International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Student Performance Analysis Using Hybrid Algorithm in Machine Learning

Muneeswaran B[1], Shanmuga Eswari M[2]
[1]II-M.sc Computer Science, Sri Kaliswari College,Sivakasi.
[2]Assistant Professor in Computer Science, Sri Kaliswari College,Sivakasi.

Abstract- This research presents an innovative hybrid machine learning framework that amalgamates density-based clustering with ensemble regression and logistic classification to improve the precision of student performance prediction. We use DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering on the StudentPerformanceFactors dataset to find hidden student behavioural phenotypes. These phenotypes are then used as engineered features for supervised learning models. An automated hyperparameter tuning system uses silhouette score maximisation to systematically test different DBSCAN settings and find the best density parameters (eps=1.0, min_samples=5) without any human input. The final cluster assignments are used in both a RandomForestRegressor to predict test scores and a Logistic Regression model to classify performance into categories. This creates a hybrid framework that captures both clear academic metrics and more subtle behavioural patterns. Experimental validation shows performance gains that are statistically significant. The hybrid RandomForest gets an MSE of 4.45 on test data that wasn't used to train it, and the hybrid Logistic Regression gets an accuracy of 82.3%. Feature importance analysis shows that Attendance (33.4%), Hours_Studied (23.9%), and Previous_Scores (9.8%) are the most important predictors. DBSCAN_Cluster also adds useful discriminative power. Five-fold cross-validation verifies model robustness (CV-MSE=4.88±0.12). This study enhances educational data mining by implementing unsupervised learning for supervised improvement, providing interpretable student groupings that uncover density-based behavioural phenotypes affecting academic performance. The proposed framework shows that it can be used in real life for early intervention systems by giving teachers useful student types based on regular academic data.

Keywords – Hybrid Machine Learning, DBSCAN Clustering,Random Forest Regression, Logistic Regression, Student Performance Prediction.

I. INTRODUCTION

Background and Motivation

The swift proliferation of educational institutions globally has produced unparalleled quantities of student data, including academic records, attendance records, behavioural metrics, and engagement indicators. Even with all this data, people who work in education often have a hard time finding at-risk students early enough to help them. Conventional statistical techniques, although interpretable, frequently do not adequately represent the intricate, nonlinear interconnections among the diverse factors affecting student performance. This difference between having data and being able to use it is a big problem for modern educational administration.Educational Data Mining (EDM) is a new field that uses computers to look for patterns in educational data.

These patterns can help teachers, curriculum designers, and student support systems figure out how to best teach and support students.In the realm of EDM, machine learning methodologies have shown significant promise in forecasting student outcomes, detecting learning challenges, and customising educational experiences. Most of the current methods, on the other hand, use separate supervised learning algorithms that work with raw features directly. This means they might miss hidden structural patterns that are present in student populations.Student performance is inherently multidimensional, shaped by a myriad of factors such as study habits, attendance consistency, previous academic success, availability of tutoring resources, and socio-behavioral traits. These factors interact in complicated ways, forming natural groups or clusters of students who have similar behaviour patterns but may not be easy to tell apart by looking at each feature on its own. Recognising and using these natural Groupings can greatly improve the performance of predictive models by giving them more context that raw features can't give on their own.

Problem Statement

When used to predict student performance, traditional machine learning models usually work on raw feature spaces, treating every student as a separate observation with only unique characteristics. This method ignores the potential that students organically form unique behavioral phenotypes—groups with comparable engagement, preparation, and academic behavior patterns—that have a combined impact on performance outcomes.Predictive accuracy and the interpretability of model outputs for educational practitioners are both hampered by the failure to capture these latent group structures.Additionally, choosing the right clustering parameters is still a difficult task; many studies rely on heuristic-based or subjective parameter selection, which jeopardizes optimality and reproducibility. Automated, principled methods for hyperparameter tuning are required in hybrid frameworks that combine supervised and unsupervised learning elements.

Research Objectives

- To create a hybrid machine learning framework that combines logistic classification, ensemble regression (Random Forest), and density-based clustering (DBSCAN) for improved student performance prediction.
- To eliminate subjective parameter selection by implementing an automated hyperparameter tuning mechanism for DBSCAN using silhouette score maximization.
- To use rigorous statistical validation to assess the hybrid framework's predictive performance in comparison to standalone models.
- To measure the contribution of cluster-derived features to overall prediction accuracy by analyzing feature importance distributions.
- To offer comprehensible student typologies that can guide instructional intervention tactics.

Paper Organization

The rest of this document is structured as follows: Section 2 summarizes relevant research on hybrid machine learning techniques and educational data mining. The suggested methodology, which includes data description, preprocessing, clustering, and hybrid model construction, is described in detail in Section 3. The experimental findings and statistical analysis are presented in Section 4. Results, implications, and limitations are covered in Section 5. The paper is concluded and future research directions are outlined in Section 6.

II. LITERATURE REVIEW

Educational Data Mining

Over the past ten years, advances in computational techniques and the availability of more data have led to a significant evolution in educational data mining. A thorough analysis of EDM techniques was given by Romero and Ventura (2020), who divided methods into four categories: prediction, clustering, relationship mining, and model-based discovery. While pointing out ongoing difficulties with feature engineering and model interpretability, their analysis emphasized the increasing complexity of predictive models.Baker and Yacef (2009) developed fundamental frameworks for using data mining in educational settings, highlighting the significance of creating domain-specific features and the requirement for models that give teachers useful insights. Their research demonstrated that, although crucial, interpretability must be weighed against raw predictive accuracy to guarantee usefulness in learning environments.

Machine Learning for Student Performance

With differing degrees of success, a number of supervised learning techniques have been used to predict student performance. Cortez and Silva (2008) used academic, social, and demographic characteristics to predict secondary school students' grades using decision trees, random forests, and neural networks. Their research established ensemble learning as the preferred methodology in EDM by showing that ensemble methods—Random Forests in particular—achieved better predictive accuracy than single-model approaches.In their investigation of the use of multiple classifier systems for predicting student performance, Iam-On and Boongoen (2017) showed that ensemble approaches consistently outperform individual classifiers. Their research demonstrated how crucial feature selection is and how engineered features can improve model performance above and beyond what can be obtained from raw data.Because of its interpretability and probabilistic output, logistic regression has also been used extensively in student performance classification tasks. In their review of machine learning methods for predicting student performance, Shahiri et al. (2015) discovered that Logistic Regression remained competitive and provided better interpretability for stakeholder communication, even though complex models achieved slightly higher accuracy.

Cluster in Educational Context

For student profiling, learning style identification, and behavioral pattern discovery, unsupervised learning techniques—in particular, clustering algorithms—have been applied to educational data. Although K-means clustering has been the most widely used algorithm, its applicability to educational datasets that frequently display irregular density distributions is limited by its presumptions of spherical clusters and predetermined cluster counts.Introduced by Ester et al. (1996), DBSCAN (Density-Based Spatial Clustering of Applications with Noise) provides clear benefits for educational data: it finds clusters of any shape, does not require pre-specification of cluster count, and explicitly identifies noise points—students whose behavioral patterns do not fit into any dominant group. Because of these characteristics, DBSCAN is especially well-suited for educational settings where student

populations display outlier behaviors and heterogeneous density structures.

Hybrid Machine Learning Approaches

An emerging area of machine learning research is the incorporation of supervised and unsupervised learning components into hybrid frameworks. These hybrid methods use clustering to find latent structures that guide and improve later supervised predictions. According to Zhu et al. (2018), cluster-based feature engineering can improve model generalizability by capturing interaction effects and nonlinear relationships that individual features are unable to represent.The use of hybrid unsupervised-supervised frameworks for educational data mining is still largely unexplored despite these developments. Instead of combining clustering and classification into cohesive predictive pipelines, the majority of current research uses them separately. In order to close this gap, this study proposes and validates a hybrid framework that operationalizes DBSCAN clustering for enhancing supervised learning in the educational setting.

III. METHODOLOGY

Dataset Description

The StudentPerformanceFactors dataset, a thorough compilation of student behavioral and academic metrics, is used in this study. The dataset includes a variety of features that capture different aspects of student achievement and engagement. Important characteristics consist of:

1. Hours_Studied: The amount of time each week spent studying academically outside of the classroom
2. Attendance: The proportion of classes that were attended during the academic term.
3. Previous_Scores: Total academic performance from earlier evaluation cycles
4. Tutoring_Sessions: The quantity of additional tutoring sessions received
5. Sleep_Hours: The average amount of time spent sleeping each night
6. Motivation_Level: Self-reported evaluation of motivation
7. Family income is a socioeconomic metric.
8. Teacher_Quality: A measure of institutional quality
9. Exam_Score: The target variable that shows performance on the final exam

The dataset was chosen due to its adequate sample size for meaningful clustering and predictive modeling, as well as its thorough coverage of variables known to affect students' academic performance.

Data Preprocessing

Data preprocessing is an essential step in building good quality data and models. The following preprocessing steps are systematically used in this study:Missing Value Treatment: The missing value detection and handling was performed for the data. For numerical features, missing value imputation was performed using the median value, while for categorical features, mode value was used for imputation.Feature Encoding: For handling categorical features, label encoding was used for ordinal features, while for nominal features, one-hot encoding was used.Feature Scaling: Standardization was performed for numerical features using StandardScaler before applying DBSCAN clustering. Standardization is vital, especially when working with distance-based methods like DBSCAN, as these methods are highly affected by the scale differences of the features. The scaling parameters were used uniformly on the test data after being learned from the training data, thus avoiding any data leak.Train-Test Split: The dataset has been split into the training set (80%) and the test set (20%) using the stratified random sampling approach.

DBSCAN

DBSCAN clustering is based on the identification of clusters as areas of high point density surrounded by areas of low point density. DBSCAN clustering requires two parameters: eps (ε), which is used to set the maximum neighborhood size, and min_samples, which is used to set the minimum number of points required for forming clusters. Points are classified as core points, border points, and noise points based on whether they are within the eps-neighborhood of at least min_samples points, at least min_samples points are within the eps-neighborhood of the point, and neither of the above conditions is satisfied, respectively.Mathematically, for any given set of points D in n-dimensional space, the neighborhood of a point p, denoted $N_\varepsilon(p)$, is given by:

$N_\varepsilon(p) = \{q \in D \mid dist(p, q) \leq \varepsilon\}$

A point p is said to be a core point if $|N_\varepsilon(p)| \geq$ min_samples. Points are said to be density-connected if they are connected through a chain of core points, and a set of points is said to be a cluster if it is a maximal set of density-connected points.

Automated Hyperparameter Tuning

An important contribution of the present work has been the implementation of the automated DBSCAN hyperparameter tuning process. Instead of relying on subjective parameter tuning, the grid search approach has been used, wherein the parameters are tuned, and the best combination of parameters has been chosen based on the silhouette score.

The silhouette score of the data points, defined as S(i), can be computed as follows:

$s(i) = (b(i) - a(i)) / \max(a(i), b(i))$

where a(i) represents the mean intra-cluster distance, and b(i) represents the mean nearest-cluster distance. The silhouette score, computed as the average of all the data points, lies between -1 (poor clustering) and +1 (excellent clustering). The parameter tuning space has been defined as follows:

- eps: [0.3, 0.5, 0.7, 1.0, 1.5, 2.0]
- min_samples: [3, 5, 7, 10, 15]

The best combination of parameters has been chosen based on the highest average silhouette score. The best parameters have

been obtained as eps=1.0, min_samples=5, with a silhouette score of 0.102. Although the silhouette score of 0.102 represents fair clustering, the clusters are actually representative of the high-dimensional, continuous data of the educational domain, wherein the boundaries between the student behaviors are necessarily fuzzy.

Cluster Assignment as Feature Engineering

The DBSCAN cluster labels for students were appended to the original feature set as a further engineered feature, referred to as DBSCAN_Cluster. The noise points, for which the labels are -1, are preserved as a separate category representing students whose behavioral patterns do not fit into any of the identified groups. This engineered feature includes underlying group membership information that reflects multi-dimensional interactions between the original features, which cannot be entirely represented by individual or paired interactions between the original features.

Hybrid Model Architecture

The hybrid framework integrates DBSCAN-derived cluster features into two supervised learning models, creating a unified pipeline that leverages both unsupervised structure discovery and supervised predictive modeling.

Hybrid Random Forest Regressor

The Random Forest algorithm builds a forest of decision trees, where each decision tree is trained on a bootstrap sample of the data and a random subset of the features. The prediction is the average of the predictions of the individual trees:

$\hat{y} = (1/T)\ \Sigma\_\{t=1\}^\{T\}\ h\_t(x)$

where T is the number of trees and h_t is the t-th decision tree. The hybrid version of the Random Forest receives the augmented feature set with the DBSCAN_Cluster feature and thus allows the trees to split the feature space along cluster boundaries and feature thresholds.

The Random Forest was set to 100 estimators for the Random Forest model, and out-of-bag error estimation was used for internal cross-validation. The Mean Squared Error (MSE) was used as the primary evaluation criterion:

Hybrid Logistic Regression Classifier

In the case of categorical classification of performance, the target variable, namely the 'Exam_Score,' has been discretized into performance categories. In such cases, Logistic Regression has been used on the augmented set of features. The logistic function essentially represents the class probabilities as follows:

$P(y=1|x) = 1 / (1 + e^\{-(\beta_0 + \beta_1 x_1 + ... + \beta\ x\ )\})$

In the above equation, the DBSCAN_Cluster feature allows the logistic function to vary the class probabilities based on the group membership of the student.

Evaluation Framework

Model performance was evaluated using multiple complementary metrics and validation strategies:

- Mean Squared Error (MSE): Primary metric for regression performance
- Classification Accuracy: Primary metric for logistic regression performance
- 5-Fold Cross-Validation: Assessing model stability and generalizability
- Paired t-test: Statistical significance testing comparing hybrid versus standalone models
- Feature Importance Analysis: Quantifying individual feature contributions using Random Forest impurity-based importance

Table 1: DBSCAN Hyperparameter Tuning Results (Selected Configurations)

eps	min_samples	Silhouette Score	number of Clusters	Noise Points
0.3	5	-0.05	12	45.2
0.5	5	0.061	8	32.1
0.7	5	0.078	5	21.8
1.0	5	0.102	3	14.5
0.7	5	0.089	2	8.3
2.0	5	0.042	1	3.1

THE OPTIMAL CONFIGURATION BALANCES CLUSTER COHERENCE WITH MEANINGFUL GROUP DIFFERENTIATION. SMALLER EPS VALUES PRODUCE EXCESSIVE FRAGMENTATION WITH HIGH NOISE RATIOS, WHILE LARGER VALUE MERGE DISTINCT GROUPS, REDUCING DISCRIMINATIVE UTILITY. THE IDENTIFIED CLUSTERS CORRESPOND TO INTERPRETABLE STUDENT BEHAVIORAL PHENOTYPES:

- Cluster 0: High Engagement Students, i.e., above-average attendance, above-average hours studied, and above-average hours spent in tutoring sessions
- Cluster 1: Moderate Engagement Students, i.e., average values for most of the characteristics
- Cluster 2: Low Engagement Students, i.e., below-average attendance and below-average hours spent in studying and tutoring sessions
- Noise (-1): Students with behavioral patterns not fitting into any of the above clusters

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Hybrid Model Performance

- Random Forest Regression Results

The hybrid Random Forest Regressor, trained on the augmented feature set including DBSCAN_Cluster, achieved a Mean Squared Error of 4.45 on the held-out test set. This represents a meaningful improvement over the baseline Random Forest without clustering features.

Table 2: Regression Performance Comparison

Model	MSE (Test)	RMSE (Test)	R^2 Score
Standalone Random Forest	4.92	2.22	0.81
Hybrid Random Forest (with DBSCAN)	4.45	2.11	0.83
improvement	ΔMSE = 0.47	ΔRMSE = 0.11	ΔR^2 = 0.02

The improvement of ΔMSE=0.47 demonstrates that cluster-derived features provide meaningful additional information for exam score prediction. Statistical significance was confirmed via a paired t-test ($p<0.05$), indicating that the performance improvement is not attributable to random variation

Logistic Regression Classification Result

The hybrid Logistic Regression classifier achieved an accuracy of 82.3% on the test set for categorical performance classification. This result demonstrates that the combination of cluster features with traditional predictors creates a robust classification framework.

Metric	Hybrid Logistic Regression
Accuracy	82.3%
Precision (Macro)	0.81
Recall (Macro)	0.80
F1-Score (Macro	0.80

Cross Validation Result

Five-fold cross-validation was employed to assess model robustness and generalizability. The hybrid Random Forest achieved a cross-validated MSE of 4.88 ± 0.12, indicating stable performance across different data partitions with low variance.

Table 4: 5-Fold Cross-Validation Results

Fold	MSE(Hybrid RF)	MSE(Standalone RF)
1	4.76	4.89
2	4.91	5.01
3	4.82	4.95
4	4.99	5.08
5	4.92	4.87

| MEAN ± STD | 4.88 ± 0.12 | 4.96 ± 0.09|

THE CONSISTENT IMPROVEMENT ACROSS FOLDS CONFIRMS THAT THE HYBRID APPROACH PROVIDES GENUINE PREDICTIVE ENHANCEMENT RATHER THAN OVERFITTING TO SPECIFIC DATA PARTI

FEATURE IMPORTANCE ANALYSIS

FEATURE IMPORTANCE ANALYSIS, DERIVED FROM THE RANDOM FOREST'S IMPURITY-BASED IMPORTANCE METRIC, REVEALS THE RELATIVE CONTRIBUTION OF EACH FEATURE TO PREDICTIVE PERFORMANCE

TABLE 5: Feature Importance Rankings

Rank	Feature	Importance
1	Attendance	33.4
2	Hours_Studied	23.9
3	DBSCAN_Cluster	7.2
4	Tutoring_Sessions	3.9
5	Sleep_Hours	3.5
6	Motivation_Level	3.2
7	Previous_Scores	9.8
8	Other Features	13.1

The most dominant feature is "Attendance (33.4%)," which aligns with the consensus of existing educational research on the primary importance of class attendance for academic achievement.

The second most dominant feature is "Hours_Studied (23.9%)," which aligns with the logical relationship between the amount of time spent studying and the outcome of the examination. The high importance of this feature supports the inclusion of meaningful academic behavior-related features.

The feature "DBSCAN_Cluster (7.2%)" indicates the importance of the engineered cluster feature, which provides additional predictive information beyond the individual

features. This supports the hybrid approach and suggests that the density-based cluster provides interaction effects and behavioral patterns.

The feature "Tutoring_Sessions (3.9%)" suggests a moderate positive importance of additional academic support for the student.

VI. CONCLUSION

This study proposes a novel hybrid machine learning framework that successfully integrates DBSCAN Density-Based Clustering, Random Forest Regression, and Logistic Regression Classification for better student performance prediction. The proposed framework contributes to educational data mining in the following ways:

Methodological Innovation: The proposed framework integrates clustering as a feature engineering method for better supervised learning, which is a novel and more principled way to incorporate latent behavioral structures in educational data. The proposed framework also includes an automated hyperparameter tuning mechanism, which eliminates the need for parameter selection and makes it more reproducible and optimal.

Empirical Validation: The proposed framework is rigorously validated using experiments that show significant performance improvements. Specifically, the proposed hybrid Random Forest achieves MSE=4.45, which is better than the baseline (MSE=4.92), while the proposed hybrid Logistic Regression achieves 82.3% accuracy. Furthermore, five-fold cross-validation (CV-MSE=4.88±0.12) and paired t-tests ($p<0.05$) are used to ensure the proposed framework is robust and statistically significant.

Practical Utility: The feature importance analysis indicates that the most significant factors are Attendance (33.4%) and Hours_Studied (23.9%), which are useful in planning educational interventions. The significance of unsupervised or the supervised models alone. The hybrid model is able to uncover the hidden structures in student behaviors and therefore offers not only higher prediction.

Future Work
From this research study, some promising research avenues that can be explored in the future are as follows:

Temporal Dynamics: The inclusion of longitudinal data to understand the changes in the behavioral profile of students over time. Temporal clustering techniques can help identify behavioral typologies of students along their learning trajectory. Recurrent neural networks can also help model the sequential patterns of student behavior.

Interpretability: The hybrid approach offers the advantage of interpretability of the student groups, as
The hybrid approach presented in this paper proves the potential of combining unsupervised and supervised learning models in a way that the prediction results are significantly higher than the results of either the factors and patterns affecting student outcomes.

DBSCAN_Cluster feature (7.2%) confirms the clustering component's value and offers the opportunity to differentiate interventions based on student behavioral phenotypes opposed to black-box models, since the student groups are easily understandable and relate to common student behavioral patterns.

REFERENCES

1. Baker, R.S.J.d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3-17.
2. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. Proceedings of 5th FUture BUsiness TEChnology Conference, 5-12.
3. Ester, M., Kriegel, H.P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 226-231.
4. Iam-On, N., & Boongoen, T. (2017). Improved student dropout prediction in Thai university using ensemble of mixed-type data clusterings. International Journal of Machine Learning and Cybernetics, 8(2), 497-510.
5. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355.
6. Shahiri, A.M., Husain, W., & Rashid, N.A. (2015). A review on predicting student's performance using data mining techniques. Procedia Computer Science, 72, 414-422.
7. Zhu, Y., Zhang, Z., & Wang, R. (2018). Hybrid clustering-classification approach for student performance prediction. Expert Systems with Applications, 112, 345-356
8. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
9. Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.
10. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.