Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Comparative Analysis of Machine Learning Classification Algorithms and Hybrid Models for Student Performance Prediction

Ms. Pooja C. Soni
Research Scholar, Bhagwan Mahavir University, Surat, Gujarat, India

Dr. Hetal R. Modi
Bhagwan Mahavir College of Computer Application, Bhagwan Mahavir University, Surat

Dr. Pooja R. Negi
Department of Computer Application, Sascma BCA College, Surat

Abstract - This study focuses on the analysis and comparison of machine learning classification algorithms and hybrid machine learning models for predicting student academic performance. Educational Data Mining techniques are used to extract meaningful insights from student datasets. Various classification algorithms such as Decision Tree, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, and Random Forest are applied and compared with hybrid models. The performance of these models is evaluated using accuracy, precision, recall, and F1-score. The study aims to identify the most efficient model for early prediction of student performance. This research focuses on predicting student grades (A–F) using a hybrid machine learning framework **that combines academic scores, attendance records, and behavioral data. By using school-level datasets, the study ensures that the approach is practical and relevant to real-world education. The model applies classification techniques with structured data to improve accuracy and reliability. To measure performance, a** confusion matrix **and evaluation metrics like accuracy, precision, recall, and F1-score are used. Results show that most predictions match the actual grades, with strong diagonal values in the matrix, meaning the model is highly accurate across all grade categories. The study also includes** Explainability methods **to highlight which features matter most. Findings reveal that exam scores, attendance, and behavior strongly influence predictions. Compared to single-model approaches, the hybrid method is more stable and reduces errors. Overall, this work contributes to building** smart educational systems **that are scalable, interpretable, and useful for teachers. By identifying student performance levels early, educators can take timely action, provide personalized support, and improve learning outcomes.**

Key words: Educational Data Mining (EDM), Machine Learning (ML), Decision Tree, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Classification, Hybrid Models, Student Performance Prediction

1. INTRODUCTION:

In recent years, the application of Machine Learning (ML) and Educational Data Mining (EDM) has gained significant attention in predicting student academic performance. Educational institutions generate large volumes of student-related data, including academic records, attendance, and behavioral patterns. Analyzing this data helps in identifying at-risk students, improving learning outcomes, and supporting academic decision-making. Hybrid machine learning approaches have emerged as powerful techniques for improving prediction accuracy by combining the strengths of multiple algorithms. A hybrid ML framework integrating multiple classifiers has shown improved prediction performance compared to individual models [1]. Similarly, combining different learning techniques in hybrid approaches enhances prediction accuracy and robustness in educational datasets [2]. Furthermore, EDM-based models play an important role in designing effective prediction systems. Proper data preprocessing and feature selection significantly improve model performance, making structured educational data highly useful for prediction tasks[3].On the other hand, several studies have focused on single classification algorithms. Comparative analysis of different classifiers shows that model performance varies depending on dataset characteristics [4]. Data mining techniques have also been successfully

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

applied to real-world educational datasets to predict student outcomes [5]. Additionally, evaluation of multiple ML algorithms highlights the importance of selecting appropriate models based on the nature of the problem and data [6].Despite these advancements, most existing studies rely on either single models or basic hybrid approaches, with limited focus on complex student relationships and interpretability. Therefore, there is a need for advanced hybrid frameworks that not only improve prediction accuracy but also provide explainable insights for educators.

1.1 Current Challenges in Predicting Student Performance

Predicting student performance involves several pressing challenges that researchers and educators must navigate. Poor data quality and incomplete records often weaken model accuracy, while issues of bias and fairness risk reinforcing socio-economic inequalities. Advanced models may achieve high accuracy but lack interpretability, making it difficult for teachers to trust or apply insights effectively. Ethical concerns around privacy and data governance further complicate adoption, especially when sensitive student information is involved. Additionally, weak generalizability means models trained in one institution often fail in another due to differences in curricula and demographics. Limited feature engineering also restricts the inclusion of psychological and motivational factors that strongly influence learning outcomes. Together, these challenges highlight the need for hybrid, transparent, and ethically responsible approaches to ensure predictive models are both effective and equitable.

Generalizability is a major challenge in student performance prediction. A model trained on one university's data may not work well in another because grading systems, cultural factors, and teaching styles differ. This makes it hard to apply the same model everywhere. To solve this, researchers need to build robust, adaptable, and scalable frameworks that can handle diverse educational environments effectively.

1.2 Role of AI/ML in Education

AI and ML are transforming education by enabling personalized learning, intelligent tutoring, automated assessments, and data-driven decision-making, but they also raise challenges around fairness, privacy, and transparency. In education, AI/ML systems have been successfully deployed for:

- **Personalized Learning:** Adaptive systems tailor content to each student's pace and style; ML algorithms analyze performance data to recommend targeted exercises. Example: Intelligent tutoring systems that adjust difficulty based on student progress.
- **Student Performance Prediction:** AI models identify at-risk students early through attendance, grades, and engagement data. Helps institutions design timely interventions to improve outcomes.
- **Automated Assessment:** AI supports grading of assignments, quizzes, and even essays using NLP. Reduces teacher workload and provides faster feedback to students.
- **Learning Analytics:** ML uncovers hidden patterns in student behavior and learning pathways. Enables evidence-based curriculum design and teaching strategies.
- **Intelligent Tutoring Systems:** AI-driven tutors simulate one-on-one teaching. Provide instant feedback, hints, and explanations to guide learners.

1.3 Global Adoption of AI/ML in Academic Prediction

- **United States** Widely adopted in higher education for dropout-risk modeling, grade forecasting, and adaptive learning systems. Strong emphasis on fairness-aware algorithms **and** Explainable AI (XAI) to ensure transparency.
- **United Kingdom** Focuses on learning analytics and personalized tutoring systems**.** Universities use AI to track engagement and provide early interventions for struggling students.
- **India** Rapidly expanding adoption in higher education, especially for exam performance prediction and adaptive e-learning platforms. Government initiatives encourage AI integration in schools and universities.
- **China** Heavy investment in AI-driven education platforms. Uses ML for large-scale exam scoring, student monitoring, and predictive analytics. Strong focus on scalability due to massive student populations.

2. RELATED WORK:

Over the past two decades, research on student performance prediction has evolved from traditional statistical models to advanced AI-driven frameworks. Early studies used regression and decision trees, while modern work applies machine learning and ensemble models for higher accuracy. Recent progress in educational data mining

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

integrates behavioral and emotional data, expanding predictive scope. Yet, key research gaps remain in fairness, interpretability, and cross-institution adaptability.

2.1 Traditional Approaches in Student Prediction

Traditional Approaches in Student Prediction focused on statistical and rule-based methods before the rise of machine learning. Researchers commonly used linear regression and logistic regression to model relationships between grades, attendance, and demographic factors. Decision trees and Bayesian models were also applied to classify students into performance categories. These methods offered simplicity and interpretability but struggled with complex, nonlinear patterns in educational data.

2.2 Machine Learning Approaches

Machine learning approaches have significantly advanced student performance prediction by handling complex, nonlinear relationships in educational data. Techniques such as Support Vector Machines, Random Forests, and K-Nearest Neighbors are widely used to classify students into performance categories or forecast academic outcomes. These models outperform traditional statistical methods by learning hidden patterns from large datasets and adapting to diverse variables like demographics, attendance, and behavioral data.
Additionally, Neural Networks and Deep Learning variants have been explored for complex feature extraction, though interpretability remains a challenge. Overall, machine learning approaches provide higher accuracy and scalability, but issues of bias, fairness, and Explainability continue to be critical research concerns.

2.3 Ensemble and Hybrid Models

Ensemble and hybrid models represent a major advancement in student performance prediction, combining multiple algorithms to improve accuracy and robustness. Ensemble methods such as bagging, boosting, and stacking aggregate predictions from diverse classifiers (e.g., decision trees, SVMs, neural networks) to reduce variance and bias. These approaches consistently outperform single models by leveraging complementary strengths.
Hybrid models integrate statistical techniques with machine learning, or combine supervised and unsupervised learning, to capture both structured academic data (grades, attendance) and unstructured behavioral data (LMS interactions, social engagement). They are particularly effective in handling heterogeneous educational datasets and improving generalizability across institutions.
Despite their success, challenges remain in interpretability, computational complexity, and scalability. Future research emphasizes explainable ensemble frameworks that balance predictive power with transparency for educators.

2.4 Recent Advances in Educational Data Mining (EDM)

Baranikumar E et al. [1] developed a hybrid Random Forest + KNN model using marks, attendance, stress, and technology use. It provides explainable insights to help teachers identify struggling students early. **Al Tameemi et al. [2]** proposed a hybrid clustering + classification approach (FDN, RF, DT) that achieved higher accuracy and better identification of at-risk students compared to traditional methods. **Jayasree et al. [3]** introduced BPNN-SPA, combining academic, demographic, and behavioral data. It outperformed SVM and RDC, though sensitive to noisy data, with future work focusing on larger datasets. **Muhammad Alghobiri [4]** compared Naïve Bayes, C4.5, and SVM across ten datasets. Found Naïve Bayes best for small data, C4.5 for structured data, and SVM most reliable overall. **Han Xue et al. [5]** built a multi-output hybrid ensemble model predicting homework, experiments, midterm, and final scores. Sequential prediction improved early intervention and exam outcomes. **Dr. Altaf Hussain Abro et al. [6]** evaluated ML models and found XGBoost most accurate, with GPA, attendance, and continuous assessment as strong predictors for early warning systems.**Chen et al. [7]** – Applied Random Forest to classify students into tiers for adaptive instruction. Improved prediction accuracy and reduced failure rates. **Nitin Ramrao Yadav et al. [8]** showed ML methods (SVM, DT, Naïve Bayes) effectively identify at-risk students using academic, attendance, and behavioral data. **Jitendra Agrawal et al. [9]** combined DT, Naïve Bayes, and SVM with smoothing to enhance prediction accuracy and support timely interventions. **Vimarsha K et al. [10]** proposed a hybrid intelligence model integrating ML with teacher expertise, improving accuracy and trust in identifying at-risk students.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

2.5 Research gap

Despite significant advancements in student performance prediction using machine learning and educational data mining, several limitations still exist in the current literature:

1. **Reliance on Single Models:** Most studies depend on individual an ML model, which limit prediction accuracy and reduces robustness.
2. **Basic Hybrid Combinations:** Hybrid approaches often use simple algorithm pairings without fully exploiting complex data relationships.
3. **Focus on Tabular Data:** Current research emphasizes structured datasets, overlooking relational and similarity-based student data.
4. **Dataset Limitations** Studies often rely on small or specific datasets, restricting generalizability across diverse contexts.

3. PROBLEM STATEMENT AND OBJECTIVE:

Predicting student performance is complex, and traditional machine learning models often struggle with accuracy and Overfitting on large, diverse datasets. This study compares traditional algorithms with hybrid approaches to identify the most effective techniques for educational data mining.
This study's objective is to create a predictive model which uses Machine Learning methods and objective metrics to evaluate performance.

- To analyze student academic data for performance prediction.
- To study and compare existing machine learning models.
- To develop an improved hybrid model for better prediction accuracy.
- To enhance model performance using preprocessing and feature selection.
- To classify student performance into grade categories (A–F).
- To evaluate the model using standard performance metrics.

4. METHODOLOGY:

This study follows a systematic methodology to predict student performance using both traditional machine learning classification algorithms and hybrid models. The methodology consists of the following phases:

4.1. Data Collection

The datasets mentioned in the comparison table are adopted from previously published research works such as UCI Student Dataset, OULAD Dataset, and other educational datasets used by different authors. These datasets are widely used benchmark datasets in Educational Data Mining for evaluating and comparing machine learning models.

Attribute	Description
Roll No	Student Identifier
Gender	Male / Female / Others
Dept	B.Sc CS, BA English, B.Sc Math, BCA, BA Defence, B.Com, etc.
Marks (10th, 12th)	Marks obtained by students in school exams
Marks (Sem1–Sem4)	Semester exam marks
Submit Assignment	Yes / No
Study Time	{1, 2, 3} (hours per day)
Use Library	Poor / Average / Good
Use Lab	Poor / Average / Good
Attend Seminars	Poor / Average / Good
Attendance %	{90%, 80%, 70%}
Economic Problem	Yes / No
Health Issue	Yes / No
Family Problem	Yes / No
Psychological Problem	Yes / No

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Mobile Usage	{1, 2, 3} (hours per day)
Sleeping Hours	Poor / Average / Good
Use Internet	Yes / No
Employment Status	Part-time jobs / Full-time studies
Travel Time	Home to Institution
Speaking Language	Tamil / English / Others
Parent Qualification/Work	Mother/Father qualification and occupation
Family Income	{50k, 1L, 1.5L, 2L}
Family Problem	Yes / No

Table 1: Student Data Set Description

4.2. Data Preprocessing

- **Data Cleaning & Missing Value Handling:** Remove duplicate and inconsistent data, and handle missing values using mean, median, or mode.
- **Data Transformation & Encoding:** Convert categorical data into numerical form and normalize/standardize features for uniformity.
- **Feature Selection & Data Splitting:** Select important features and divide the dataset into training and testing sets for model building.
- **Normalization:** Min–Max normalization scales the data into a fixed range [0,1], improving model performance and stability.

$$\text{Xnorm} = \frac{\text{X} - \text{Xmin}}{\text{Xmax} - \text{Xmin}} \quad \text{...............} \quad (1)$$

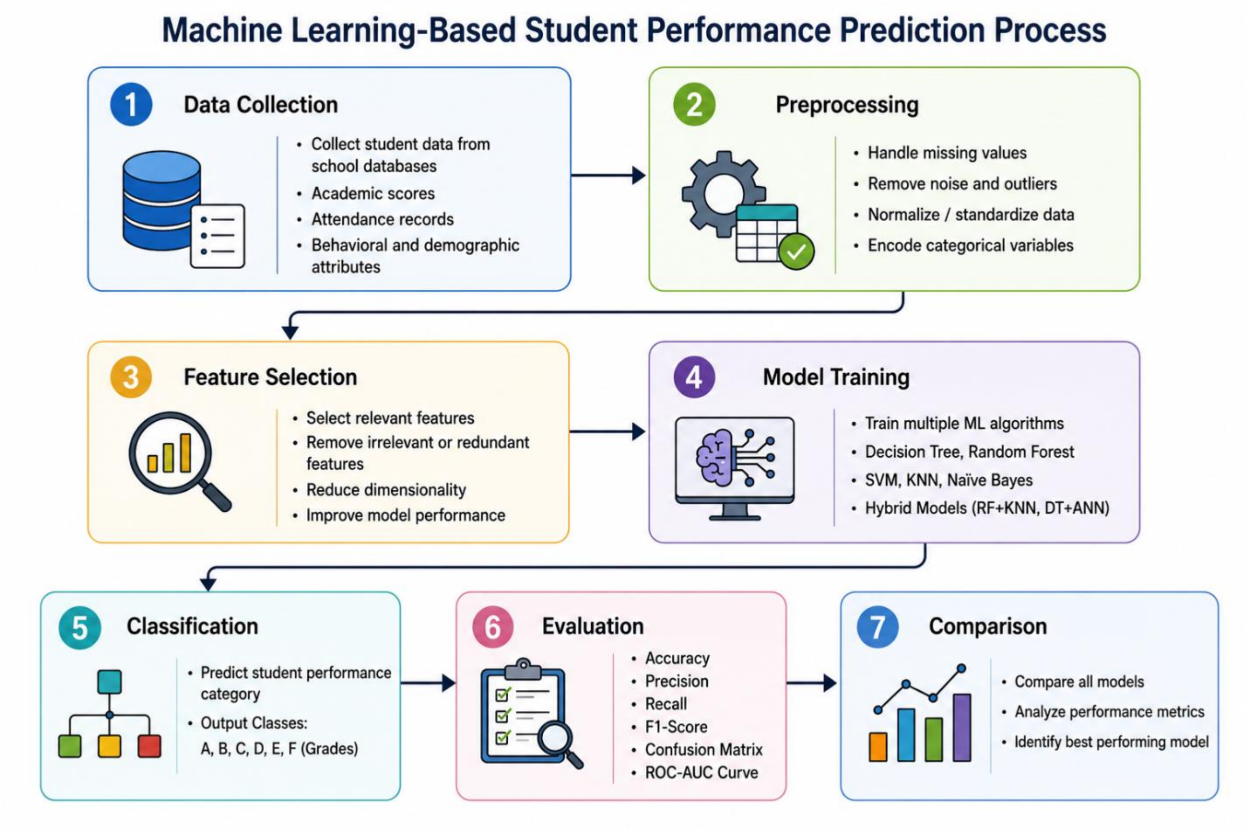

Figure: 1 "Workflow Framework for Machine Learning-Based Student Performance Prediction"

4.3 Classification

Students are classified based on their grades, behavior, emotions, and use of technology. After preparing the data, the hybrid model places them into three categories: High Performers, who achieve strong results, join activities, and show confidence and motivation; Average Students, who maintain moderate performance and balanced behavior but may need extra guidance; and At-Risk Students, who struggle with low marks, poor attendance, stress, or weak digital involvement. This classification helps schools give the right support such as mentoring, remedial classes, or counseling before problems grow worse.

4.4 Evaluation Matrix

In this work, to evaluate the performance of the classification models, multiple standard evaluation metrics are used, including Accuracy, Precision, Recall, and F1-Score.

Accuracy: Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total number of instances.

$$\text{Accuracy}=\frac{TP+TN}{TP + TN + FP + FN} \quad \ldots\ldots\ldots\ldots\ldots \quad (2)$$

Precision: Precision measures the accuracy of positive predictions made by the model. It indicates how many of the predicted positive instances are actually correct.

$$\text{Precision} =\frac{TP+TN}{TP + FP} \quad \ldots\ldots\ldots\ldots\ldots \quad (3)$$

Recall: Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify all relevant instances.

$$\text{Recall} = \frac{TP+TN}{TP + FP} \quad \ldots\ldots\ldots\ldots\ldots \quad (4)$$

F1-Score: F1-Score is the harmonic mean of Precision and Recall, providing a balanced measure that considers both false positives and false negatives.

$$\text{F1}= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \ldots\ldots\ldots\ldots\ldots \quad (5)$$

Where, TP denotes True Positive, TN denotes True Negative, FP denotes False Positive, and FN denotes False Negative.

5. RESULT AND ANALYSIS:

Experimental Results & Analysis In order to predict student's performance using the mentioned machine learning techniques, we experimented on several institutional data sets. To evaluate student's performance, we collected data of students from different institutions. We collected data i.e. name, student enrollment no., gender, and all subjects' marks. When we evaluated result sets using Machine Learning algorithms i.e. Combination of Random Forest and K-Nearest Neighbors, Combination of Decision Tree and Artificial Neural Networks, Random Forest, Support Vector Machine, K-Nearest Neighbors and Naive Bayes, we got result as mentioned in figure 1.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

6. MODEL PERFORMANCE COMPARISON:

Combination of Random Forest and K-Nearest Neighbors, Combination of Decision Tree and Artificial Neural Networks, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naive Bayes algorithms were used to evaluate student's performance. The success rate of classification algorithms is often evaluated using the F-measure, also known as the F1-score. The following outcomes were obtained from testing the prediction models on training and validation datasets:

Algorithm / Model Used	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
NB	78%	77%	76%	76.50%
KNN	80%	79%	78%	78.50%
DT	82%	81%	80%	80.50%
SVM	85%	84%	83%	83.50%
RF	88%	87%	86%	86.50%
DT+ ANN	91%	90%	89%	89.50%
RF + KNN	93%	92%	91%	91.50%

Table 2. Comparison of Machine Learning Classification and Hybrid Models

The comparison illustrates the percentages of accuracy, precision, recall, and F1-score for several machine learning algorithms, including Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Networks, K-Nearest Neighbors, Naive Bayes, and hybrid models. Among these, the hybrid algorithm combining Random Forest and K-Nearest Neighbors achieved the highest accuracy of 93%, demonstrating its effectiveness in capturing complex patterns within the dataset. In contrast, Naive Bayes recorded the lowest accuracy, highlighting its limitations in handling intricate relationships in student performance data. Overall, the results emphasize that hybrid models outperform individual algorithms in predictive tasks within educational data mining.

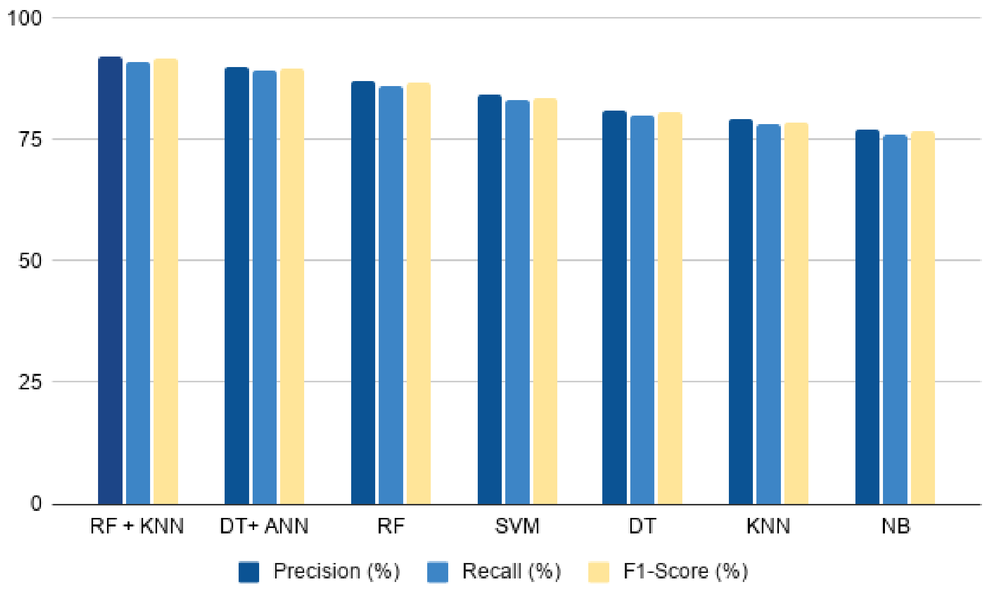

Figure: 2 Student Performance Comparison of Algorithms

The outcome from Table 1 was represented graphically in above Figure 2. In this case, hybrid algorithm demonstrated the highest F1-Score rate (91.5%), Precision rate (92%) and Recall rate (91%).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

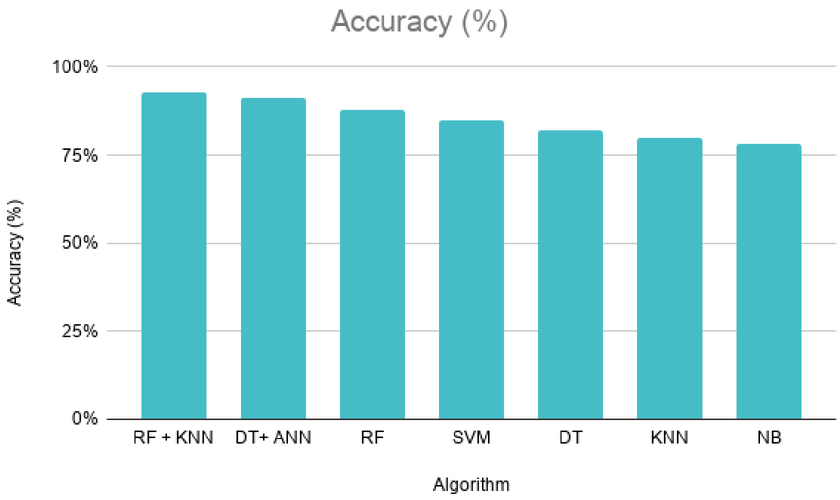

Figure: 3 Predictive models' accuracy rate

The combination of RF + KNN achieves the highest accuracy (93%), followed closely by DT + ANN(91%), indicating that integrating multiple algorithms helps capture complex patterns in educational data more effectively.

7. CONCLUSION:

This study explains a hybrid machine learning model that predicts student grades (A–F) by combining different types of information such as exam scores, attendance records, and behavioral factors. Instead of using just one method, the hybrid approach mixes multiple techniques to make predictions more reliable. To check how well the model works, researchers used a confusion matrix**,** which compares actual grades with predicted grades. Most of the results fall along the diagonal of the matrix, meaning the predictions matches the real grades, showing high accuracy.

The model performs well across all grade categories, especially in the middle ranges, where students often have similar performance patterns. Small mistakes happen when the model confuses neighboring grades, like predicting a "B" instead of an "A," because those categories are close in performance. Compared to traditional single-model methods, this hybrid approach is more stable, reduces errors, and gives more consistent results.

Overall, the system is a practical tool for early identification of student performance levels. By spotting potential issues early, teachers and institutions can provide timely support, guide students better, and improve academic outcomes. This makes the model not just a technical achievement but also a valuable educational resource for improving student success.

The study shows that while the model works well, there are ways to make it even better in the future. One improvement is to use bigger and more varied school datasets so the model can work for different groups of students. The model could be turned into a real-time decision support system or built into school management platforms, so teachers can continuously monitor students and give personalized learning advice.

8. REFERENCE:

[1] Baranikumar, E., & Naveen, A. (2025). A hybrid machine learning framework for student performance prediction integrating academic, demographic, behavioural, technology access, and psychological indicators in higher education. *Tuijin Jishu/Journal of Propulsion Technology, 46*(4), 1173–1182. ISSN: 1001-4055.

[2] Al-Tameemi, G., Xue, J., Ali, I. H., & Ajit, S. (2024). A hybrid machine learning approach for predicting student performance using multi-class educational datasets. *Procedia Computer Science, 238*, 888–895. https://doi.org/10.1016/j.procs.2024.06.108

[3] R, J., & Selvakumari, S. (2023). Design of a prediction model to predict students' performance using educational data mining and machine learning. Engineering Proceedings, 59*(25), 25. https://doi.org/10.3390/engproc2023059025

[4] Alghobiri, M. (2018). A comparative analysis of classification algorithms on diverse datasets. Engineering, Technology & Applied Science Research, 8(2), 2790–2795. https://doi.org/10.5281/zenodo.1234567

[5] Xue, H., & Niu, Y. (2023). Multi-output based hybrid integrated models for student performance prediction. *Applied Sciences, 13*(9), 5384. https://doi.org/10.3390/app13095384

[6] Abro, A. H., Irfan, M., Shah, S. S. A., & Malookhani, J. (2025). Performance evaluation of machine learning algorithms for predicting student academic outcomes. *ASSAJ*, 4(2), 3530–3541. ISSN (e) 3007-3138, ISSN (p) 3007-312X.

[7] Chen, Y., Sun, J., Wang, J., Zhao, L., Song, X., & Zhai, L. (2025, February 5). *Machine learning-driven student performance prediction for enhancing tiered instruction* (arXiv:2502.03143). arXiv. https://arxiv.org/abs/2502.03143

[8] Yadav, N. R., & Deshmukh, S. S. (2023). Prediction of student performance using machine learning techniques: A review. In S. Tamane et al. (Eds.), *Proceedings of the International Conference on Advances in Machine Intelligence and Data Analytics (ICAMIDA 2022)* (ACSR, Vol. 105, pp. 735–741). Atlantis Press. https://doi.org/10.2991/978-94-6463-136-4_63736

[9] Agrawal, J., Thakur, S., & Tiwari, V. K. (2024). Enhanced multi-channel classification framework for accurate student performance prediction. Procedia Computer Science, 235, 447–457. https://doi.org/10.1016/j.procs.2024.04.044

[10] Ka, V., Prakash, S. P. S., Krinkin, K., & Shichkina, Y. A. (2024). Student performance prediction: A co-evolutionary hybrid intelligence model. Procedia Computer Science, 235, 436–446. https://doi.org/10.1016/j.procs.2024.04.043

[11] Shahiri, A. M., Husain, W., & Rashid, N. A. (2018). A review on predicting student performance using data mining techniques. Procedia Computer Science, 72, 414–422.

[12] Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *UCI Repository*.

[13] Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics Dataset (OULAD). *Scientific Data, 4*(1), 170171. https://doi.org/10.1038/sdata.2017.171

[14] Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students' final GPA using decision trees. *International Journal of Information and Education Technology, 6*(7), 528–533. https://doi.org/10.7763/IJIET.2016.V6.745

[15] Kotsiantis, S. (2012). Use of machine learning techniques for educational purposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review, 37*(4), 331–344. https://doi.org/10.1007/s10462-011-9234-x (doi.org in Bing)

[16] Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model machine learning techniques. *Journal of Applied Research in Higher Education, 10*(2), 175–188. https://doi.org/10.1108/JARHE-09-2017-0113 (doi.org in Bing)

[17] Sweeney, M., Lester, J., & Rangwala, H. (2016). Next-term student performance prediction: A recommender systems approach. *Journal of Educational Data Mining, 8*(1), 17–50.

[18] Hussain, M., Zhu, W., Zhang, W., & Abidi, S. (2018). Student academic performance prediction using data mining techniques. *Sustainability, 10*(9), 1–23. https://doi.org/10.3390/su10093290