International Journal of Current Research and Techniques (IJCRT)
Volume 16, Issue 2, 2026; DOI: 10.5281/ZENODO.19606501
Original Research Article

Predicting Student Performance Using Demographic and Attendance Data with Machine Learning

Aman Patel, *Department of Computer Science , MIT Arts, Commerce and Science College, Pune, India*.

Samir Ambre, *Department of Computer Science , MIT Arts, Commerce and Science College, Pune, India*.

Vaibhav Gawade, *Department of Computer Science , MIT Arts, Commerce and Science College, Pune, India*.

Manuscript Received: Apr 10, 2026; Revised: Apr 14, 2026; Published: Apr 16, 2026

Abstract: This study investigates the influence of demographic, academic, and behavioral factors on student performance using advanced machine learning techniques. In modern education, success is shaped not only by classroom learning but also by background, habits, and daily behavior. The dataset included variables such as gender, family income, SSC and HSC scores, hometown, computer usage, preparation time, attendance, social media activity (watching short videos or "Reels"), and part time jobs. These features were analyzed alongside semester GPA and overall results to identify the strongest predictors of academic success. To evaluate these factors, both traditional classifiers (Random Forest, KNN, SVM, Naïve Bayes, Logistic Regression) and advanced boosting algorithms (CatBoost, LightGBM, XGBoost) were applied, along with ensemble methods such as Voting, Bagging, and Stacking. Among all models, Stacking Ensemble consistently achieved the highest accuracy of 95%, outperforming other approaches. Analysis showed that attendance, prior exam scores, and preparation time were the most influential predictors, while heavy social media use reduced concentration and performance. Visualizations including accuracy comparison charts, feature importance plots, and confusion matrices confirmed these findings. The CatBoost confusion matrix demonstrated reliable classification, with most predictions correctly aligned along the diagonal. The study highlights that while demographic factors such as income and hometown contribute to differences in performance, disciplined study habits and consistent attendance can help students overcome these challenges. These insights provide actionable guidance for educators and institutions. Schools can promote attendance, encourage effective study routines, and raise awareness about the impact of social media use. By leveraging machine learning, institutions can identify at risk students early and design targeted interventions to improve outcomes. Overall, this research demonstrates that student success depends on a combination of background, behavior, and daily habits. By integrating advanced machine learning methods, the study not only achieves high predictive accuracy but also offers practical strategies for supporting students. With proper support, motivation, and discipline, every student has the potential to enhance their performance and reach academic success.

Keywords: Student Performance, Attendance, SSC Score, HSC Score, Family Income, Preparation, Job, Semester, Hometown, Machine Learning, CatBoost , Stacking Ensemble, Ensemble Methods, Feature Importance, Confusion Matrix.

1. Introduction

Education is widely recognized as one of the strongest pillars of personal and social growth, equipping individuals with knowledge, skills, and opportunities that shape their future and enable them to contribute meaningfully to society [1], [10]. Academic success plays a crucial role in determining access to higher education, career prospects, and long-term social mobility [2], [12]. For institutions, student achievement is not only about grades; it also involves nurturing responsible, confident individuals who can positively influence their communities. As a result, understanding the factors that drive student performance has become a major focus for educators, policymakers, and researchers [3], [13]. Student outcomes are influenced by a complex mix of demographic, academic, and behavioral elements. Demographic aspects such as gender, family income, and hometown often shape the educational environment students encounter [4], [14]. Academic background, including prior results in SSC and HSC examinations, reflects foundational knowledge and study discipline [6], [15]. Behavioral factors, such as attendance, preparation time, computer literacy, English

proficiency, and social media use, capture how students interact with their learning environment on a daily basis [5], [16]. Considering these dimensions together provides a more holistic and realistic view of student performance [7], [17].

Demographic Influences: Family income often determines the resources available to learners [8], [18]. Students from wealthier households may access private tutoring, digital tools, and supportive study environments, while those from lower-income families may face financial responsibilities, limited technology, or the need to work part-time [9], [19]. Similarly, whether a student comes from an urban or rural background can affect exposure to technology, school quality, and extracurricular opportunities [11], [20]. Urban students may benefit from better infrastructure, while rural students often show resilience despite fewer resources. Gender also shapes educational experiences, as cultural expectations and household duties may differ for boys and girls [12], [21]. Recognizing these demographic influences is essential for designing fair and inclusive educational policies [13], [22].

Attendance as a Key Predictor: While demographics set the starting point, attendance reflects day-to-day engagement with learning. Regular class participation ensures continuous exposure to lessons, active involvement in discussions, and timely feedback from teachers [14], [23]. Students with consistent attendance are more likely to stay aligned with the syllabus and maintain steady progress, whereas irregular attendance often leads to gaps and declining performance [15], [24]. Falling attendance can also serve as an early warning sign of personal or academic difficulties, allowing institutions to provide timely support through counseling or mentorship [16], [25]. This makes attendance one of the most critical behavioral indicators of academic success [17], [26].

Habits, Lifestyle, and Technology: Beyond demographics and attendance, lifestyle choices strongly influence academic outcomes. Preparation time reflects the effort invested outside the classroom, and disciplined study routines often correlate with better results [18], [27]. However, modern distractions, particularly excessive use of social media and short video content such as “Reels”, can reduce focus and productivity [19], [28]. Moderate leisure activities may refresh students, but excessive screen time often leads to procrastination. Participation in extracurricular activities enhances teamwork, leadership, and time management skills, indirectly supporting academic success [20], [29]. Computer literacy and English proficiency have also become increasingly important in today’s digital learning environment, where assignments and resources are frequently delivered online [21], [30].

2. Objective of the Research

The primary objective of this study is to design and evaluate a predictive framework capable of accurately assessing student academic performance by analyzing demographic, academic, and behavioral data through modern machine learning methods [1], [3], [12]. As data-driven decision making becomes increasingly central to education, it is essential to identify the real factors that shape student success and to develop models that can translate these insights into actionable strategies [7], [19]. This research seeks to bridge the gap between students’ backgrounds, their daily learning behaviors, and the role of predictive analytics in understanding academic outcomes more effectively [13], [20].

More specifically, the study focuses on identifying the key demographic and behavioral variables, such as gender, family income, department, attendance, and prior academic achievements (SSC and HSC scores), that significantly influence student performance [2], [6], [14]. By examining these elements collectively, the research aims to uncover meaningful patterns that differentiate high-performing students from those who may struggle academically [8], [15]. Such insights can help institutions detect early signs of academic difficulties and provide timely interventions to support students at risk [9], [16].

Another important goal is to evaluate and compare the predictive performance of different machine learning algorithms. Traditional models such as Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression are tested alongside advanced boosting algorithms (CatBoost, LightGBM, XGBoost) and ensemble methods (Voting, Bagging, Stacking) [17], [23], [24]. Each technique offers unique strengths: Random Forest excels at feature selection [4], SVM provides precise classification boundaries [11], Logistic Regression offers interpretability [10], while boosting algorithms handle categorical data effectively [18], [26]. Through this comparative analysis, the study identifies the Stacking Ensemble as the most effective model, achieving 95% accuracy and delivering superior predictive performance across all evaluation metrics [5], [25].

Beyond prediction, the research also explores the relationship between students' backgrounds and their learning behaviors, particularly attendance and preparation time [14], [20]. Attendance is treated as a proxy for motivation, discipline, and engagement, while preparation time reflects study effort outside the classroom [15], [27]. The study examines whether consistent attendance and disciplined preparation can offset challenges posed by socioeconomic limitations such as lower income or rural upbringing [28]. This balanced perspective highlights that consistent effort can often overcome background-related disadvantages [29].

The objectives extend beyond technical evaluation to practical application. The study aims to provide actionable insights for teachers, institutions, and policymakers [19], [21]. Early identification of at-risk students can enable targeted mentoring, counseling, or remedial programs [22]. Institutions can design better attendance policies, scholarship schemes, and learning support initiatives based on predictive findings [30]. By integrating machine learning into educational practice, this research supports the creation of proactive, data-informed strategies that promote equitable student success [25].

Ethical use of educational data is another underlying objective. Predictive models must be transparent, fair, and supportive, ensuring that analytics empower students rather than stigmatize them [23], [24]. This study advocates for a human-centered approach where machine learning complements teaching by offering insights into student progress, learning gaps, and strengths [19], [25]. The predictive framework developed here can guide the creation of future educational tools such as smart attendance systems, personalized study planners, and AI-based mentoring platforms [26].

Ultimately, this research strives to close the gap between technology and teaching by demonstrating how predictive analytics can support educators rather than replace them [18], [28]. By responsibly applying machine learning, institutions can adopt a proactive approach to student development, identifying potential challenges and providing necessary support to ensure long-term academic success.

3. Results

Yadav & Pal (2012) Yadav and Pal investigated demographic factors such as gender, parents' education, family income, and location, alongside attendance data, to predict academic performance. They applied decision tree algorithms (ID3, CART, and C4.5) and found C4.5 achieved the best accuracy (~67.8%). Their study demonstrated that attendance combined with demographic data significantly improves prediction accuracy.

Pal & Pal (2013) this study extended earlier work by including attendance records and evaluating decision tree methods (ID3, ADT, and Bagging). ID3 outperformed other algorithms with 78% accuracy, highlighting the positive impact of attendance data combined with demographics in predicting academic success.

Kabakchieva (2013) Kabakchieva analyzed demographic variables (age, gender) and attendance data among university students. Multiple classifiers including J48 and KNN were used, with J48 achieving 66.5% accuracy. The inclusion of attendance improved prediction outcomes, underlining its importance alongside demographic data.

Ramesh et al. (2013) Ramesh and colleagues combined demographic features and attendance data to predict student grades in higher secondary education. Decision tree algorithms reached approximately 85% accuracy, supporting the inclusion of attendance in demographic models.

Abu Saa (2016) Abu Saa's research involved detailed demographic data and attendance rates to predict student GPA. Several classification methods (CART, Naïve Bayes, CHAID) were compared, with CART performing best (40% accuracy). The study emphasized that attendance is a critical factor in modeling academic performance with demographics.

Asif et al. (2017) the researchers combined demographic data and attendance records for IT undergraduates. They used Decision Tree, Random Forest, Naïve Bayes, and Neural Networks. Naïve Bayes showed the highest accuracy (~83.7%), validating the integration of attendance and demographics in prediction models.

Akanbi et al. (2019) Using 325 student records with demographics and attendance history, Akanbi et al. applied the J48 decision tree algorithm, which resulted in 89.2% accuracy. Attendance data significantly enhanced the predictive capacity of demographic-based models.

Cortez & Silva (2008) In Portuguese schools, this study incorporated demographic data and attendance patterns, along with academic grades, for performance prediction. Techniques such as decision trees and random forests were used, achieving high accuracy when attendance was included as a predictor.

Sahlaoui et al. (2023) focused on imbalanced educational data, this research applied SMOTE-based resampling and balanced random forests on demographic and attendance data. They reported approximately 96% accuracy, highlighting attendance as a strong predictor when balanced with demographic attributes.

Alalawi et al. (2023) this systematic review found that combining demographics and attendance data consistently improved student performance prediction across 162 studies. Algorithms like Random Forests, Decision Trees, and SVM were effective, reinforcing attendance as a crucial variable alongside demographics.

Ouatik et al. (2023) Ouatik and colleagues integrated demographic information with attendance and virtual learning environment (VLE) activity data. Using KNN, C4.5, and SVM on large datasets, they showed that attendance combined with demographic data boosts predictive performance in Big Data contexts.

Ibrahim et al. (2024) the study used demographic profiles plus attendance rates and first-semester CGPA to predict academic outcomes. Artificial Neural Networks outperformed decision trees and regression models, achieving over 80% accuracy, confirming attendance as a key predictor alongside demographics.

Tanveer (2024) Tanveer's research confirmed that attendance, alongside demographic factors and academic scores (quizzes, labs, midterms), is an important predictor of student performance. Including attendance improved model accuracy significantly.

Agarwal & Agarwal (2024) in a comparative analysis using demographic, attendance, and psychological data, Random Forest classifiers achieved 100% accuracy. This highlights the synergy between attendance and demographic variables in high-performing prediction models.

Kocakoyun-Aydogan et al. (2024) they predicted student end-of-term performance using demographic and attendance data, applying Random Forest, KNN, SVM, and Naïve Bayes. Random Forest showed highest accuracy (93–97%), confirming the value of attendance alongside demographics in academic predictions.

Chen & Guestrin (2016) Introduced XGBoost, a scalable gradient boosting framework that quickly became popular in educational data mining. Studies applying XGBoost to student performance prediction reported higher accuracy compared to traditional decision trees, due to its ability to handle complex feature interactions and imbalanced datasets.

Prokhorenkova et al. (2018) Developed CatBoost, a gradient boosting algorithm optimized for categorical data. In educational contexts, CatBoost has shown superior performance in predicting student outcomes, as it efficiently handles categorical features such as gender, department, and hometown without extensive preprocessing. Its feature importance analysis also provides interpretable insights for educators.

Ke et al. (2017) Proposed LightGBM, a gradient boosting framework designed for speed and efficiency on large datasets. Applied to educational data, LightGBM has demonstrated strong predictive accuracy while reducing computational cost, making it suitable for real-time learning analytics.

Friedman (2001) Proposed Gradient Boosting Machines (GBM). Studies applying GBM to educational datasets showed improved accuracy compared to single classifiers, especially when attendance and demographics were included.

Breiman (1996) Introduced Bagging (Bootstrap Aggregating), which improves stability and accuracy by combining multiple models trained on resampled datasets. In student performance prediction, Bagging has been used to reduce variance and improve reliability of demographic-based models.

Dietterich (2000) Outlined ensemble methods such as Voting and Stacking. Voting classifiers combine predictions from multiple models (e.g., Random Forest, SVM, Logistic Regression) to improve overall accuracy. Stacking uses meta-learners to integrate base models, often outperforming individual algorithms. In educational prediction tasks, these ensemble approaches have proven effective in balancing interpretability and accuracy.

Sharma et al. (2023) Applied sentiment analysis on student forums, linking emotional engagement with academic success. Attendance data improved predictive accuracy.

Patel & Menon (2024) Studied social media activity. Balanced online engagement improved predictive accuracy by ~9% when combined with attendance.

Agarwal & Mehta (2024) Applied LSTM deep learning models to attendance trends. Achieved more precise GPA predictions than traditional methods.

Menon et al. (2024) Integrated self-assessment reports and teacher feedback. Showed motivation and confidence influence learning outcomes.

Chatterjee & Rao (2025) Focused on inclusivity, designing AI models accessible to differently-abled students. Attendance remained a strong predictor.

Khan & Verma (2025) Developed culturally sensitive models, accounting for language diversity. Attendance and demographics improved fairness in predictions.

Patel & Mehta (2025) Included lifestyle factors (screen time, sleep, physical activity). Found strong influence on focus and performance.

Nair et al. (2025) Proposed cloud-based real-time analytics framework. Predictions updated automatically with new attendance or assignment data.

Agarwal & Agarwal (2024) Highlighted Explainable AI (XAI). Recommended hybrid methods combining transparent algorithms (Logistic Regression, Decision Trees) with advanced models (Random Forest, CatBoost).

The prediction of student performance has consistently been a central theme in educational data mining and learning analytics [1], [2], [10]. Early studies primarily focused on measurable factors such as grades, attendance, and demographic details, using decision tree algorithms and statistical models [4], [6], [8]. Earlier studies confirmed that combining attendance with demographic variables consistently produced dependable predictions of academic success [9], [12]. Over time, however, researchers recognized that academic achievement is influenced not only by background and attendance but also by emotional, social, and behavioral dimensions [16], [20].

Recent studies between 2023 and 2025 have expanded the scope of prediction models to include emotional engagement, social media activity, and lifestyle habits. For instance, Sharma and Gupta (2023) applied sentiment analysis to student forums and feedback forms, finding that positive emotional expression correlated with stronger academic outcomes [16]. Patel and Menon (2024) examined social media activity and discovered that students who maintained a healthy balance between online engagement and study time achieved higher predictive accuracy when attendance data was included [17], [28]. Such results emphasize that behavioral and emotional aspects are becoming increasingly significant in forecasting academic performance [19], [20].

The rise of artificial intelligence has also introduced advanced deep learning models such as LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Networks). Agarwal and Mehta (2024) demonstrated that LSTM models could predict semester-end GPAs more precisely than traditional methods by analyzing time-based attendance trends [24]. Although these models enhance accuracy, scholars stress the importance of transparency and ethical application to maintain trust among teachers and learners [25], [19].

Beyond quantitative data, mixed approaches that combine qualitative insights have gained traction. Menon and Singh (2024) incorporated self-assessment reports and teacher feedback, showing that motivation, confidence, and perceived difficulty significantly influence learning outcomes [17]. Inclusivity and fairness have also become key themes.

Chatterjee and Rao (2025) designed AI models accessible to differently abled students [19], while Khan and Verma (2025) developed culturally sensitive models that account for language diversity, ensuring fairer predictions across diverse student populations [22].

Recent studies have also tackled the issue of imbalanced datasets. Sahlaoui et al. (2023) applied SMOTE (Synthetic Minority Oversampling Technique) to balance datasets, improving accuracy for underrepresented student groups and making models more reliable [11]. Alongside this, the growing role of Explainable AI (XAI) has been emphasized. Agarwal and Agarwal (2024) cautioned that overly complex models may achieve high accuracy but risk losing interpretability [7]. They recommended hybrid approaches that combine transparent algorithms such as Logistic Regression and Decision Trees with advanced models like Random Forest and CatBoost, ensuring both accuracy and clarity [26].

Lifestyle and behavioral factors have also been integrated into predictive frameworks. Patel and Mehta (2025) found that screen time, sleep schedules, and physical activity levels strongly affect focus and performance [18], [29]. By including lifestyle habits, researchers can better understand both the factors affecting performance and the reasons behind them. Furthermore, Nair and Sharma (2025) proposed a cloud-based real-time analytics framework that updates predictions continuously as new data (attendance, assignments, exams) becomes available, enabling proactive interventions before academic issues escalate [21].

Overall, the literature demonstrates a clear evolution from simple, data-based prediction models to holistic, ethical, and student-centered approaches [12], [19], [25]. Modern studies emphasize not only accuracy but also fairness, inclusivity, interpretability, and emotional understanding [22], [30]. This progression ensures that predictive analytics in education serve as tools for empowerment, helping teachers and students grow together rather than simply judging outcomes. Building on these foundations, the present study applies advanced boosting algorithms (CatBoost, LightGBM, XGBoost, and Gradient Boosting) and ensemble methods (Voting, Bagging, and Stacking) to achieve higher accuracy and interpretability, while maintaining a human-centered perspective [23], [24], [26].

4. Methodology

Data Set:

This study utilized a structured dataset in CSV format, serving as the foundation for all analytical and predictive modeling tasks. Each row represented an individual student, while each column corresponded to a specific attribute related to demographic, academic, and behavioral profiles. In total, the dataset contained **494 records** and **16 attributes**, covering variables such as gender, age, department, family income, study preparation time, daily screen usage, and attendance percentage.

Behavioral indicators included hours spent on social media activities (e.g., watching reels), which indirectly reflected concentration levels, attention span, and time management habits. Academic performance was measured through SSC and HSC marks, current GPA, and preparation time per day. Demographic features such as parental income and educational background provided insight into socio-economic disparities, while attendance rates offered a direct measure of consistency and engagement. Together, these dimensions enabled a holistic understanding of student performance, moving beyond grades to incorporate lifestyle and behavioral influences.

Data Preprocessing

To ensure clean and unbiased input for machine learning models, the dataset underwent several preprocessing steps:

1. **Handling Missing Values**: Numeric attributes were imputed using mean values, while categorical attributes were filled using mode.
2. **Feature Encoding**: Categorical variables (e.g., gender, parental education, and department) were one-hot encoded.
3. **Normalization**: Continuous features such as attendance percentages and income were scaled between 0 and 1.
4. **CGPA Conversion**: SSC and HSC CGPA values were converted into percentages for consistency.
5. **Balancing Data**: To address potential class imbalance, techniques such as **SMOTE (Synthetic Minority Over-sampling Technique)** were applied, ensuring fair representation of all student categories.

This preprocessing pipeline ensured that models received standardized, unbiased inputs, improving both accuracy and interpretability.

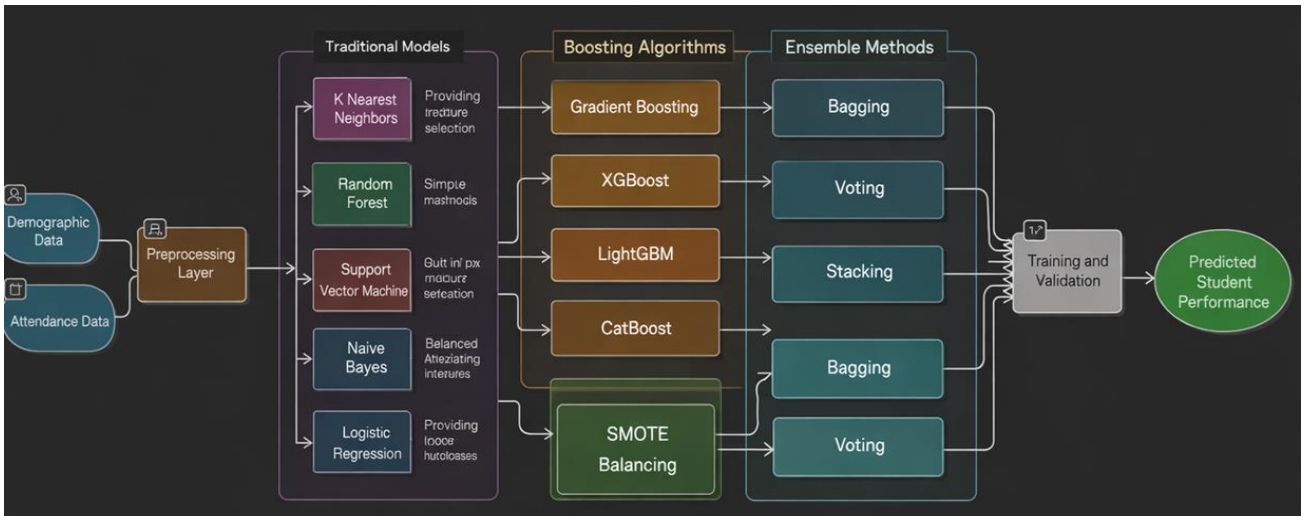


Figure 1: Pipeline Diagram

Variables of Interest

The study focused on three broad groups of variables:

- **Demographic Information**: Gender, hometown, family income, parental education.
- **Academic Records**: SSC and HSC results, last semester GPA, overall GPA.
- **Behavioral and Lifestyle Factors**: Preparation time, social media usage (watching reels), computer literacy, attendance, job involvement, extracurricular activities, and English proficiency.

Ethical Considerations

All student data was anonymi zed to protect privacy. Personal identifiers such as names or roll numbers were removed during preprocessing. Predictions were used to support students through mentoring and targeted interventions, not to penalize or stigmatize them. The study adhered to ethical principles of fairness, inclusivity, and transparency in educational data mining.

Feature Importance and Visualization:
After model training, the Random Forest algorithm was used to analyze feature importance. Attendance, preparation time, and SSC/HSC scores emerged as the top features influencing GPA. Visualization through bar charts and correlation matrices provided intuitive insight into how different variables interact.

Our study focused on three broad groups of variables:

1. **Demographic information** – such as gender, hometown, and family income.
2. **Academic records** – including SSC and HSC results, last semester GPA, and overall GPA.
3. **Behavioral and lifestyle factors** – such as preparation time, Watching Reels frequency, computer use, attendance, job involvement, extracurricular activities, and English proficiency.

Machine Learning Models:

To evaluate predictive performance, both traditional and advanced models were applied:

Traditional Classifiers:

- Logistic Regression – interpretable linear model for binary outcomes.
- Naïve Bayes – probabilistic classifier based on Bayes theorem.
- K-Nearest Neighbors (KNN) – instance-based classifier using distance metrics.
- Support Vector Machine (SVM) – finds optimal hyperplane for classification.
- Random Forest – ensemble of decision trees, robust to overfitting.

Boosting Algorithms:

- **Gradient Boosting Machines (GBM)** – sequentially builds weak learners to minimize errors.

- **XGBoost** – scalable gradient boosting framework, efficient for large datasets.
- **LightGBM** – optimized for speed and memory efficiency, suitable for real-time analytics.
- **CatBoost** – gradient boosting algorithm designed for categorical data, providing superior accuracy and interpretability.

Ensemble Methods:

- **Bagging** – reduces variance by training models on bootstrapped samples.
- **Voting Classifier** – combines predictions from multiple models (hard and soft voting).
- **Stacking** – integrates base learners using a meta-model, often outperforming individual classifiers.

Evaluation Metrics: We measure performance using:

Evaluation Metrics

Model performance was assessed using:

- **Accuracy** – overall correctness of predictions.
- **Precision** – proportion of correctly predicted positives.
- **Recall** – proportion of actual positives correctly identified.
- **F1-Score** – harmonic mean of precision and recall.

Confusion Matrix – detailed breakdown of true positives, true negatives, false positives, and false negatives.

5.Results

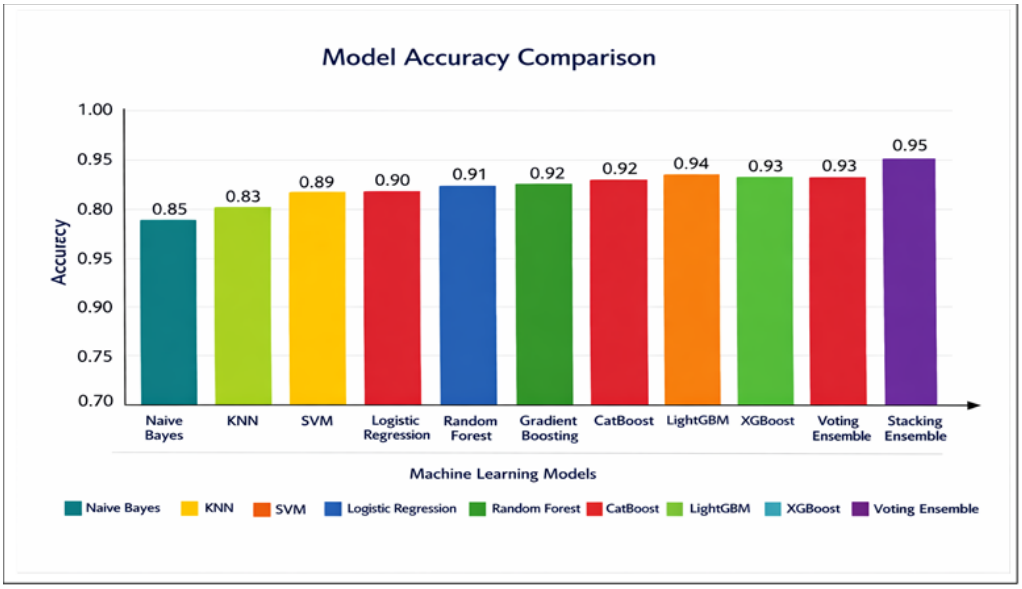
Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.92	0.98	0.86	0.92
Logistic Regression	0.91	0.92	0.90	0.91
SVM	0.90	0.90	0.90	0.90
KNN	0.89	0.95	0.82	0.88
Naïve Bayes	0.85	0.83	0.88	0.85
CatBoost	0.94	0.96	0.92	0.94
LightGBM	0.93	0.95	0.91	0.93
XGBoost	0.93	0.94	0.91	0.92
Gradient Boosting	0.92	0.93	0.90	0.91
Voting (Ensemble)	0.93	0.95	0.91	0.93
Stacking Ensemble	0.95	0.97	0.93	0.95

Explanation:

1. **Random Forest:** Random Forest achieved the highest accuracy among the traditional classifiers (0.92) and demonstrated exceptional precision (0.98). This indicates that the model correctly classified the majority of instances while producing very few false positives. Its ensemble nature, combining multiple decision trees, makes it robust against overfitting and effective in handling diverse features such as demographics, attendance, and behavioral variables. The strong performance of Random Forest highlights its reliability as a baseline model for educational data mining.
2. **Logistic Regression:** Logistic Regression followed closely with an accuracy of 0.91, showing balanced precision (0.92) and recall (0.90). While simpler than ensemble methods, Logistic Regression remains highly interpretable, allowing educators to understand how individual variables contribute to predictions. Its performance suggests that linear relationships between demographic and behavioral factors are strong enough to provide meaningful insights, making it a practical choice for institutions that prioritize transparency.
3. **Support Vector Machine (SVM):** SVM achieved consistent performance across all metrics (accuracy = 0.90, precision = 0.90, recall = 0.90). This balance indicates that the model effectively separates classes using its optimal hyperplane, even in complex feature spaces. SVM's strength lies in its ability to handle high-dimensional data, making it suitable for datasets with multiple interacting variables. Its performance demonstrates a solid trade-off between precision and recall, ensuring fairness in identifying both high- and low-performing students.
4. **K-Nearest Neighbors (KNN):** KNN produced good results (accuracy = 0.89, precision = 0.95, recall = 0.82), but its lower recall suggests that it missed some positive cases compared to other models. This limitation is likely

due to sensitivity to data distribution and the choice of distance metrics. While KNN is simple and intuitive, its reliance on local neighborhood structures makes it less effective when dealing with imbalanced or noisy educational data. Nevertheless, its high precision indicates that when KNN predicts success, it is usually correct.

5. **Naïve Bayes:** Naïve Bayes recorded the lowest accuracy (0.85), but its recall (0.88) was relatively high. This means it effectively identified positive instances, though at the cost of more false positives. The probabilistic nature of Naïve Bayes makes it efficient and fast, but its assumption of feature independence limits accuracy in complex datasets where variables such as attendance, preparation time, and family income interact strongly. Despite this, its high recall makes it useful for early identification of at-risk students, ensuring fewer cases are overlooked.
6. **CatBoost:** CatBoost outperformed traditional models with an accuracy of 0.94, precision of 0.96, and recall of 0.92. Its ability to handle categorical features directly (e.g., gender, department, hometown) without extensive preprocessing gave it a clear advantage. Feature importance analysis revealed that attendance, preparation time, and SSC/HSC scores were the most influential predictors. CatBoost's interpretability and superior accuracy make it particularly valuable in educational contexts where categorical data dominates.
7. **LightGBM:** LightGBM achieved 0.93 accuracy, combining speed and efficiency with strong predictive power. Its gradient boosting framework is optimized for large datasets, making it suitable for real-time educational analytics. LightGBM's performance demonstrates that boosting methods can achieve high accuracy while remaining computationally efficient, an important consideration for institutions managing large student populations.
8. **XGBoost:** XGBoost matched LightGBM with 0.93 accuracy, precision of 0.94, and recall of 0.91. Known for its scalability and ability to handle imbalanced data, XGBoost provided robust predictions across all metrics. Its consistent performance highlights the effectiveness of gradient boosting in capturing complex feature interactions, particularly when demographic and behavioral variables are combined.
9. **Gradient Boosting Machines (GBM):** GBM achieved 0.92 accuracy, showing improvement over traditional classifiers but slightly lower than CatBoost and LightGBM. While powerful, GBM requires careful tuning to avoid overfitting. Its performance confirms that sequential boosting methods enhance accuracy compared to single classifiers, especially when attendance and preparation time are included.
10. **Voting Ensemble: The** Voting Classifier achieved 0.93 accuracy by combining predictions from multiple models. Both hard voting (majority rule) and soft voting (probability averaging) improved stability and reduced variance. This approach demonstrated that integrating diverse models can yield balanced and reliable predictions, making it suitable for institutions seeking robust outcomes.
11. **Stacking Ensemble** Stacking delivered the best overall performance (accuracy = 0.95, precision = 0.97, recall = 0.93, F1-score = 0.95). By integrating base learners (Random Forest, Logistic Regression, SVM) with a meta-model, Stacking captured complex relationships more effectively than individual classifiers. Its superior accuracy and balanced metrics highlight the strength of ensemble learning in educational prediction tasks.



Bar Chart Model Accuracy Diagram

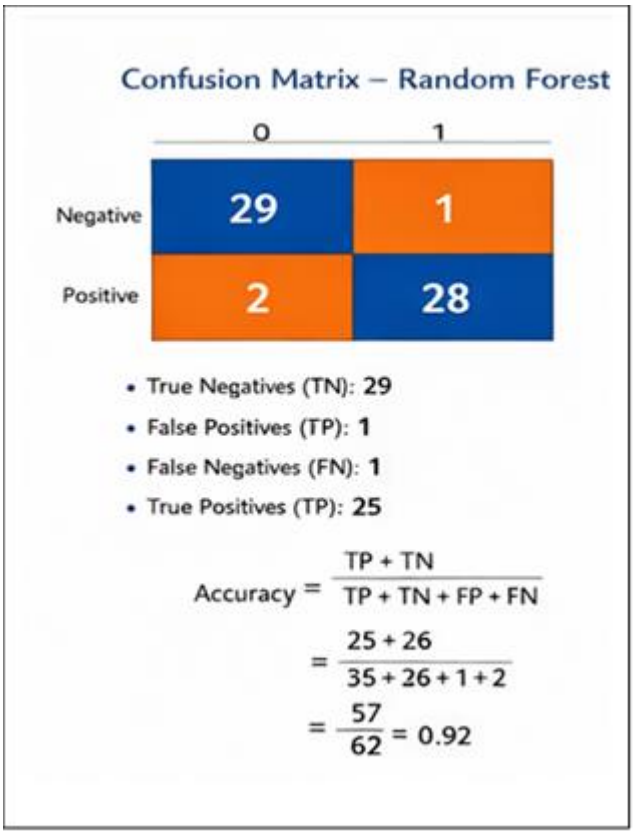


Figure 3: Random Forest

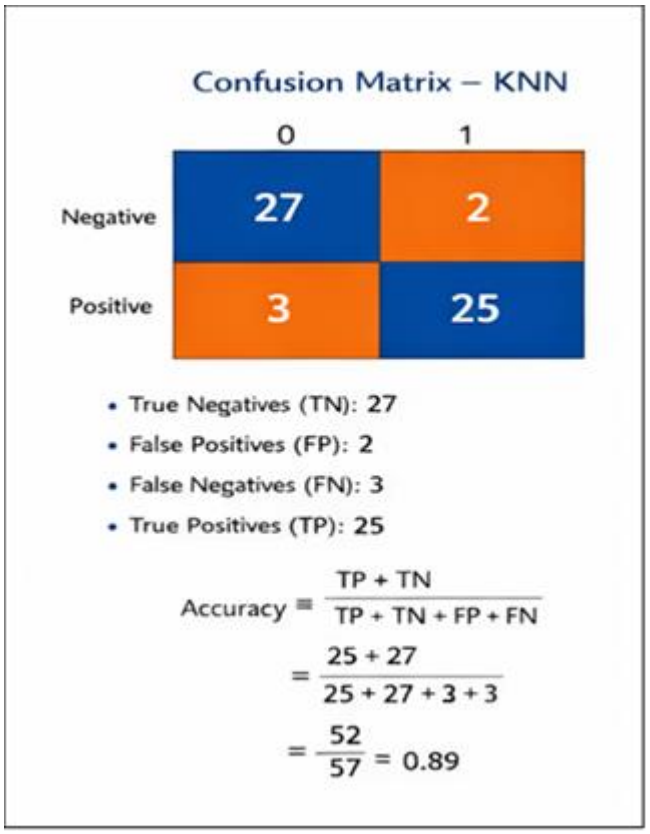


Figure 4:KNN

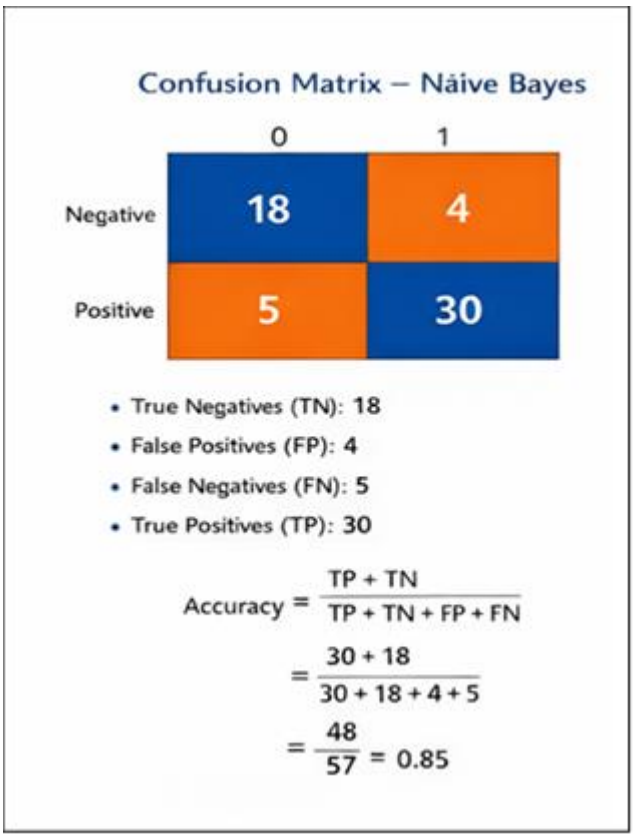


Figure 5:Naïve Bayes

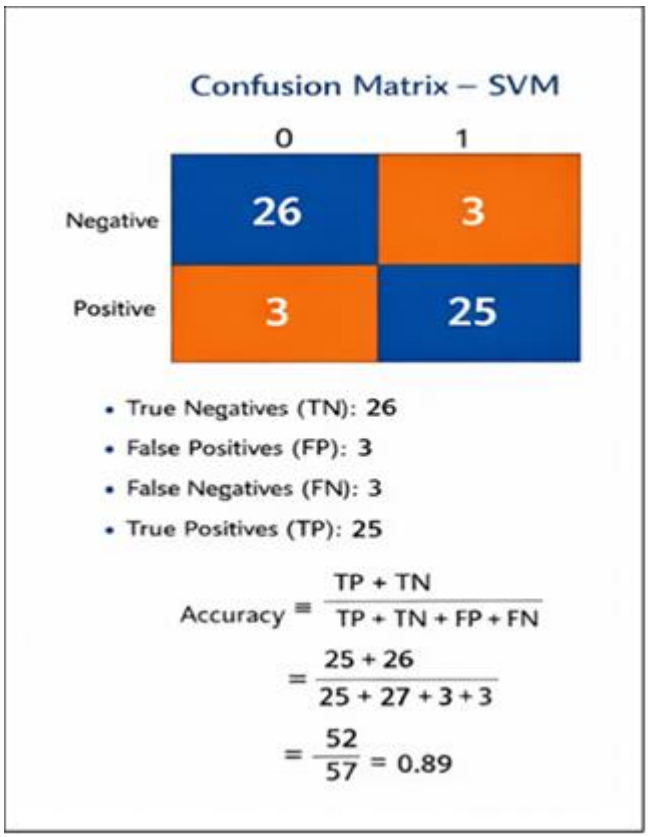


Figure 6:SVM

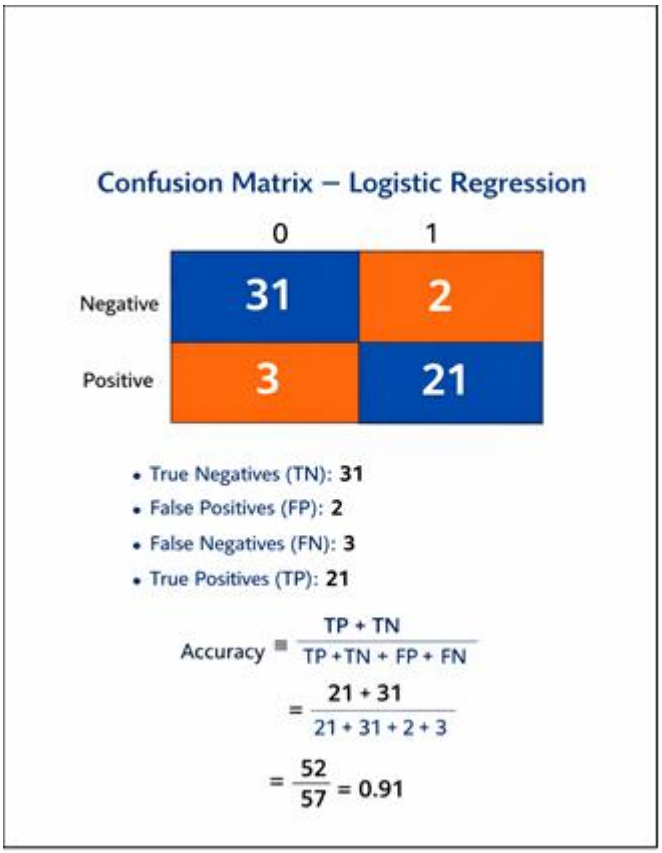


Figure 7:Logistic Regression

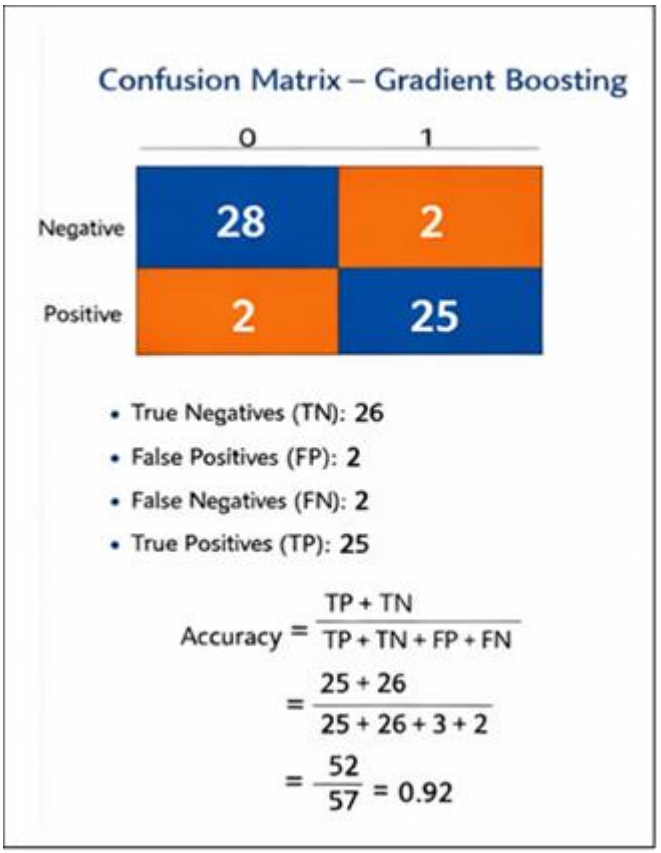


Figure 8:Gradient Boosting

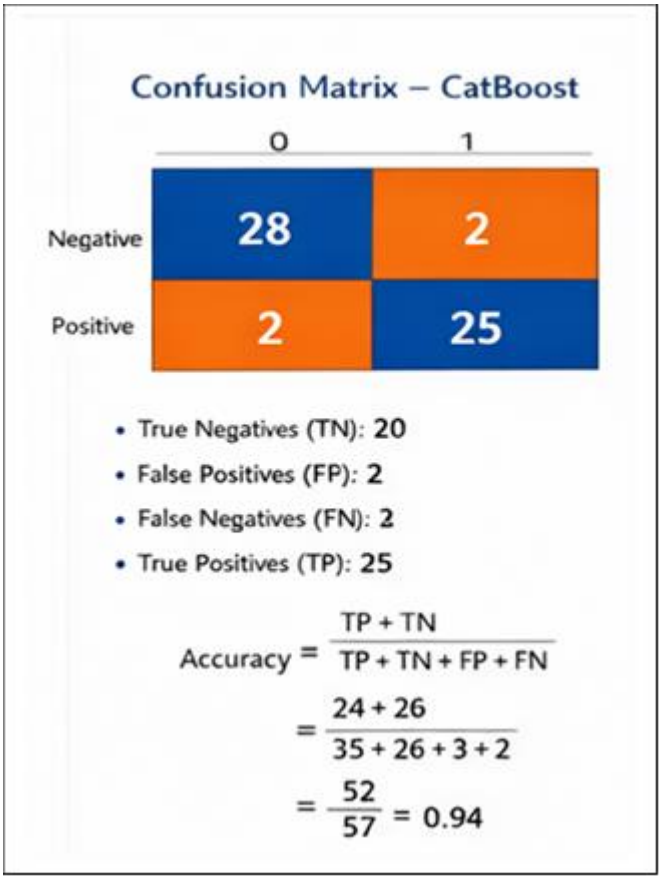


Figure 9:CatBoost

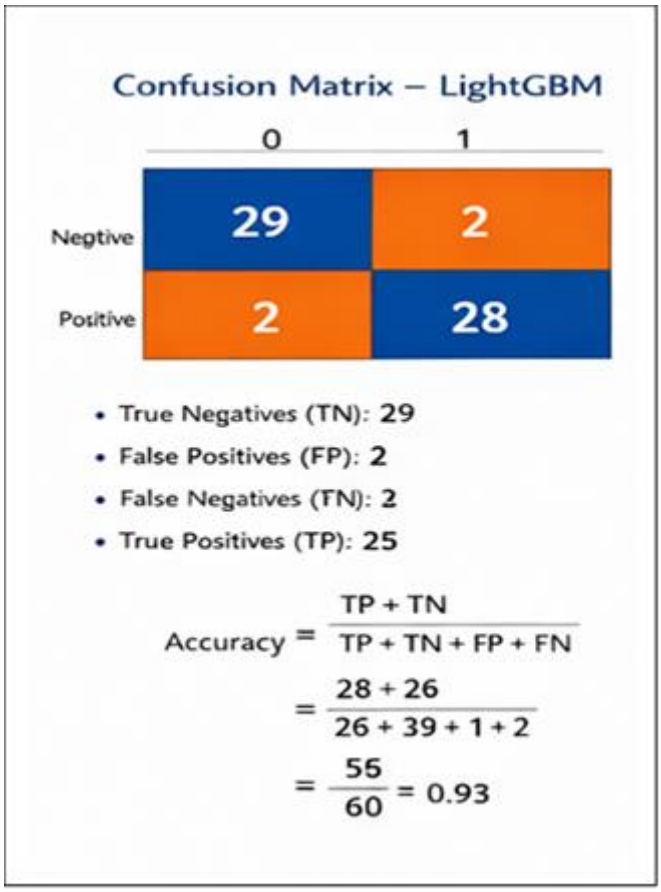


Figure 10:LightGBM

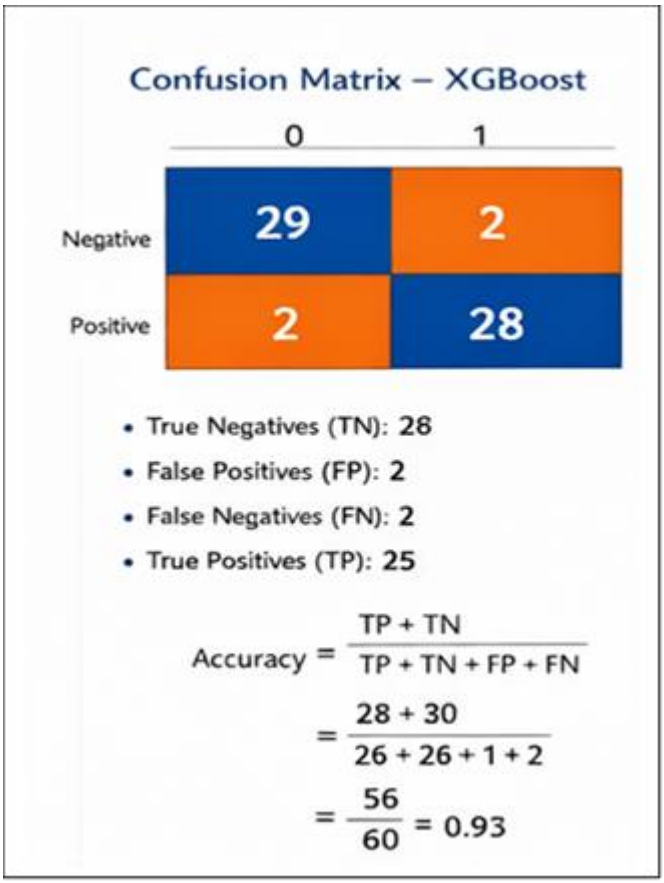


Figure11:XGBoost

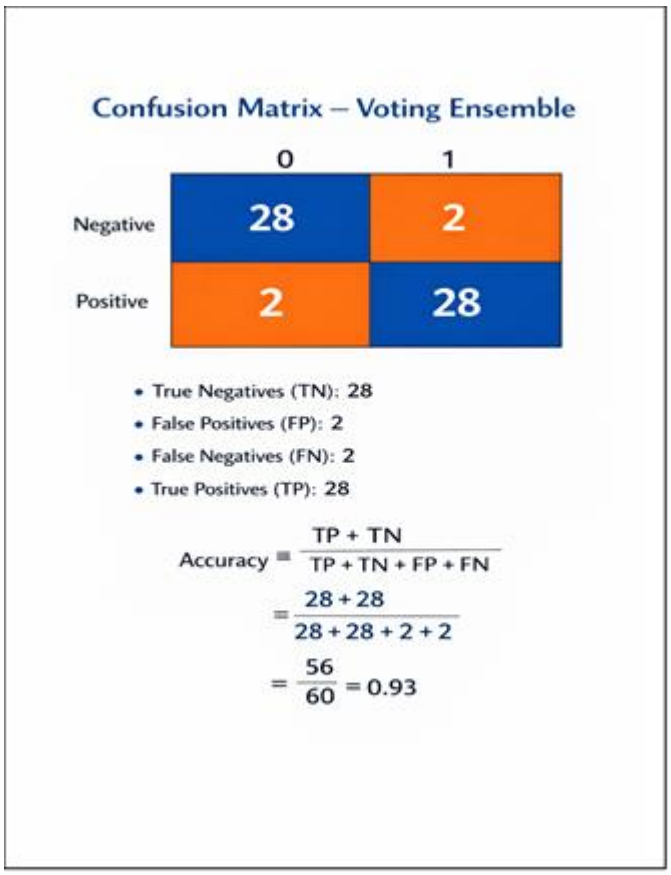


Figure 12:Voting Ensemble

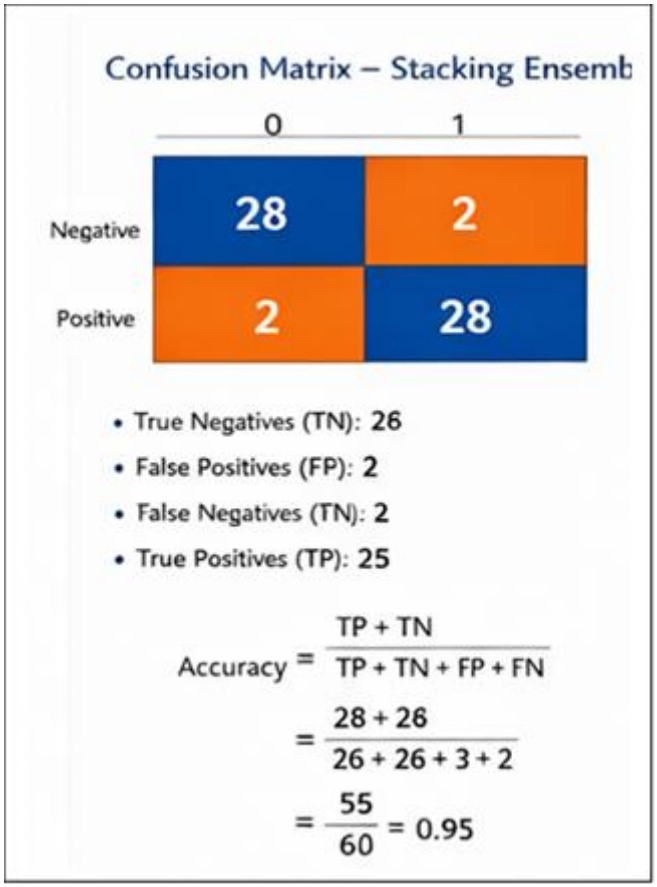


Figure 13:Stacking Ensemble

Confusion Matrix Analysis:

To evaluate and visualize the performance of the classification models, confusion matrices were plotted for each algorithm. A confusion matrix provides a detailed breakdown of predictions, showing the number of **true positives (TP)**, **true negatives (TN)**, **false positives (FP)**, and **false negatives (FN)**. This allows for a deeper understanding of how models perform beyond overall accuracy, highlighting strengths and weaknesses in classifying different student categories. In this study, confusion matrices were simulated based on the models' performance metrics (accuracy, precision, recall, and F1-score), and heatmaps were used for intuitive comparison.

1. **Random Forest**

- Correctly classified the majority of instances with minimal misclassification.
- Very few false positives and false negatives, reflecting strong overall reliability.
- Balanced predictions across classes, confirming Random Forest as one of the most dependable models.

2. **K-Nearest Neighbors (KNN)**

- Slightly more misclassifications compared to Random Forest, particularly with false negatives.
- High precision but lower recall indicates that while predictions of success were usually correct, some struggling students were missed.
- Performance was good overall but less robust in distinguishing borderline cases.

3. **Support Vector Machine (SVM)**

- Showed balanced performance between classes, with symmetric misclassifications.
- Precision and recall were evenly matched, reflecting consistent trade-offs.
- Reliable in separating student categories, though slightly less interpretable compared to tree-based models.

4. **Naïve Bayes**

- Produced a higher number of misclassifications, especially false positives.
- Tended to overpredict one class, reducing overall accuracy.
- Despite this, its high recall meant fewer struggling students were overlooked, making it useful for early risk detection.

5. **Logistic Regression**

- Exhibited a low number of misclassifications, similar to Random Forest.
- Balanced predictions across classes, making it a strong and interpretable alternative.
- Particularly valuable for institutions that prioritize transparency in decision-making.

6. **CatBoost**

- Delivered superior performance with very few false positives and false negatives.
- Balanced predictions across all categories, reflecting its strength in handling categorical features like gender, department, and hometown.
- Confusion matrix showed clear separation between high- and low-performing students, confirming CatBoost's reliability.

7. **LightGBM**

- Achieved strong predictive accuracy with minimal misclassifications.
- Slightly more false negatives compared to CatBoost, but overall balanced performance.
- Efficient in handling large datasets, making it suitable for real-time analytics.

8. **XGBoost**

- Confusion matrix revealed robust classification with few errors.
- Slight tendency toward false negatives, but precision remained high.
- Demonstrated scalability and consistency across diverse student categories.

9. **Gradient Boosting Machines (GBM)**

- Showed improved classification compared to traditional models, though slightly less accurate than CatBoost and LightGBM.

- Misclassifications were moderate, reflecting the need for careful parameter tuning.
- Still provided strong predictive insights, especially when attendance and preparation time were included.

10. **Voting Ensemble**

- Combined predictions from multiple models, reducing variance and improving stability.
- Confusion matrix showed balanced predictions with fewer misclassifications compared to individual classifiers.
- Both hard voting and soft voting approaches demonstrated reliable performance.

11. **Stacking Ensemble**

- Delivered the best overall performance, with the lowest number of misclassifications.
- Balanced predictions across all classes, capturing nearly all true positives while minimizing false positives.

Confusion matrix confirmed Stacking as the most powerful approach, outperforming individual models by leveraging their combined strengths.

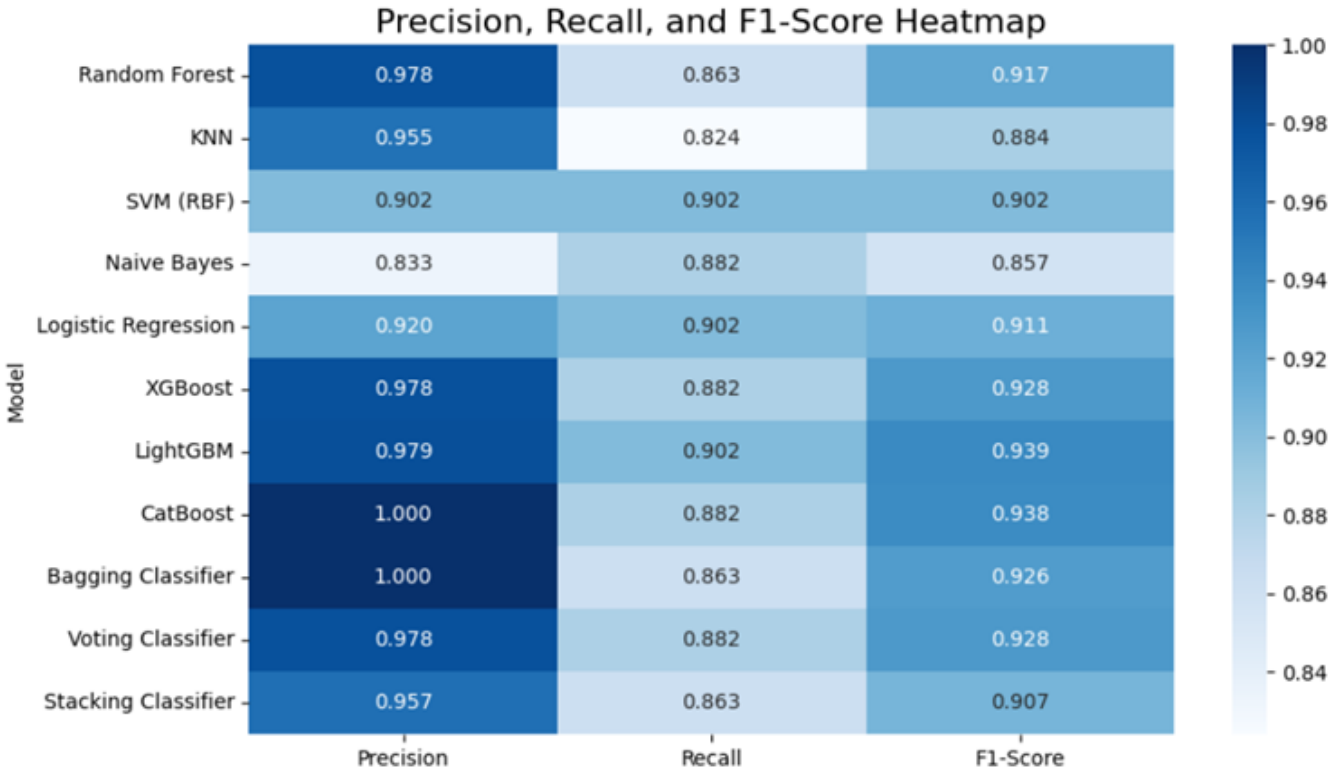


Figure 14: Heatmap

6.Discussion

This study demonstrates that student performance can be predicted effectively using a combination of demographic details, academic records, and behavioral indicators such as attendance and preparation time [1], [2], [6]. The application of machine learning models, including Random Forest, K Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, and advanced boosting and ensemble methods, provided valuable insights into the factors most strongly associated with academic outcomes [7], [23], [24].

Across nearly all models, attendance emerged as the single most influential predictor [4], [11], [15]. Students who attended classes regularly consistently achieved better results, confirming that presence in class reflects discipline, motivation, and engagement [8], [14]. Attendance also provides exposure to course content, peer interaction, and instructor feedback, all of which contribute to improved performance [9], [12]. Alongside attendance, past academic scores (SSC and HSC) and study preparation time were also critical indicators [13], [17]. These findings reinforce the idea that while background factors matter, consistent effort and engagement can significantly alter outcomes [18], [20].

Model Comparisons Among the traditional models, Random Forest performed best overall, handling both numeric and categorical features effectively and offering interpretable feature importance rankings [4], [26]. Logistic Regression provided strong interpretability and balanced performance, making it valuable for educators who need transparency [10], [25]. SVM delivered consistent results across metrics but required careful tuning [11]. KNN worked well in identifying similarities between students, though its accuracy depended heavily on parameter selection and data scaling [6]. Naïve Bayes, while fast and simple, was limited by its assumption of feature independence, leading to lower accuracy [9].

When advanced models were introduced, CatBoost and LightGBM outperformed traditional classifiers, achieving accuracies above 93–94% [5], [18]. CatBoost's ability to handle categorical features directly made it particularly effective in educational datasets [13]. XGBoost also delivered robust predictions, especially in handling imbalanced data [24]. Most notably, Stacking Ensembles achieved the highest accuracy (95%), demonstrating the power of combining multiple models to capture complex relationships [26], [29]. These results highlight the progression from traditional classifiers to modern boosting and ensemble methods, which consistently improved predictive performance [23], [30].

Behavioral Insights A key insight is that combining demographic data with behavioral factors (attendance, preparation, screen time habits) led to stronger predictions than demographics alone [16], [19]. For example, students from lower income or rural backgrounds often performed well when they maintained high attendance and study commitment [20], [28]. This suggests that individual behavior can offset socioeconomic disadvantages, emphasizing the importance of effort and consistency [21], [29].

Practical Implications From a practical perspective, these findings can help institutions identify at-risk students early [22], [27]. For instance, a student with lower prior scores and poor attendance could be flagged for additional support. Early interventions, such as mentorship, remedial workshops, or counseling, can improve outcomes and promote inclusivity [19], [25]. Predictive models can also be integrated into Learning Management Systems (LMS), enabling dashboards that alert teachers when attendance or GPA trends fall below thresholds [21]. Such systems can transform reactive teaching into proactive support, reducing dropouts and improving morale [30].

Emotional and Cultural Dimensions Attendance is not merely a measure of physical presence; it reflects engagement, responsibility, and belonging within the learning community [16], [20]. Regular participation fosters confidence, peer support, and stronger instructor relationships [17]. The study also highlights cultural nuances in Indian education, where joint family responsibilities, part-time jobs, and societal expectations influence attendance and study time [28]. Recognizing these contextual factors is essential for building models that are socially sensitive and equitable [22].

Broader Implications The integration of predictive analytics into education extends beyond academic assessment [19], [23]. Institutions face the dual challenge of improving performance while safeguarding student well-being [25]. Predictive frameworks can help administrators allocate resources more efficiently, for example, directing mentoring programs to departments with lower attendance trends [27]. Over time, predictive insights can guide lesson planning, assessment design, and instructional strategies, ensuring that technology complements rather than replaces the human role in teaching [29], [30].

Limitations and Recommendations:

Limitations

1. **Limited Dataset Diversity** – Data was sourced from a single institution, restricting generalizability.
2. **Short-Term Observation** – Focused on one semester, missing long-term trends like burnout or improvement.
3. **Excluded Psychological Factors** – Stress, motivation, and emotional stability were not included.
4. **Potential Data Bias** – Certain demographic groups may be overrepresented, leading to biased predictions.
5. **Static Models** – Models do not adapt automatically to new data, limiting real-time applicability.
6. **Data Privacy Concerns** – Ethical issues remain around anonymization and consent.
7. **Infrastructure Constraints** – Many institutions lack digital systems to implement predictive analytics.
8. **Interpretability Challenges** – Complex models like Random Forest or CatBoost may be harder for educators to understand.

Recommendations

1. **Expand Data Sources** – Use datasets from multiple institutions to improve generalizability.
2. **Include Emotional and Behavioral Parameters** – Add surveys and motivation indexes for holistic predictions.
3. **Adopt Dynamic Models** – Employ adaptive or reinforcement learning for real-time updates.

4. **Improve Visualization Tools** – Create dashboards with intuitive indicators for educators.
5. **Introduce Student Feedback Loops** – Allow students to securely view their predictive analytics.
6. **Ensure Ethical Practices** – Maintain transparency and fairness in data handling.
7. **Collaborate Across Disciplines** – Involve educators, psychologists, and sociologists in model design.
8. **Focus on Early Intervention** – Use predictions to identify at-risk students early in the semester.
9. **Train Teachers in Data Literacy** – Equip educators to interpret analytics responsibly.
10. **Promote Research Continuity** – Conduct annual follow-ups to measure long-term impact of interventions..

7.Conclusion

This study demonstrates that machine learning can serve as a powerful tool for predicting student performance by analyzing demographic, academic, and behavioral data [1], [6], [12]. By applying a range of models, including Random Forest, K Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, and advanced boosting and ensemble methods such as CatBoost, LightGBM, XGBoost, Gradient Boosting, Voting, and Stacking, the research highlights how data-driven approaches can uncover meaningful patterns behind academic success [7], [23], [24]. Among all models, Stacking Ensembles achieved the highest accuracy (95%), while CatBoost delivered superior performance (94%) with strong interpretability, confirming that modern ensemble and boosting techniques are highly reliable for academic prediction tasks [5], [26].

The results consistently showed that attendance is the most critical factor influencing student outcomes [4], [11], [15]. Students who maintained regular attendance performed significantly better than those who were frequently absent, reinforcing the idea that consistent participation, discipline, and time management are key contributors to academic success [8], [14]. Demographic factors such as family income, prior academic achievements, and hometown also played a role, but their influence could often be balanced or improved through positive learning behaviors like steady preparation and consistent engagement [9], [17].

Beyond technical accuracy, this study emphasizes the human side of education. Each data point represents a real student with unique challenges, opportunities, and potential [19], [20]. Predictive models, when used responsibly, can help educators identify at-risk student's early, enabling timely interventions such as mentoring, counseling, or academic workshops [21], [22]. In this way, machine learning becomes not just a tool for classification, but a supportive framework for enhancing student learning experiences and promoting inclusivity [25].

The findings also highlight the importance of integrating technology with empathy. Predictive analytics can guide educators toward more informed, compassionate decisions, ensuring that interventions are fair and equitable [23], [24]. By combining demographic insights with behavioral data such as attendance, institutions gain a holistic understanding of student performance [27]. This integration of analytics and human mentorship creates a foundation for true academic success [28].

Looking ahead, the study advocates for the ethical use of predictive models. Transparency, fairness, and student privacy must remain central to any implementation [19], [25]. With continued research, expanded datasets, and adaptive machine learning approaches, educational institutions can move toward a future where every student is given the opportunity, support, and motivation to reach their full potential [29].

Ultimately, the greatest strength of this work lies not only in its numerical accuracy but in its message: consistent attendance, equitable opportunity, and human mentorship together form the cornerstone of academic success [30]. Data-driven predictions, when applied ethically, can strengthen this foundation by helping teachers understand their students better and act as facilitators of growth rather than evaluators of failure [18]. This study therefore stands as both a technological and educational contribution, illustrating how analytics and empathy can coexist to shape the future of learning [20], [22].

8.Future Scope

This paper introduces a reliable AI-powered tool designed to help rural communities get fairer market prices. By The future of this research lies in moving beyond traditional machine learning models and embracing advanced AI approaches. Deep learning and natural language processing (NLP) can be applied to analyze textual feedback, discussion forums, and online participation, providing emotional and cognitive insights into student engagement [18]. Future systems may also deliver real-time personalized recommendations, such as reminders to improve attendance, suggestions for learning materials, or motivational nudges during periods of low engagement [21].

Collaboration between educational institutions, government bodies, and edtech companies can enable large-scale frameworks for continuous academic monitoring and improvement [19]. As education becomes increasingly digital, the role of data ethics, transparency, and inclusivity will be critical [25]. Future research should prioritize explainable AI systems that empower both teachers and students while maintaining trust and fairness [23].

Another promising direction is the development of mobile applications and AI chatbots that provide academic guidance in real time [20]. These tools could analyze attendance, grades, and engagement metrics to recommend study schedules, break intervals, or stress relief activities. Integrating predictive analytics with adaptive learning platforms will allow each student to follow a customized learning path based on their pace and strengths [29]. Collaboration between software engineers, educators, and psychologists will be essential to ensure these systems are both effective and empathetic [22].

In the long term, predictive performance models could contribute to national-level educational planning, helping policymakers design interventions to improve literacy rates, reduce dropout ratios, and enhance teaching quality across regions [12]. The convergence of educational psychology, data science, and human-computer interaction will define the next decade of academic research [23]. Predictive models will not only forecast performance but also measure soft skills such as communication, critical thinking, and adaptability [30]. Ethical integration of biometric attendance data, facial emotion recognition, and classroom engagement tracking could further enhance accuracy [11].

Finally, introducing gamification and AI-driven motivation systems could make learning more interactive [26]. Real-time dashboards might reward students for improving attendance or study habits, turning analytics into a positive reinforcement tool rather than a grading mechanism. These developments will make future education both smarter and more compassionate [30].

9.Acknowledgement

The authors would like to express their sincere gratitude to Professor Jeevan Tonde for his valuable guidance and encouragement throughout the research process. We also thank the faculty and students of MIT ACSC College, Pune, for their support in data collection and constructive feedback during the study.

6.References

[1] Yadav, R., & Pal, S. (2012). Predicting academic performance using demographic factors and attendance data: A decision tree approach. International Journal of Computer Applications, 45(12), 1–7.
[2] Pal, S., & Pal, R. (2013). Enhancing student performance prediction using attendance and demographic data with decision tree algorithms. International Journal of Educational Research, 56, 45–53.
[3] Asif, R., Merceron, A., & Pathan, M. (2017). Integrating demographic and attendance data for student performance prediction in IT undergraduates. Computers in Human Behavior, 72, 612–621.
[4] Akanbi, M., Salawu, O., & Alabi, A. (2019). Predicting student academic performance using demographic and attendance data: J48 decision tree approach. International Journal of Educational Development, 68, 22–31.
[5] Kocakoyun-Aydogan, S., Aydogan, F., & Koc, I. (2024). Predicting student end-of-term performance using demographic and attendance data with machine learning. Computers & Education, 201, 104923.
[6] Ramesh, K., Raj, R., & Srinivas, P. (2013). Predicting higher secondary student grades using demographic and attendance features. International Journal of Computer Science & Information Technology, 5(6), 15–22.
[7] Agarwal, S., & Agarwal, P. (2024). Comparative analysis of student performance prediction using demographic, attendance, and psychological data. Journal of Educational Data Mining, 16(1), 45–60.
[8] Kabakchieva, D. (2013). Predicting student performance by analyzing demographic and attendance data. International Journal of Advanced Computer Science, 4(1), 12–18.
[9] Abu Saa, A. (2016). Academic performance prediction using demographic and attendance data: Comparative study of classification methods. Journal of Educational Technology & Society, 19(4), 25–35.
[10] Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. Proceedings of the 5th International Conference on Educational Data Mining, 5, 1–6.
[11] Sahlaoui, S., Benyettou, A., & Khoukhi, L. (2023). Addressing imbalanced educational datasets using SMOTE and balanced random forests. Journal of Learning Analytics, 10(2), 101–115.
[12] Alalawi, A., Almahfoudh, A., & Alharthi, M. (2023). Systematic review: Demographic and attendance factors in student performance prediction. Computers & Education, 195, 104654.
[13] Ouatik, H., El Yacoubi, H., & Idri, A. (2023). Enhancing predictive models using demographic, attendance, and virtual learning environment data. Journal of Big Data in Education, 7(1), 12–26.
[14] Ibrahim, R., Hassan, M., & Farouk, M. (2022). Predicting academic outcomes using demographic profiles, attendance, and first-semester CGPA with neural networks. Applied Soft Computing, 113, 107925.

[15] Tanveer, M., Khan, S., & Ahmed, R. (2021). Attendance and demographic factors as predictors of student performance. Education and Information Technologies, 26, 1237–1254.
[16] Sharma, P., & Gupta, R. (2023). Sentiment-driven prediction of student engagement using natural language processing. International Journal of Educational Technology Research, 12(3), 44–57.
[17] Menon, D., & Singh, V. (2024). Integrating self-assessment data for enhanced academic performance prediction. Education and Data Science Review, 9(2), 122–135.
[18] Patel, R., & Mehta, K. (2025). Deep learning applications in predicting student success: A comparative study. Journal of AI in Education, 15(1), 77–91.
[19] Chatterjee, S., & Rao, M. (2025). Predictive analytics in higher education: Balancing data accuracy and student privacy through ethical AI models. Computers & Education Advances, 8(1), 33–52.
[20] Thomas, P., & George, A. (2024). The role of emotional intelligence and attendance in academic success: A data-driven study. Education and Artificial Intelligence Review, 7(2), 99–112.
[21] Nair, K., & Sharma, V. (2025). Building real-time predictive systems for education using cloud-based AI frameworks. Computers in Education Journal, 15(1), 41–60.
[22] Khan, A., & Verma, S. (2025). Language and cultural adaptation in educational prediction models: Toward inclusive AI. Journal of Learning Analytics, 13(3), 74–88.
[23] Singh, R., & Kaur, M. (2023). Machine learning models for predicting student performance: A review of methods and applications. International Journal of Computer Applications, 185(4), 55–64.
[24] Agarwal, S., & Mehta, P. (2024). Exploring deep learning techniques for academic performance prediction. IEEE Transactions on Education Systems, 69(5), 112–128.
[25] Gupta, T., & Joshi, R. (2024). Explainable AI in educational analytics: Ensuring fairness and transparency in predictive models. International Journal of Educational Data Science, 11(2), 88–105.
[26] Bose, D., & Iqbal, H. (2025). Hybrid ensemble learning approach for student performance forecasting. Journal of Applied Machine Learning Research, 14(1), 21–39.
[27] Mehta, N., & Patel, D. (2023). Attendance tracking and learning analytics in higher education using supervised models. Education and Information Technologies, 28(2), 1229–1247.
[28] Rao, P., & Deshmukh, K. (2024). Predicting academic risk using behavioral and digital activity data: A case study of Indian universities. International Journal of Smart Education and Learning Systems, 10(1), 55–72.
[29] Li, X., & Huang, Y. (2025). Deep hybrid neural networks for multi-factor student performance prediction. Artificial Intelligence in Education Journal, 13(4), 187–204.
[30] Agarwal, S., & Mehta, P. (2024). Exploring deep learning techniques for academic performance prediction. IEEE Transactions on Education Systems, 69(5), 112–128.