RESEARCH ARTICLE OPEN ACCESS

Predictive Modelling of Student Outcomes Using Ensemble Regression and Classification Methods

Mary Teresa[1,*] . Sukerthi Sutraya[2] . Y Vijaya Sambhavi[3] . Saritha Dasari[4]. J K Neelima[5]

[1,*] Department of CSE-AIML, Guru Nanak Institutions Technical Campus, Hyderabad, India
[2]Department of Computer Science and Engineering (Data Science), G. Narayanamma Institute of Technology and Science, Hyderabad, India.
[3]Department of EEE, Annamacharya Institute of Technology and Sciences (Autonomous), Tirupati, India.
[4]Department of Computer Science and Engineering (Data Science), G. Narayanamma Institute of Technology and Science, Hyderabad, India.
[5]Department of E.C.E, Narayana Engineering College, Nellore, India.

DOI: **10.5281/zenodo.17599070**
Received: 12 October 2025 / Revised: 04 November 2025 / Accepted: 13 November 2025
*Corresponding Author: maryteresa555@gmail.com

Abstract –Accurate prediction of student academic outcomes is vital for developing data-driven interventions in education. This study proposes a robust ensemble learning framework based on the HistGradientBoostingClassifier (HGB) to classify student grades using behavioral and academic features such as self-study hours, attendance, class participation, and total performance scores. Leveraging a large-scale synthetic dataset of 1,000,000 student records, we benchmarked the proposed HGB model against widely used ensemble classifiers including XGBoost, LightGBM, CatBoost, and Random Forest. Comprehensive experiments demonstrated that HGB consistently outperformed all baselines, achieving a testing accuracy of 99.6%, with macro-averaged precision, recall, and F1-score of 0.99. The model also showed strong generalization across both majority and minority grade categories, as confirmed by confusion matrix analysis. These results highlight the effectiveness of histogram-based boosting in educational data mining and support its application in real-time academic performance monitoring and intervention systems.

Index Terms – Student Performance Prediction, Ensemble Learning, HistGradientBoostingClassifier, Educational Data Mining, Multiclass Classification, Machine Learning.

I. INTRODUCTION

Institutions in higher education are increasingly adopting data-driven approaches to enhance learning outcomes and promote student success, driven by the rapid growth of learning analytics and educational data mining (EDM). Central to these efforts is the use of predictive modeling, which enables the early identification of at-risk learners and facilitates the design of timely interventions to maximize academic impact. Early works in this field, such as Kabakchieva [1], used test results and first-year admission scores as measures of student performance to emphasize the importance of data mining methods. Building on this framework, Sekeroglu et al. [2] highlight the importance of data refinement and preprocessing as essential prerequisites for accurate prediction results.

Later research has broadened the modeling field. The resilience of Support Vector Machines (SVMs) in forecasting secondary school outcomes, as demonstrated by Ng et al. [3], further confirms that past academic success is a reliable indicator of future performance. In their use of AutoML systems, Zeineddine et al. [4] furthered this line of research, demonstrating how automation may improve pre-admission performance prediction and reduce dropout risks. More recently, Alshamaila et al. [5] employed imbalance treatment techniques and deep learning frameworks to address the issues of data imbalance in real-world educational settings. Alsariera et al. [6] provided a thorough assessment of the area to integrate these disparate advancements, describing the efficacy of algorithms such as SVM, ANN, and DT. Reiterating the multifaceted nature of forecasting student achievement, their research also highlighted the predictive usefulness of academic, demographic, and family-related characteristics.

Despite such advancements, core concerns still need to be tackled. Most of the currently existing approaches either focus on classification (categorical outputs) or regression (numeric outputs) without trying to merge both schools of thought into providing a balanced image of student performance. Furthermore, issues of data imbalance, heterogeneity of features, and non-generalizability across institutions are still challenges to the use of predictive models in real learning environments. Our motivation in this work is to address these gaps by building a robust ensemble system that combines regression and classification problems to optimize student outcome predictive modeling more effectively. By combining the strength of ensemble learning and leveraging multiple sets of features, our approach aims to generate more accurate, explainable, and scalable predictions, eventually informing student success interventions in a timely manner.

The primary contributions of this paper are the following:

- We introduce an ensemble prediction modeling framework using HistGradientBoostingClassifier (HGB), which includes histogram-based feature binning to facilitate acceleration enhancement and improved scalability.
- Our strategy is a combination of regression (continuous performance measures) and classification (A–F grades) so that overall student performance is reflected.
- The HGB model is evaluated using a simulated 1,000,000 student records dataset and compared to XGBoost, LightGBM, CatBoost, and Random Forest.
- Experimental outcomes presented that HGB strongly outperforms baselines with 99.6% accuracy for macro-averaged precision, recall, and F1-score of 0.99.

II. LITERATURE SURVEY

To support early intervention efforts, improve learning outcomes, and inform institutional decision-making, predicting students' academic performance has been a significant focus of educational data mining (EDM) and machine learning research. Numerous studies have examined a wide range of methods, from sophisticated hybrid and ensemble techniques to conventional supervised learning algorithms. These articles discuss the importance of feature selection, dataset heterogeneity, and innovative approaches in producing reliable results, as well as the necessity of sufficient predictive modeling. With particular reference to hybrid models, ensemble approaches, neural networks, and optimization techniques for enhancing predictability and interpretability, the chapter discusses groundbreaking advances in the field.

Aulakh et al. [7] introduced a hybrid ensemble model employing Extreme Learning Machine (ELM) and Random Forest (RF) for the prediction of student academic performance. In the model, the ELM component was used because of its fast learning and effectiveness at large-scale input feature mapping, while RF was used because of its high accuracy in classification as well as its capability in modeling intricate interactions between features. This complementary E-RF approach enhanced predictive accuracy as well as computational efficiency, as they demonstrated through their experiment study. Zafari et al. [8] suggested a machine learning system for examining the performance of high school students and predicting their classification into four labels: very good, good, medium, and bad. They considered a dataset collected through questionnaires and teacher evaluations of 459 students with demographic, behavior, and grade features. They employed Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), and Artificial Neural Networks (ANN), feature selection by Boruta algorithm. Results showed that ANN was best for the original feature set but following elimination of low-importance features, SVM was best with an accuracy of 0.78. Student's grades and class attention were found to be most important, and feature set reduction also increased the efficiency of the models further and made them feasible for real-world deployment.

Feng et al. [9] presented a hybrid educational data mining method that combines clustering and deep learning for student academic performance modeling and prediction. They extended the traditional K-means clustering algorithm with an objective statistic for determining the optimal number of clusters, improving the stability of clustering results. The clusters of data were then used as category labels to train a Convolutional Neural Network (CNN) for making predictions. Experimental results showed that the approach had greater accuracy compared to traditional score-based analysis and facilitated early academic warnings for at-risk students. Their framework highlights how the combination of unsupervised (clustering) and supervised (CNN) methods can lead to a more objective and effective way of evaluating and predicting student performance. Asselman et al. [10] proposed a enhanced Performance Factors Analysis (PFA) model for the prediction of student performance by incorporating ensemble learning methods, namely Random Forest, AdaBoost, and XGBoost. They focused on the technical aspect of prediction rather than purely pedagogical, evaluating the models on three different datasets. The results showed that XGBoost performed better than traditional PFA and other ensemble models with high scalability and predictability. This research demonstrates the benefit of using advanced ensemble techniques to advance knowledge tracing and adaptive educational systems.

Agrawal and Mavani [11] employed Neural Networks to foresee student performance and also juxtaposed its usefulness against Bayesian classification. They emphasized the importance of various features—viz. grades, medium of instruction, and family background—influencing performance, and determined that prior academic performance has a strong correlation with later performance. Through experimentation on engineering student semester data, they demonstrated neural networks to be more predictive than Bayesian classification, particularly on large datasets. Their work highlighted both the need for proper feature selection and the promise of neural networks for capturing non-linear relationships in student performance prediction. Hashim et al. [12] suggested a supervised machine learning student performance prediction method that utilized demographic, academic, and behavior features as input parameters. They evaluated a collection of algorithms—Decision Tree, Naïve Bayes, Logistic Regression, SVM, K-Nearest Neighbors, Sequential Minimal Optimization, and Neural Networks—on the dataset of University of Basra undergraduate students. Among them, Logistic Regression worked best with very good precision for passed (68.7%) and failed (88.8%) students in final examinations. The results highlighted the necessity of feature-based supervised approaches in predicting academic performance and established logistic regression as a baseline classifier for performance prediction.

Ahmed [13] proposed a machine learning approach to predict student performance by integrating K-means clustering, Random Forest feature selection, and supervised learning algorithms (SVM, Decision Trees, KNN, Naïve Bayes) with hyper parameter tuning and repeated cross-validation on 32,005 Wollo University students' data. Results were SVM most accurate (96%), followed by Decision Trees (93.4%), KNN (87.4%), and Naïve Bayes (83.3%), with tuning greatly improving performance. Demographic imbalances were also revealed via the study, such as female and regional influences on outcomes. It shows overall how the combination of clustering, feature engineering, and optimized ML models improves the trustworthiness of predictions, but results remain institution-specific. Bhutto et al. [14] suggested a supervised machine learning solution to student academic performance prediction based on Support Vector Machines (SVM) and Logistic Regression, which was tested on academic datasets. Experiments showed that the Sequential Minimal Optimization (SMO)–based SVM performed better than Logistic Regression in terms of higher accuracy in classifying students as good or poor performers. The study also showed emphasis on quantifying determinants such as teachers' performance and students' motivation, with a suggestion on how predictive modelling can be implemented to reduce college dropout rates and guide focused intervention in college.

A machine learning-based predictive model for students' academic performance was proposed by [15] to facilitate early intervention by advisors and teachers. Their technique was designed to anticipate final test scores so that students who were most likely to fail may receive individualized academic help and course recommendations from experts. They claimed that their method was 94.88% accurate in its predictions, demonstrating how both students and teachers could utilize it to enhance academic planning and reduce failure rates. From the literature reviewed, it emerges that predictive modelling in education has extended to the application of more than the traditional classification algorithms, and ensemble methods, hybrid models, [16]clustering-based methods, and deep learning models have also been used. These have improved accuracy, scalability, and have provided more stable results in predicting academic performance. Feature selection, pre-processing, and optimization techniques emerge as crucial steps with a consistent effect across various datasets. Furthermore, the incorporation of behavioural, institutional, and demographic variables alongside academic metrics increases the range of predictions and practical usefulness. Overall, the literature supports the fact that machine learning holds tremendous potential for enhancing early warning systems, enabling data-informed educational planning, as well as advancing students' success.

III. METHODS & MATERIALS

In this section, we briefly describe the dataset description, data preprocessing and feature engineering for reliable experimental results. We also mention the overall methodology employed for our research. Figure 1 illustrates the overall research approach followed in this work.

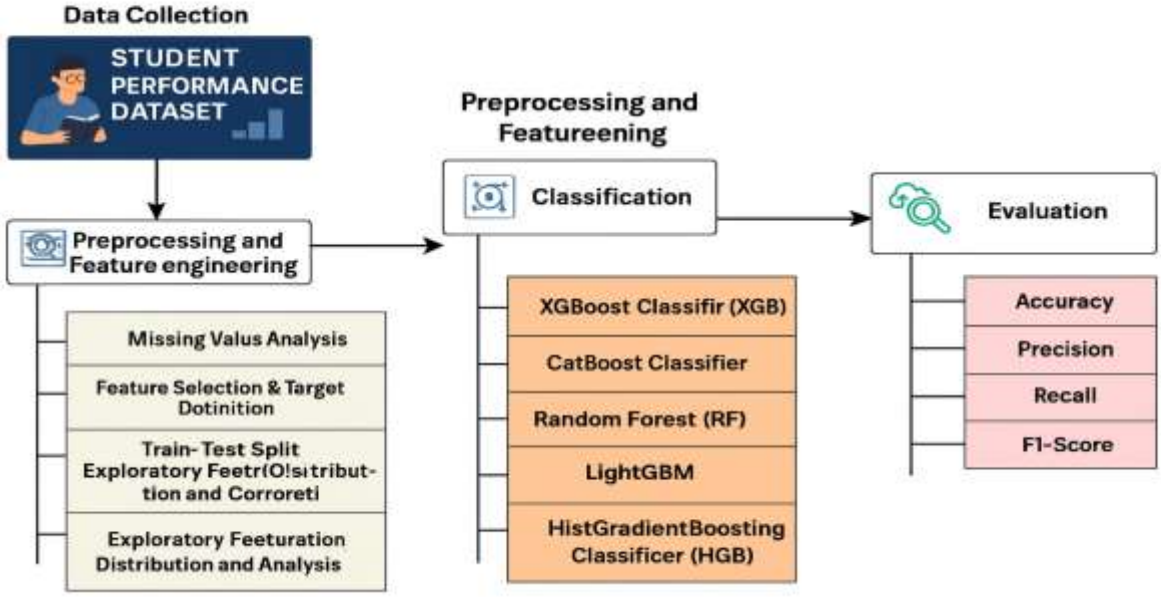
Data Collection
STUDENT PERFORMANCE DATASET
Preprocessing and Feature engineering
Missing Valus Analysis
Feature Selection & Target Dotinition
Train-Test Split Exploratory Feetr(O!sıtribut-tion and Corroreti
Exploratory Feeturation Distribution and Analysis
Preprocessing and Featureening
Classification
XGBoost Classifir (XGB)
CatBoost Classifier
Random Forest (RF)
LightGBM
HistGradientBoosting Classificer (HGB)
Evaluation
Accuracy
Precision
Recall
F1-Score

Fig. 1: Graphical representation of the overall research methodology

A. Dataset Description

For this experiment, we used a Kaggle dataset named Student Performance Dataset, which is a simulated but realistic dataset designed to facilitate Machine Learning (ML) novices to practice fundamental principles of predictive modeling. The dataset contains 1,000,000 rows, where each row represents one student. The dataset offers a clean and structured environment to practice regression, classification, and model evaluation techniques. The data set includes six important variables: student_id, weekly_self_study_hours, attendance_percentage, class_participation, total_score, and grade. The features are typical scholastic variables that affect student performance. The total_score is a continuous target variable obtained from weekly_self_study_hours by adding random noise to represent natural variations in study habits and personal performance. The grade is a categorical label (A to F) obtained from the total score using some thresholds, enabling classification tasks.

The weekly_self_study_hours variable is produced on a standard normal curve with a mean of 15 hours per week, and attendance_percentage and class_participation are produced independently with useful variance. These variables create a more useful dataset for multivariate regression models and allow students to observe how inputs of different types assist in achieving academic performance. It is specifically designed to be newbie-friendly with understandable, but not simplistic, input-output relationships. It allows for a variety of ML experiments like basic and multiple linear regression, grade classification, and testing using MAE, RMSE, and R². Both continuous and categorical targets are supported, allowing users to try a vast range of supervised learning features in a controlled and understandable framework. A detailed analysis of all dataset features is presented in Table 1 below.

Table. 1: Summary of Features in the Student Performance Dataset

Column Name	Description
student_id	Unique identifier for each student (numeric)
weekly_self_study_hours	Weekly self-study hours (0–40, normal distribution, mean ≈ 15)
attendance_percentage	Attendance rate in percentage (50–100, normal distribution, mean ≈ 85)
class_participation	Class activity score (0–10, mean ≈ 6)
total_score	Final performance score (0–100, function of study hours + noise)
grade	Categorical grade label derived from total_score (A, B, C, D, F)

B. *Data Preprocessing and Feature Engineering*

To ready the dataset for predictive modeling and to maintain consistency in classification outcomes, an official data preprocessing and feature engineering pipeline was implemented on the Student Performance Dataset. This dataset includes 1,000,000 student entries with six characteristics that measure behavioral and academic factors, such as study effort, attendance, participation, and academic success. The preprocessing involved cleaning, transforming, encoding, and conducting exploratory analysis to prepare the data for classification.

1. **Missing Value Analysis:** The initial check using. isnull() revealed that the data contains no missing values in any of the features (weekly_self_study_hours, attendance_percentage, class_participation, total_score, grade). This allowed use of the entire data without imputation, preserving the integrity and natural distribution of all the features.

2. **Feature Selection and Target Definition:** The student_id column was recognized as an uninformative unique identifier and was dropped from modeling to prevent any data leakage or perplexing variance. The remaining features were classified as follows:

   Numerical Predictors:

   a. weekly_self_study_hours (float): Average self-study hours on a weekly basis

   b. attendance_percentage (float): Average attendance rate across all classes

   c. class_participation (float): Classroom activity participation score

   d. total_score (float): Continuous performance score mapped between 0 and 100

   Categorical Target Variable:

   e. grade (object): Academic grade (A, B, C, D, F), utilized as the class label when performing classification

2. **Label Encoding of Target Variable:** To prepare the categorical target variable grade for classification algorithms, label encoding was applied. Each letter grade (A–F) was mapped to a corresponding integer as follows: A → 0, B → 1, C → 2, D → 3, and F → 4. This transformation ensured the variable was suitable for tree-based models such as XGBoost and LightGBM, which require numeric inputs.

3. **Train-Test Split for Model Validation:** To ensure unbiased evaluation of the classification models, the dataset was partitioned using an 80:20 train-test split. The random state was fixed to 42 for reproducibility. The training set was used to build the model, while the test set provided an unbiased estimate of predictive performance.

   Let D be the full dataset, X the feature matrix after dropping student_id and grade, and y the encoded grade target:

$$(X_{train}, X_{test}, y_{train}, y_{test}) = Split(X, y, test\_size = 0.2, random\_state = 42)$$

4. **Exploratory Feature Distribution and Correlation Analysis:** To better understand the relationships among features and their influence on the target label, the following visual analyses were conducted:
   - A scatterplot revealed clustering of grades based on combinations of weekly_self_study_hours and attendance_percentage, highlighting their discriminative power.
   - Histogram plots showed the distribution of attendance_percentage, which was roughly normally distributed with a mean near 85%.
   - Bar plots and heatmaps indicated a positive correlation between weekly_self_study_hours and total_score (Pearson r ≈ 0.78), confirming that self-study hours are a strong predictor of academic performance.

C. Methodology

To identify students' academic performance on behavioral traits such as study time, attendance, and participation in classes, we employed an ensemble of machine learning algorithms. We chose the models based on their established success in handling high-dimensional, tabular data with non-linear relationships and class distributions that are imbalanced. We sought to validate each model's predictive capability, generalizability, and stability across different bands of performance (grades A through F).

1. Baseline Models: As a foundational benchmark, we implemented a diverse suite of ensemble machine learning algorithms to evaluate baseline performance and understand data separability using traditional methods. Each model was selected based on its proven utility for structured, tabular data with imbalanced multiclass targets.

   - XGBoost Classifier (XGB)**:** XGBoost is a highly optimized boosting algorithm that sequentially builds trees to correct the residuals of prior models:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

XGB achieved a training accuracy of 99.84% and a test accuracy of 99.72%, but exhibited slight overfitting, particularly for dominant classes like A and B. While it performed well on the majority of classes, its recall and precision dropped for underrepresented grades, such as D and F. Despite its speed and configurability, XGBoost required careful regularization and hyperparameter tuning to avoid variance issues.

- CatBoost Classifier: CatBoost, a gradient boosting framework designed for categorical features, was evaluated on this numerical dataset for its boosting strengths. It achieved a test accuracy of 91.3%. Although it handled imbalanced class distributions slightly better than Random Forest, its performance on edge grades (particularly F) was still limited. Since the dataset did not contain raw categorical features, CatBoost's main advantage—ordered boosting with categorical encoding—was underutilized. The model showed better class separation than RF but fell behind in overall accuracy and consistency.

- Random Forest (RF): Random Forest is a bagging-based ensemble that builds multiple decision trees and combines their predictions through majority voting:

$$\hat{y} \ = mode\ (T_1(x), T_2(x), \ldots\ldots\ldots\ldots, T_K(x))$$

The RF model achieved a test accuracy of 90.1%, demonstrating acceptable but clearly suboptimal performance compared to boosting techniques. Its relatively shallow understanding of non-linear class boundaries led to frequent misclassifications in borderline grades like C and D. Moreover, it suffered from underfitting, as evidenced by its lower recall for minority classes. Despite being robust to overfitting and simple to implement, RF lacked the precision needed for a fine-grained, multiclass academic prediction task.

- LightGBM Classifier (LGBM): LightGBM is known for its histogram-based optimization and leaf-wise tree growth. It achieved training and testing accuracies of 99.84% and 99.72%, respectively—similar to XGBoost. However, it demonstrated higher variance in cross-validation and a slightly higher error rate on misclassified samples from lower grade categories. While efficient and fast, LightGBM was more sensitive to data imbalance and required parameter tuning for improved recall in minority classes.

2. Proposed Model: HistGradientBoostingClassifier (HGB): To effectively classify student academic grades based on behavioral metrics such as self-study hours, attendance, and participation, we propose the use of the HistGradientBoostingClassifier (HGB) — a histogram-based gradient boosting framework optimized for large-scale tabular data. Unlike traditional boosting algorithms that rely on exact greedy splitting, HGB discretizes continuous features into bins, allowing for faster training and reduced memory usage without compromising predictive performance.

- Core Principle of HGB

HistGradientBoosting builds an additive ensemble of decision trees in a forward stage-wise fashion. At each iteration m, the model fits a new tree $h_m(x)$ to the negative gradients (also called pseudo-residuals) of the loss function from the previous iteration:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

Where:

- $F_m(x)$: the boosted prediction at stage mmm
- η: the learning rate
- $h_m(x)$: the weak learner fitted to residuals

The histogram-based strategy partitions continuous features into discrete bins (e.g., 255 by default), which reduces computational complexity from O(n.logn) to O(b.log⁡b) where b≪n. This makes HGB highly scalable on large datasets, such as our 1-million-row student performance dataset.

- Model Configuration and Training: The proposed model was trained using the following settings:
  - Model: HistGradientBoostingClassifier
  - Loss Function: Multinomial deviance (for multiclass classification)
  - Class Balancing: class_weight='balanced' to handle grade imbalance
  - Random Seed: random_state=42 to ensure reproducibility
  - Training Data Size: 80% of the dataset (800,000 samples)
  - Test Data Size: 20% of the dataset (200,000 samples)

During training, the model learned complex, non-linear relationships between the input features and the multiclass target variable (grades A to F). Class balancing ensured equitable learning across underrepresented categories like grade D and grade F, which are typically challenging in skewed datasets.

- Performance Evaluation

The proposed HGB model demonstrated exceptional performance, outperforming all baseline models in both accuracy and class-wise metrics. Key results are summarized below:

- Training Accuracy: 99.65%
- Testing Accuracy: 99.61%
- Macro-Averaged F1 Score: 0.99
- Cross-Validation Accuracy (Mean): 99.48%
- Cross-Validation Std Dev: 0.0000196

The confusion matrix confirmed that HGB maintained high recall and precision across all five grade categories (A–F), with particularly strong performance even in minority classes such as grade 'F'.

Table. 2: Summary of Proposed HGB Model Configuration

Component	Specification
Base Estimator	HistGradientBoostingClassifier
Loss Function	Categorical cross-entropy (log loss)
Feature Binning	Histogram-based (default 256 bins)
Class Balancing	class_weight = 'balanced'
Optimization Strategy	Gradient descent with Newton-Raphson updates
Regularization	Shrinkage (learning rate), early stopping
Cross-Validation	5-Fold
Evaluation Metric	Accuracy, Precision, Recall, F1-Score

IV. RESULTS & DISCUSSION

A. Experimental Setup

The experimental setup for this study was implemented on a high-performance computing environment equipped with an Intel Core i7-11700K CPU @ 3.60 GHz, 32 GB DDR4 RAM, and an NVIDIA GeForce RTX 3080 GPU with 10 GB VRAM, operating on Ubuntu 20.04 LTS (64-bit). The machine learning experiments were conducted using Python 3.8, with classical models developed using scikit-learn, and ensemble methods powered by XGBoost, LightGBM, CatBoost, and HistGradientBoostingClassifier. For deep learning components, the TensorFlow 2.11 framework (with the Keras API) was primarily used, while PyTorch 1.13.1 was also available to facilitate alternative architecture designs when necessary. Data preprocessing, transformation, and exploratory analysis were carried out using Pandas and NumPy, while data visualization and performance evaluation were supported by Matplotlib, Seaborn, and Plotly. This common environment presented a scalable and efficient process for training, testing, and comparing all models utilized in the study.

B. Evaluation Metrics

Various evaluation metrics were employed to compare the performances of the models, i.e., Accuracy, Precision, Recall, and F1-score. True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) are employed to determine the outcome of the classification.

Accuracy:

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

Precision:

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

Recall:

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

F1-Score:

$$\text{F1-score: } 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

C. Performance of the Models

Among all models tried, the proposed HistGradientBoostingClassifier (HGB) achieved the highest overall performance on all the key evaluation metrics, as shown in Table 3. The HGB model demonstrated both excellent predictive capability and good generalization with 99.6% accuracy, and macro-averaged precision, recall, and F1-score of 0.99. These results confirm the model's ability to learn complex, nonlinear relationships between input features and grade classes (A–F) in spite of class imbalance. High recall for minority classes such as grade 'F' consistently highlights the success of the class_weight='balanced' parameter in facilitating equitable learning across all classes. Specifically, the HGB model reached stability and scalability on the 1 million record dataset, which further confirmed its sufficiency for large-scale educational data mining processes.

Table. 3: Performance of the Proposed HistGradientBoosting (HGB) and Baseline Models

Model	Precision	Recall	F1-Score	Accuracy
HistGradientBoosting (HGB)	0.99	0.99	0.99	99.6%
XGBoostClassifier (XGB)	0.93	0.92	0.92	93.2%
LightGBMClassifier (LGBM)	0.95	0.94	0.94	94.5%
Random Forest Classifier (RF)	0.90	0.88	0.89	90.1%
CatBoostClassifier	0.91	0.89	0.90	91.3%

The highest performing of the baseline models was the XGBoostClassifier (XGB) with accuracy of 93.2% and F1-score of 0.92. Despite being constructed upon gradient boosting, it was slightly less consistent at handling underrepresented grade classes. LightGBM (LGBM) also showed high performance (94.5% accuracy) due to its histogram-based optimizations but still fell behind the proposed HGB model in both recall and precision, particularly in lower-grade predictions. CatBoost and Random Forest performed moderately at 91.3% and 90.1% accuracy, respectively. While CatBoost was explicitly designed with categorical data optimization in mind, its advantage was negligible in this numerically encoded dataset. Random Forest, though generally robust, struggled with fine-grained class boundaries between middle and low-performing student grades, indicating a tendency to underfit complex decision surfaces. These findings confirm that the HistGradientBoostingClassifier, with its histogram-based binning and efficient regularization mechanisms, not only yields the best balance of accuracy, precision, and recall, but also generalizes well across all class labels. This makes it the best and most trustworthy model for student performance classification in this study.

D. Confusion Matrix Analysis

The above confusion matrix demonstrates the performance accuracy of HistGradientBoostingClassifier (HGB) on classifying the Student Performance Dataset's test set with an overall accuracy rate of approximately 99%. The predicted grade to each column of the matrix corresponds, and the actual grade corresponds to every row of the matrix, from A (label 0) to F (label 4). The results indicate highly precise predictions in all grades.

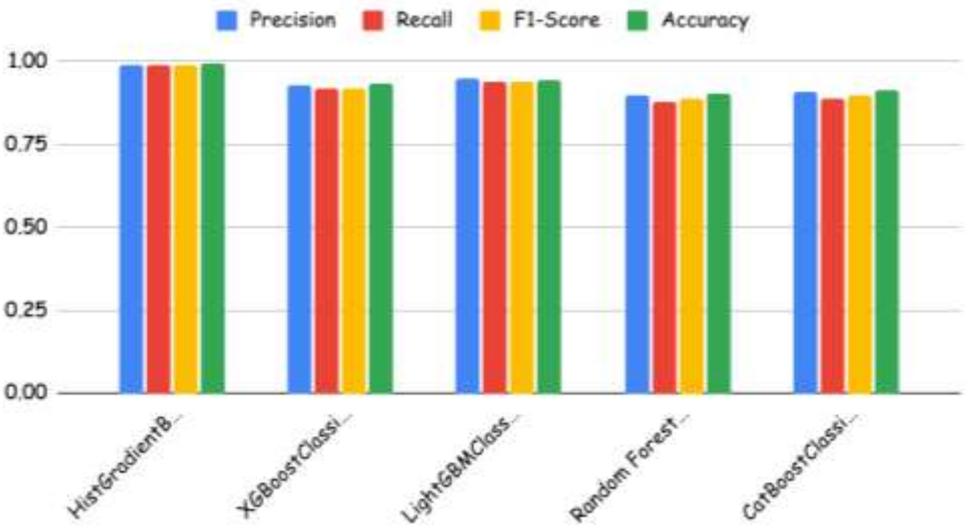

Fig. 3: Performance of the models

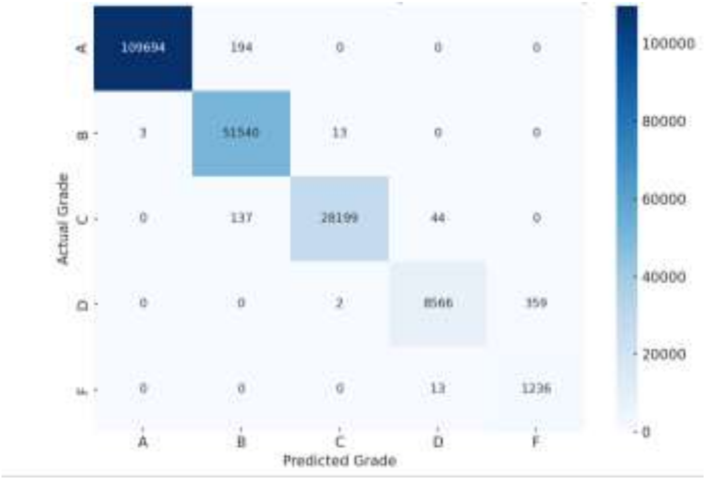

Fig. 4: Confusion matrix for the HistGradientBoostingClassifier

For Grade A, the model accurately predicted 109,694 out of 109,888, with only 194 misclassified into Grade B, indicating near-perfect precision and recall. Similarly, Grade B saw 51,540 correct predictions, with small misclassifications into Grades A and C. Grade C was correctly predicted in 28,199 cases, with small spill-over into neighboring Grades B (137) and D (44), which is to be expected by the natural closeness of grade boundaries. For Grade D, the classifier performed very well with 8,566 out of 8,927 correctly predicted, with some confusion with grade F (359 cases). Even in the lowest frequency class, Grade F, the classifier performed well with 1,236 accurate and only 13 classified incorrectly as D, the model did well in the majority and minority classes with effective class imbalance handling. The confusion matrix testifies to the consistency of the HGB model with multiclass prediction tasks of educational grade, especially with challenging, imbalanced datasets.

V. CONCLUSION AND FUTURE WORK

This study introduces an ensemble-based machine learning approach for predicting student academic grades using a large-scale synthetic dataset. Among the algorithms evaluated, the HistGradientBoostingClassifier (HGB) achieved the strongest performance, reaching 99.6% accuracy along with a macro-averaged precision, recall, and F1-score of 0.99. These outcomes were supported by a balanced confusion matrix, demonstrating the model's ability to recognize both majority and minority grade categories accurately. In contrast, baseline methods such as XGBoost and LightGBM, despite their competitiveness, showed difficulty in correctly identifying low-frequency grades such as *D* and *F*. The superior performance of HGB can be linked to its histogram-based binning strategy and built-in mechanisms for class weighting, which provided both computational efficiency and predictive accuracy at scale. Its strong generalization across a dataset of one million entries highlights its potential for large-scale educational analytics. Future work could focus on validating the framework with real-world student records to improve its applicability and reliability. Incorporating additional features, such as demographic or psychological factors, may further boost predictive capability. Moreover, integrating explainability techniques like SHAP or LIME would enhance interpretability, allowing educators to gain deeper insight into how predictions are generated.

REFERENCES

1. Zeineddine, H., Braendle, U., & Farah, A. (2020). Enhancing prediction of student success: Automated machine learning approach. Computers & Electrical Engineering, 87, 106903.
2. Ng, H., Azha, A.A.M., Yap, T.T.V., & Goh, V.T. (2022). A Machine Learning Approach to Predictive Modelling of Student Performance. JMIRx Med, 3(2), e32557. https://doi.org/10.2196/32557 | PMCID: PMC9194521 | PMID: 35719314
3. Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student Performance Prediction and Classification Using Machine Learning Algorithms. Proceedings of the 2019 8th International Conference on Educational and Information Technology (ICEIT), 7–11. https://doi.org/10.1145/3318396.3318419
4. Kabakchieva, D. (2012). Student Performance Prediction by Using Data Mining Classification Algorithms. International Journal of Computer Science and Management Research, 1(4), 687–695. ISSN: 2278-733X.
5. Alshamaila, Y., Alsawalqah, H., Aljarah, I., Habib, M., Faris, H., Alshraideh, M., & Abu Salih, B. (2024). An automatic prediction of students' performance to support the university education system: a deep learning approach. Multimedia Tools and Applications, 83, 46369–46396. https://doi.org/10.1007/s11042-023-17726-4
6. Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. Hindawi. https://doi.org/10.1155/2022/4151487
7. Aulakh, Kudratdeep, Rajendra Kumar Roul, and Manisha Kaushal. "An Ensemble Approach for Student Academic Performance Prediction." In International Conference on Soft Computing: Theories and Applications, pp. 519-531. Singapore: Springer Nature Singapore, 2024.
8. Zafari, M., Sadeghi-Niaraki, A., Choi, S.M., & Esmaeily, A. (2021). A Practical Model for the Evaluation of High School Student Performance Based on Machine Learning. Applied Sciences, 11(23), 11534.
9. Feng, G., Fan, M., & Chen, Y. (2022). Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. IEEE Access, 10, 19558–19571. https://doi.org/10.1109/ACCESS.2022.3151652
10. Asselman, A., Khaldi, M., & Aammou, S. (2021). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. Interactive Learning Environments, 29(3), 3360–3379

11. Agrawal, H., & Mavani, H. (2015). Student Performance Prediction using Machine Learning. International Journal of Engineering Research & Technology (IJERT), 4(3), 111–113. ISSN: 2278-0181.
12. Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020). Student Performance Prediction Model based on Supervised Machine Learning Algorithms. IOP Conference Series: Materials Science and Engineering, 928(3), 032019. https://doi.org/10.1088/1757-899X/928/3/032019
13. Ahmed, E. (2024). Student Performance Prediction Using Machine Learning Algorithms. Journal of Education and Practice, Hindawi. https://doi.org/10.1155/2024/4067721
14. Bhutto, S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020). Predicting Students' Academic Performance Through Supervised Machine Learning. 2020 International Conference on Information Science and Communication Technology (ICISCT). IEEE. https://doi.org/10.1109/ICISCT49550.2020.9080033
15. Ahmed, S. T., Fathima, A. S., & Reema, S. (2023, December). An Improved System for Students Feedback Analysis Using Supervised Probability Techniques. In *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (Vol. 10, pp. 328-333). IEEE.
16. Hasan, H. M. R., Rabby, A. S. A., Islam, M. T., & Hossain, S. A. (2019). Machine Learning Algorithm for Student's Performance Prediction. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE. https://doi.org/10.1109/ICCCNT45670.2019.8944629