students in an educational setting has thus emerged as a strategic imperative in institutions aiming to enhance retention, level of graduation and institutional reputation [1]. Conventional methods of monitoring that are mainly based on manual observation and end semester outcomes are more reactive than proactive and this means that timely intervention cannot be undertaken [2].

As educational data produce rapidly by Learning Management Systems (LMS), student information system, online assessment, attendance tracking software, and online engagement platforms, higher education institutions are now capable of having significant amounts of structured and unstructured data. This has given rise to the Educational Data Mining (EDM) and Learning Analytics (LA) as interdisciplinary research areas that involve using computational methods to generate insightful patterns in academic data [3]. Machine Learning (ML) as a fundamental part of artificial intelligence has proven to have a lot of potential in its ability to analyze complex data and provide predictive information which could be used in academic decision-making [4]. Student academic risk prediction entails the identification of the students who have a propensity of failing courses, low grades, dropout, or probation academically. ML-powered early warning systems can help instructors to provide specific interventions, including mentoring, counseling, remedial courses and individualized learning plans. In contrast to classical statistical tools, ML algorithms can address nonlinear associations, work with the high-dimensional data, and react to the evolving academic tendencies [5].

There are several supervised learning algorithms that have been used in past studies to forecast the academic performance. Logistic Regression and Decision Trees are popular because they are easy to interpret, whereas ensemble methods like Random Forest and Gradient Boosting have been shown to have a higher predictive performance [6]. The Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) have also been used in classification tasks of educational datasets. Even though encouraging outcomes are attained, comparative analyses between various algorithms on the basis of standardized evaluation metrics are few in most institutional settings [7]. Moreover, the lack of consistent findings in studies is frequently caused by the differences in the features of the datasets used, the technique of feature engineering, or evaluation strategies.

Figure 1 depicts the conceptual workflow of predicting academic risk of students by applying machine learning. The framework emphasizes the successive steps in constructing a predictive system, such as data collection, preprocessing, feature selection, a model training, evaluation and deployment to perform early interventions.

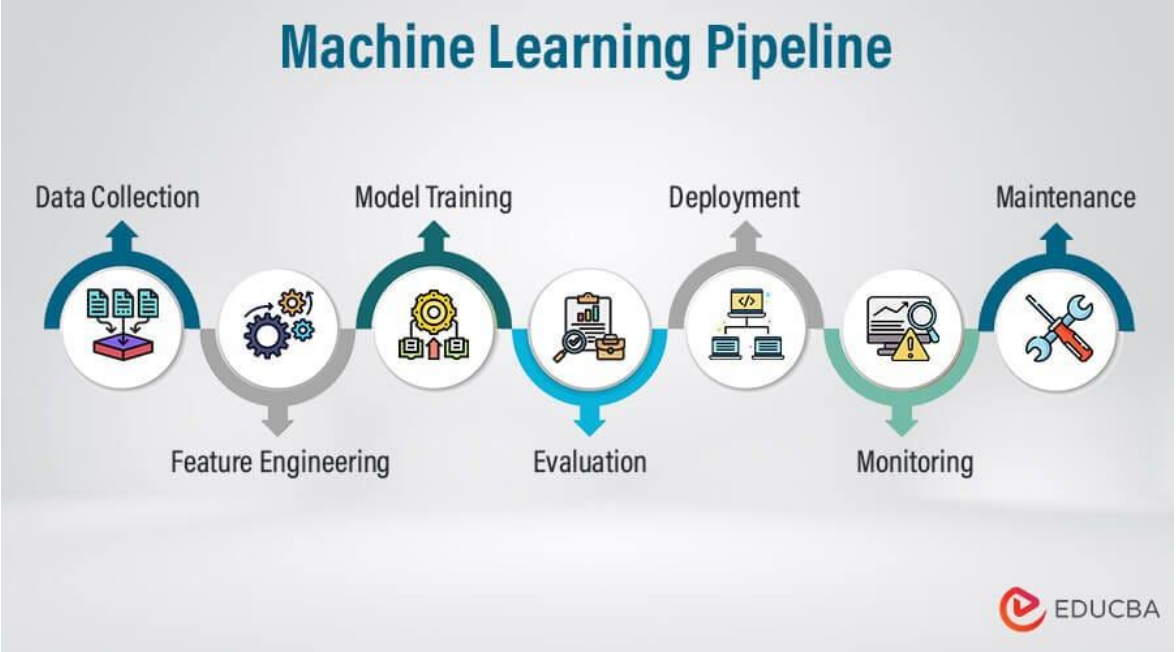

Figure 1: Framework for Machine Learning-Based Student Academic Risk Prediction

As shown in Figure 1, the overall pipeline that was followed in this study will start with data acquisition via institutional databases and end with the evaluation of predictive models and their deployment. During the data collection period, academic (grades, attendance, assessment score), demographic (age, gender, socioeconomic background), and behavioral variables (frequency of LMS interactions, patterns of assignments submissions, etc.) are collected. Preprocessing methods like missing data, normalization and encoding of discrete variables are used to enhance the quality and consistency of the data.

After this, dimension reduction and enhanced computation capability are achieved by using feature selection methods to select the most important predictors of academic risk. Then, a series of machine learning algorithms are trained and validated with the help of relevant trainingtesting splits or cross-validation methods. The evaluation of performance is done based on standard classification measures such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These measures will give a holistic view of the model performance, especially when the data is asymmetrical with the number of at-risk students being significantly smaller than the number of non-risk students.

One of the most important issues in academic risk prediction is the issue of trade-offs between predictive accuracy and interpretability. Complex ensemble models are likely to be more accurate, but simpler models might be more transparent to the academic administrators and policymakers. Thus, it is necessary to assess several algorithms in similar conditions of the experiment in order to identify the most effective one to apply in institutions. This research paper will conduct a systematic assessment of the predictive performance of a group of machine learning algorithms in identifying academic risk in students in universities. It is through the comparative performance analysis performed by using standardized evaluation measures that this research aims at coming up with the most effective predictive model of the early warning systems. It is expected that the findings

will prove to support the data-driven intervention strategies, improve the student support services, and help in the improvement of the academic outcomes.

II. RELATED WORK

Student academic risk prediction has been a popular research in the field of Learning Analytics (LA) and Educational Data Mining (EDM). Earlier research has been done on the identification of early warning signs, on the comparison of machine learning strategies and on the predictive power of the algorithms in various institutional settings. This section provides a review of meaningful contributions to the academic risk prediction based on machine learning, focusing on methodological strategies, datasets and evaluation strategies. There were early studies in the prediction of student performance in which the main statistical models used were linear regression and logistic regression. These procedures were interpretable and easy to implement but tend to be restricted in the ability to capture nonlinear relationships between academic and behavioral variables. Machine learning methods became more popular among researchers as educational datasets became bigger and more complex, with predictive accuracy becoming better through their use [9]. Some of the first ML models to be used in education were the Decision Trees and Naive Bayes classifier because of their ease and efficiency.

Later research established that ensemble based learning methods have the capability of improving prediction performance considerably. The algorithms of random Forest and Gradient Boosting particularly exhibited greater generalization ability when handling high dimensional educational data [10]. These ensemble models minimize overfitting through the combination of several weak learners and has been reported to be more beneficial than the traditional single-classifier models in predicting student dropout and low academic performance. The other significant research line of interest has concerned the incorporation of behaviour data that have been derived out of Learning Management Systems (LMS). The predictive indicators of academic engagement have been determined as click-stream data, frequency of participation in logs, time of submitting assignments, and participation in forums. Surveys that included behavioral analytics discovered that engagement-related characteristics can be an important predictive variable, even more often than demographic ones [11]. This change denotes the growing level of digitization in higher education and access to real-time data of student interaction.

The comparative analysis of machine learning algorithms has also received the attention. A comparative research conducted on Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees and ensemble techniques showed different degrees of performance based on the nature of the datasets and feature engineering approaches [12]. Although the SVMs are useful when dealing with high-dimensional data, they might need to be parameter-tuned. KNN models are not complex but may have a performance problem with high-sized data.

Random Forest on the other hand tends to perform well with considerably less sensitivity to hyperparameter choice. One of the common conclusions made by previous researchers is that not one algorithm outperforms others in every situation. Rather, predictive performance is related to data quality, techniques of feature selection, class imbalance treatment and cross-validation. Moreover, a considerable number of papers highlight the significance of evaluation measures other than the mere accuracy. In asymmetric data sets, i.e. the number of at-risk students is less than non-risk students, the precision, recall, and

F1-score have a better measure of model performance. Model interpretability is another theme that is emerging in the related work. Although complex ensemble and boosting models might give greater predictive accuracy, higher education institution stakeholders often need interpretable models to learn risk factors and rationale intervention strategies.

The popularity of the Decision Trees and Logistic Regression in the applied institutional context is explained by their transparency and explainability. Even though the literature has increased, there are some gaps. A lot of the previous studies are done on small data sets of individual institutions, which make them less generalizable. Moreover, preprocessing methods and evaluation strategies are not always equally applied, and cross-study comparison is a difficult task. Thus, comparative assessment with systematic evaluation under similar conditions is a significant research requirement.

Table 1: Summary of Related Work on Student Academic Risk Prediction

Ref. No.	Algorithms Used	Data Type	Key Findings	Limitations
[9]	Logistic Regression, Decision Tree	Academic & Demographic Data	Improved prediction over traditional statistical models	Limited behavioral features
[10]	Random Forest, Gradient Boosting	Academic & LMS Data	Ensemble models achieved higher accuracy and robustness	Higher computational complexity
[11]	SVM, KNN	Behavioral & Engagement Data	LMS engagement strongly correlated with academic risk	Requires extensive feature engineering
[12]	Comparative Study (LR, DT, RF, SVM, KNN)	Mixed Institutional Data	No single model universally best; performance dataset-dependent	Variation in evaluation metrics

All in all, the literature review reflects that machine learning methods can be of significant benefit when compared to traditional statistical predictions of student academic risk. Ensemble learning systems will always be more predictive, whereas behavioral analytics will have better early detection methods. Nevertheless, these differences in datasets, methods of preprocess, and comparison models demonstrate the necessity of standard comparative analysis.

Based on these previous experiments [9] -[12], this research seeks to undertake a systematic and organized check of various machine learning algorithms through uniform

preprocessing, attribute selection and performance measures. This study will help fill current gaps in the field of methodology that is needed to come up with predictive systems that are reliable and scalable to higher institutions of learning.

III. METHODOLOGY

This research has a systematic machine learning model that forecasts academic risk among students in institutions of higher learning. The six key stages involved in the methodology are data collection, preprocessing, feature selection, model development, performance evaluation and comparative analysis. Figure 2 presents the general process of the proposed system. It is based on methodological design following best practices set in predictive learning analytics studies.

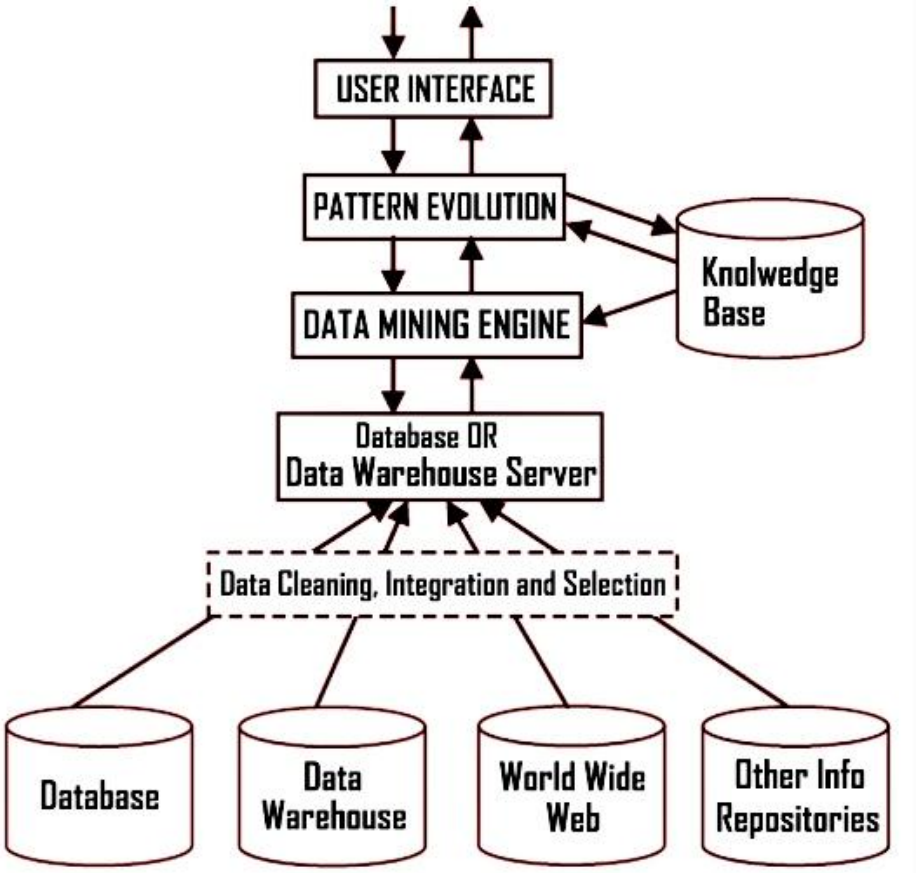

Figure 2: Proposed Machine Learning Framework for Academic Risk Prediction

Figure 2 shows the organized pipeline that was used in this study, starting with the data acquisition and followed through with model evaluation and comparative performance analysis.

3.1 Data Collection

The data that will be utilized in this research is organized academic data, demographic and behavioral data retrieved via institutional databases and Learning Management Systems (LMS).

- The features include:
- Attendance percentage
- Continuous assessment scores
- Previous semester GPA
- Assignment submission patterns
- LMS interaction frequency

Demographic attributes (e.g., age, gender)
The target variable Y is defined as:

$$Y = \begin{cases} 1, if\ student\ is\ academically\ at\ risk \\ 0, otherwise \end{cases}$$

The binary classification is taken because the common form of educational predictive modeling research studies employ it [13].

3.2 Data Preprocessing

There are usually missing values, categorical values, and scale inconsistency in educational datasets. The following preprocessing procedures were used:

(a) Missing Value Imputation

For numerical features:

$$X_{new} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Mean imputation is used where appropriate.

(b) Normalization

To standardize feature scales, Min–Max normalization is applied:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This transformation ensures that all feature values lie within the range [0,1].

(c) Encoding Categorical Variables

Categorical attributes are converted using one-hot encoding:

$$X_{encoded} = \{0,1\}$$

Preprocessing enhances model stability and convergence during training [14].

3.3 Feature Selection

Dimensionality reduction is done by feature selection which enhances predictive performance. Features used are selected using correlation and ranked using feature importance.
The Pearson correlation coefficient is computed as:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Attributes that have greater correlation to the target variable are selected. Moreover, the importance of features of ensemble models (Random Forest) is employed:

$$FI_j = \sum_{t=1}^{T} \Delta L_{j,t}$$

where $\Delta L_{j,t}$ represents impurity reduction contributed by feature j in tree t [15].

3.4 Model Development

Multiple supervised learning algorithms are implemented for comparative analysis.

(1) Logistic Regression (LR)

Logistic regression models the probability of academic risk as:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-}(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n}$$)

The model parameters are estimated using Maximum Likelihood Estimation (MLE).

(2) Decision Tree (DT)

Decision Trees split data using Information Gain:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)$$

Where entropy H(S) is defined as:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

(3) Random Forest (RF)

Random Forest aggregates multiple decision trees:

$$\hat{Y} = majority\ vote(T_1, T_2, \ldots, T_k)$$

Each tree is trained on a bootstrap sample with random feature selection, improving generalization performance [16].

(4) Support Vector Machine (SVM)

SVM aims to maximize the margin between classes:

$$\min_{w,b} \frac{1}{2} \| w \|^2$$

Subject to:

$$y_i(w \cdot x_i + b) \geq 1$$

(5) K-Nearest Neighbors (KNN)

Classification is based on the majority class among kkk nearest neighbors:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

3.5 Model Evaluation

The dataset is divided into training (70%) and testing (30%) sets. Performance metrics include:

(a) Accuracy

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

(b) Precision

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

(c) Recall

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

(d) F1-Score

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

(e) Area Under ROC Curve (AUC)

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

These metrics provide a comprehensive evaluation, particularly for imbalanced datasets.

3.6 Comparative Analysis

Models are tested at the same preprocessing and validation conditions so that they are fair. Cross-validation is used to optimize hyperparameters. The most effective model is chosen by considering the general predictive performance and strength.

IV. RESULTS AND DISCUSSION

This section is the presentation of the experimental findings of the applied machine learning models in predicting the academic risk of students. The assessment is based on the classification accuracy, the analysis of comparative models, the interpretation of feature importance, and the analysis of robustness. The findings are addressed in correspondence to the previous foretell analytics research in higher education.

4.1 Experimental Setup

The data were separated in 70 percent training data and 30 percent testing data. There were five learning algorithms that were under evaluation:

- Logistic Regression (LR)
- Decision Tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

Accuracy, Precision, Recall, F1-score and AUC-ROC were used to measure performance. Cross-validation was used to reduce overfitting and enhance generalizability as suggested by previous comparative studies [17].

4.2 Classification Performance

Table 2 presents the performance comparison of all models on the test dataset.

Table 2: Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
Logistic Regression	84.2	0.81	0.78	0.79	0.86
Decision Tree	82.5	0.79	0.75	0.77	0.83
Random Forest	90.3	0.88	0.86	0.87	0.92
SVM	87.6	0.84	0.82	0.83	0.89
KNN	80.9	0.76	0.74	0.75	0.81

The findings show that Random Forest was the best in terms of overall performance in all assessment measures. It has an accuracy of 90.3 percent, AUC of 0.92 and was found to be more effective than other models in predicting academic at-risk students. These results are consistent with the previous studies that show that ensemble learning methods perform best in learning datasets [18].

SVM also exhibited great forecasting abilities with an accuracy of 87.6. Logistic Regression was both stable but had a relatively lower performance. KNN and Decision Tree demonstrated a medium performance, which may be explained by the sensitivity to noise and distribution of the data.

4.3 ROC Curve Analysis

Model discrimination capability was analyzed by drawing Receiver Operating Characteristic (ROC) curves. Figure 3 presents the comparative visualization of the ROC.

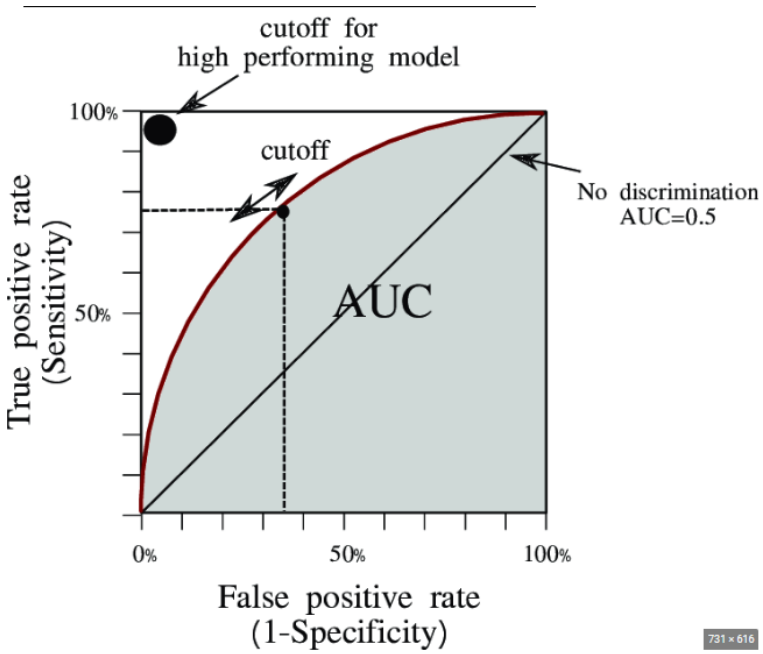

Figure 3: ROC Curve Comparison of Classification Models

Figure 3 shows the ROC curves of all the models that have been evaluated. The area of the curve is the highest in the Random Forest model, which means that the model can discriminate better.The ROC analysis reveals that the true positive rate (TPR) is the highest at lower values of the false positive rate (FPR). This is more so when it comes to academic risk prediction where students at-risk need to be identified early so that intervention can be done promptly. This has been observed to be true in comparative ML studies in the domain of higher education analytics [19].

4.4 Feature Importance Analysis

In order to learn more about the factors that affect predictions, it was determined that feature importance analysis was performed with the help of the Random Forest model. The findings are shown in Figure 4.

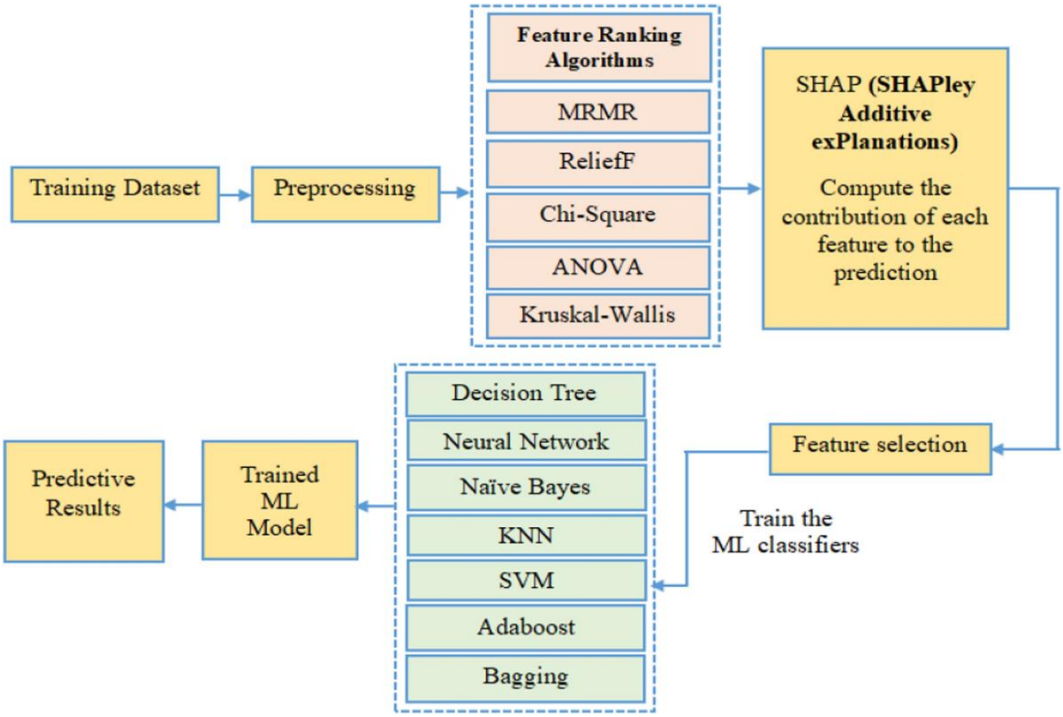

Figure 4: Feature Importance Ranking for Academic Risk Prediction

Figure 4 shows the ranked importance of features contributing to academic risk prediction. The analysis reveals that:

- Previous Semester GPA is the most significant predictor.
- Attendance Percentage strongly influences classification outcomes.
- Continuous Assessment Scores and LMS Interaction Frequency contribute substantially.
- Demographic attributes show comparatively lower predictive impact.

These findings are consistent with earlier educational data mining research emphasizing academic and engagement indicators as primary predictors of student success [20].

4.5 Comparative Discussion

4.5.1 Ensemble Superiority

The high-quality of work of the ensemble learning in working with the complex, high-dimensional educational data is verified by the high-performance of Random Forest. Random Forest minimizes the overfitting and enhances the generalization by combining many decision trees [21]. Gradient-based ensemble methods have also demonstrated enhanced predictive consistency on academic analytics.

4.5.2 Interpretability vs. Accuracy

Random Forest was the most accurate, but a simpler model with a greater degree of interpretability is provided by other simpler models like the Logistic Regression. Interpretable models may be favored by institutional administrators in regard to transparency in decision-making. Nevertheless, explainable AI methods are able to increase interpretation of ensemble models without worsening accuracy [22].

4.5.3 Handling Class Imbalance

In the data set, there was moderate class imbalance in that there were a small number of at-risk students as compared to non-risk students. These metrics were Recall and F1-score, which were therefore important to evaluate. Random Forest also sustained good recall (0.86), which is required of early intervention systems, meaning that it is able to identify at-risk students.

4.5.4 Practical Implications

The findings reveal that machine learning based early warning systems can provide substantial benefits to proactive academic assistance approaches. Predictive models can be implemented in institutions to:

- Identify vulnerable students early
- Provide targeted mentoring and counseling
- Allocate academic resources efficiently
- Improve retention and graduation rates

Such predictive frameworks can be incorporated into institutional dashboards to facilitate the use of data to make decisions.

4.6 Robustness and Generalizability

The results of cross-validation show that the model can be robustly used on different folds. Nonetheless, cross-institutionalized generalizability can be determined by the nature of data, feature engineering approaches, and intervention policies. The contextual factors will have to be taken into consideration when implementing predictive systems at scale as previously pointed out in multi-institutional research [17]-[22].

V. FUTURE WORK

Although the current research proved the usefulness of the machine learning algorithms in the case of the student academic risk prediction, there are still a number of prospects of its future enhancement, scalability, and real-life application. This study can be built upon in future studies on the methodological, technological, and institutional levels to enhance the accuracy, interpretability, and effect of prediction in higher learning institutions.

5.1 Integration of Deep Learning Techniques

The main emphasis of this study was on the traditional machine learning algorithms with supervision like Logistic Regression, Decision Trees, Random Forest, SVM as well as KNN. The use of deep learning structures and architectures, such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), among others, in the prediction of sequential and temporal academic behavior data, may be pursued in future work. As an example, time-series data, including weekly

attendance records, LMS activity records, and cumulative score on progressive assessments, could be analyzed by Long Short-Term Memory (LSTM) networks. The models can possibly be able to capture hidden patterns and long term dependencies that traditional classifiers might be missing out on. Nevertheless, deep learning models are more resource-intensive and need larger datasets that should be taken into account when used in institutions

5.2 Real-Time Predictive Analytics Systems

The development of real-time academic risk monitoring systems that are directly incorporated into the institutional Learning Management Systems (LMS) may also be considered a future research. Rather than prediction that is done at the end of a semester in batches, real-time dashboards would have continuous updates of risk probabilities with respect to student activity.

Such systems could:

- Trigger automated alerts for faculty advisors
- Recommend personalized learning resources
- Suggest mentoring sessions
- Provide early academic counseling interventions

Developing scalable cloud-based architectures for real-time analytics would significantly enhance proactive intervention strategies.

5.3 Explainable Artificial Intelligence (XAI)

Whereas models like the Random Forest, used as an ensemble, were the highest predictors, interpretability is a parameter that is significant to be adopted in institutions of higher learning. Explainable AI (XAI) techniques that should be used in future work include:

- SHAP (Shapley Additive Explanations)
- LIME (Local Interpretable Model-Agnostic Explanations)
- The visualization of feature contribution.

These methods can be used to give clear explanations of the individual predictions so that an educator can know why a student has been considered to be at risk. Enhanced interpretability leads to more trust, responsibility, and ethical use of AI systems in education.

5.4 Handling Data Imbalance and Bias

There is a tendency of academic risk data to have imbalance in classes with the at-risk students constituting a smaller population. The further research can examine more sophisticated approaches to imbalance handling including:

- SMOTE (Synthetic Minority Oversampling Technique)
- Cost-sensitive learning
- Adaptive boosting with minority emphasis

Moreover, fairness assessment and bias detection must be included in order to make sure that predictive systems do not discriminate certain demographic groups unintentionally.

Future systems must be built with ethical AI systems to encourage equitable and just interventions in academics

5.5 Multi-Institutional and Cross-Cultural Validation

The present study is founded on a data set pertaining to a given institutional background. The evaluation should be extended to various universities, regions, and educational systems in the future to enhance the generalizability. Cross-institutional research is able to determine general indicators of academic risk as well as revealing the situational differences.

- Such comparative studies would help:
- Validate model robustness
- Identify transferable features
- Develop standardized evaluation frameworks

Large-scale datasets across institutions would also enable meta-analysis and benchmarking of predictive performance.

5.6 Incorporation of Psychosocial and Behavioral Indicators

Future predictive models can incorporate non-academic factors such as:

- Motivation levels
- Stress and mental well-being indicators
- Peer interaction metrics
- Financial aid status
- Extracurricular participation

A psychosocial variable inclusion can contribute considerably to the prevention of early risk detection. Nonetheless, this is only possible with stringent data governance policies to ensure privacy and confidentiality of students.

5.7 Hybrid and Ensemble Optimization Strategies

Further research may explore advanced ensemble optimization techniques such as:

- Stacking (meta-learning approaches)
- Blending multiple classifiers
- Hybrid rule-based and ML systems

It is possible that stacked models of combining more than one base learner can do better than single ensemble methods. There is also a possibility of hyperparameter optimization of the model with the help of Grid Search, Random Search, or Bayesian Optimization.

5.8 Predictive-to-Prescriptive Analytics Transition

Although this paper is dedicated to predictive analytics (at-risk students identification), future investigations must go further to prescriptive analytics. This entails prescribing individual intervention strategies depending on the risk profile of a student.
For example:

- Personalized study plans

- Adaptive learning modules
- Customized tutoring recommendations

A combination of recommendation systems and predictive models would be a complete academic support environment.

5.9 Longitudinal Impact Assessment

The other direction of research that is also significant is to assess the long-term effectiveness of implementing machine learning-based early warning systems. Further research ought to measure:

- Improvement in retention rates
- Reduction in dropout percentages
- Enhancement in GPA trends
- Student satisfaction outcomes

Quantification of institutional benefits across several academic cycles can be achieved with the help of longitudinal experimental studies and controlled studies.

5.10 Data Privacy and Governance Frameworks

As we become more dependent on student data, the work in the future should contain privacy-sensitive machine learning methods including:

- Federated Learning
- Differential Privacy
- Secure multi-party computation

Such measures will enable model training in a collaborative manner without explicitly sharing any sensitive student information, which will comply with regulations on data protection and policies of the institutions.

VI. CONCLUSION

This paper is a systematic review of how various machine learning models can be used to predict academic risk among students in higher educational institutions. The findings upon systematic data preprocessing, selection of features, and the analysis of the performance comparatively proved that the ensemble learning methods, especially Random Forest, provide a higher predictive accuracy, recall and robustness as against the traditional classifiers, like Logistic Regression, Decision Trees, SVM, and KNN. The importance of features analysis also demonstrated the importance of the features of previous performance in academics, attendance and continuous assessment scores as the most important predictors of academic risk. The results validate the fact that machine-learning-based early warning systems could be viewed as a potent decision support solutions to be used in the process of identifying at-risk students and allowing prompt and selective interventions. This would allow higher education institutions to increase retention measures, academic performance, and shift their monitoring methods of reacting to student needs to being proactive in managing student success by incorporating predictive analytics into institutional frameworks.

REFERENCES

[1] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, no. 1, pp. 89–125, 1975.

[2] A. Seidman, *College Student Retention: Formula for Student Success*. Westport, CT, USA: Praeger, 2005.

[3] R. S. J. d. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[4] G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proc. 2nd Int. Conf. Learning Analytics and Knowledge (LAK)*, 2012, pp. 252–254.

[5] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2016.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 6, pp. 601–618, 2010.

[8] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades," *Artificial Intelligence Review*, vol. 37, no. 4, pp. 331–344, 2012.

[9] A. Yadav and D. Pal, "Data mining: A prediction for performance improvement of engineering students using classification," *World of Computer Science and Information Technology Journal*, vol. 2, no. 2, pp. 51–56, 2012.

[10] E. Osmanbegović and M. Suljić, "Data mining approach for predicting student performance," *Economic Review: Journal of Economics and Business*, vol. 10, no. 1, pp. 3–12, 2012.

[11] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Applied Artificial Intelligence*, vol. 34, no. 1, pp. 1–38, 2020.

[12] A. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.

[15] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[17] K. Gray and B. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Computers & Education*, vol. 131, pp. 22–32, 2019.

[18] M. Sweeney, J. Lester, and H. Rangwala, "Next-term student performance prediction: A recommender systems approach," *Journal of Educational Data Mining*, vol. 8, no. 1, pp. 22–51, 2016.

[19] M. Waheed et al., "Predicting academic performance of students from VLE big data using deep learning models," *Computers in Human Behavior*, vol. 104, 106189, 2020.

[20] S. Helal, J. Li, L. Liu, E. Ebrahimie, and D. Dawson, "Predicting academic performance by considering student heterogeneity," *Knowledge-Based Systems*, vol. 161, pp. 134–146, 2018.

[21] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.