Predictive Model to Analyze Real and Synthetic Data for Learners' Performance Prediction Using Regression Techniques

Shabnam, Ara S. J., Tanuja Ramachandriah, Manjula S. Haladappa

UVCE, Bangalore University, Bangalore

Abstract

Predicting learner performance with precision is critical within educational systems, offering a basis for tailored interventions and instruction. The advent of big data analytics presents an opportunity to employ Machine Learning (ML) techniques to this end. Real-world data availability is often hampered by privacy concerns, prompting a shift towards synthetic data generation. This study presents an empirical comparison of real, synthetic, and hybrid (real + synthetic) datasets in forecasting learner performance, deploying an array of regression-based ML algorithms, including Random Forest, Gradient Boosting, Support Vector Regression, XGBoost, and K-nearest Neighbor. Our methodology encompasses the generation of synthetic data via generative model, followed by the application of these algorithms to each dataset. The models are evaluated using precision metrics to assess their predictive accuracy. The study reveals that synthetic data can match real data in terms of predictive performance, with hybrid datasets achieving an accuracy of up to 87.76%, highlighting the effectiveness of combining both data types. These findings highlight the potential of synthetic data as an effective alternative when access to actual data is limited, promoting progress in educational technology and ML.

Keywords: Learners' performance prediction, educational data analytics, predictive models, privacy preservation, synthetic data generation, regression analysis

Shabnam, Ara S. J., Ramachandriah, T. & Haladappa, M. (2025). Predictive model to analyze real and synthetic data for learners' performance prediction using regression techniques, *Online Learning, 29*(1), 108–129. DOI: https://doi.org/10.24059/olj.v29i1.4390

In the dynamic landscape of online education, predicting learner performance has become a cornerstone for adaptive learning environments. Accurate predictions enable educators to implement personalized teaching strategies, leading to improved educational outcomes (Shabnam Ara et al., 2023). While traditional predictive models have relied heavily on real-world data, they face growing constraints regarding data privacy, availability, and comprehensiveness. The evolution of generative modelling has introduced synthetic data as a potential panacea to these challenges, allowing for large-scale, privacy-compliant data generation that mirrors real-world distributions.

Synthetic data, meticulously designed to mirror the statistical characteristics of actual data, offers transformative potential in the research of education. This approach circumvents the privacy and ethical challenges associated with original data while providing an improved dataset for training ML models. Despite its promise, its effectiveness when compared to real data in educational contexts has not been thoroughly explored.

Addressing this gap, our study probes the utility of synthetic data in the context of online learning portals. By deploying advanced regression ML algorithms across real, synthetic, and blended datasets, this research evaluates the predictive accuracy and reliability of synthetic data against its real counterpart. Our investigation is guided by an evaluation of synthetic data's capability to match the predictive accuracy of real data in educational settings. We further examine whether a combined dataset, integrating both real and synthetic data, establishes a more robust foundation for ML models than using either dataset independently. The findings of this research are intended to inform a pathway for leveraging synthetic data in predictive models, aiding stakeholders in educational technology to circumvent data accessibility challenges while maintaining the integrity and reliability of their analytical models.

Contribution

This research marks a significant stride in educational analytics by evaluating the authenticity of synthetic data against real data in training ML models. It introduces an innovative blend of real and synthetic datasets, revealing an enhanced predictive prowess that such integration can provide. Using the ACTGAN (Adaptive Conditional Tabular Generative Adversarial Network) based algorithm, the paper showcases advancements in generating synthetic data that maintains privacy without sacrificing the data's analytical value. These advancements have practical implications for educators and policymakers, offering data-driven insights for educational improvements while maintaining ethical standards**.**

Research Questions (RQ)

RQ 1: Does integrating various feature sets improve the accuracy of predictive models for academic performance?
RQ 2: How do real, synthetic, and hybrid data compare in their effectiveness for predicting learner performance in BL environments?

Related Work

The field of online learning is advancing quickly, with significant attention given to challenges such as data scalability, personalized learning, and privacy protection. ML-based predictive models play an essential role in creating adaptive learning systems. Yet, concerns over data privacy have led to the rise of synthetic data generation techniques, like Generative Adversarial Networks (GANs), which offer a solution for generating privacy-protected data (Goodfellow et al., 2020). This research expands on existing work by evaluating how real, synthetic, and hybrid datasets perform in predicting learner outcomes within blended learning environments, focusing on improving prediction accuracy while ensuring data privacy. The effectiveness of synthetic data in predictive modelling, particularly in the context of learner performance, is an emerging topic that has garnered interest in recent years. The increasing use of ML techniques in educational technology necessitates a deeper understanding of how different types of data—real, synthetic, and hybrid—contribute to the accuracy and reliability of predictive models. This literature review aims to explore the current state of research related to these topics, focusing on RQ1 and RQ2.

Enhancing Predictive Models

Several studies have explored the impact of various feature sets on the performance of predictive models in educational settings. For example, Alyahyan et al. (2020) provided a comprehensive analysis of the factors influencing academic success, identifying that combining cognitive, behavioral, and contextual features significantly enhances predictive accuracy. Shabnam Ara et al., 2024 explored parameters that impact learners' performers in BL environments. The findings indicated that factors like geographical area, study medium, frequent logins, active forum participation, and dedicating adequate time to learning activities positively impact learner performance. While sleep is essential for overall well-being, it appears to be less influential on academic outcomes in this specific context.

Namoun et al. (2020) performed a systematic analysis on the application of data mining techniques in predicting student performance. They emphasized the importance of integrating multiple data sources to capture the multifaceted nature of learning processes. The inclusion of both static and dynamic features, such as demographic information and real-time engagement metrics, has been shown to enhance model generalizability. The work by Bujang et al. (2021) also supports this notion, as they developed a multiclass prediction model using a variety of feature sets, including previous academic records, attendance, and engagement levels. Their findings indicate that models that integrate diverse feature sets outperform those relying on a single type of data. Moreno-Marcos et al. (2020) conducted an analysis to determine the factors that influence learners' performance, using learning analytics to gain insights. Their study identified a significant gap due to the insufficiency of data regarding learners' offline behavior. This finding underscores the necessity for comprehensive data collection that integrates both online and offline activities to develop more accurate and holistic predictive models.

Alalawi et al. (2023) performed a comprehensive review of ML application in predicting student performance. Their analysis of 162 research studies highlighted the most commonly used predictive techniques. The top five machine learning algorithms identified were Decision Trees (DT), Artificial Neural Networks (ANN), RF, and Naive Bayes. Tomasevic et al. (2020) compared various supervised ML techniques to forecast student exam outcomes. Their study highlighted the superior performance of ANNs and emphasized the importance of robust data collection and active student engagement. Murray et al. (2021) highlighted the crucial role of regression analysis in educating students about statistical methods. They developed a novel approach using multiple linear regression, which entails generating several multivariate datasets.

Yagci et al. (2022) investigated the use of ML algorithms to forecast students' academic performance. By including factors such as midterm exam scores, department information, and faculty details, their model achieved an accuracy rate of 70–75%. Similarly, Zhao et al. (2020) conducted a study to forecast student achievement using behavioral data. They worked with a relatively small dataset comprising 156 students. Their model reached an accuracy of 86.6%. This result highlights the potential of leveraging behavioral data for accurate academic performance prediction. This research illustrates the utility of ML techniques in educational contexts but also highlights the constraints associated with a limited number of predictors. The moderate accuracy suggests that incorporating additional variables and employing more sophisticated models could potentially improve predictive performance.

The studies reviewed in this section demonstrate that ML models such as RF and XGBoost are effective in predicting academic outcomes when applied to real-world educational data (Zhao et al., 2020; Yagci et al., 2022). However, these studies predominantly rely on real datasets, which can be limited by data privacy constraints as well as availability of large educational datasets. By introducing synthetic and hybrid datasets into the analysis, our study extends this body of work, providing a novel comparison of how real, synthetic, and hybrid data impact the performance of these machine learning models. Our findings reveal that hybrid datasets deliver enhanced predictive performance, effectively bridging the gap between the demand for accurate predictions, the limitations of real-world data availability, and the need to address data privacy concerns.

Underexplored Effectiveness of Synthetic Data

The application of synthetic data in ML has increasingly attracted attention, particularly in domains where real data is scarce or sensitive, such as education. Goodfellow et al. (2020) introduced GAN as a powerful tool for generating realistic synthetic data, which has since been used across multiple domains, including educational technology. Bethencourt et al. (2023) examined the compatibility and suitability of synthetic data generated by GANs for educational research. Although the study assessed the feasibility of using synthetic data, it did not fully explore the potential benefits and insights such data could provide. The findings highlight the need for methodologies that leverage synthetic data to enhance educational research by maintaining data integrity and predictive accuracy while preserving privacy.

Wang et al. (2020) exemplified the Data synthesizer, a framework for generating privacy-preserving synthetic datasets. Their work is particularly relevant in educational contexts where student data privacy is paramount. Similarly, Zhang et al. (2017) introduced PrivBayes, a technique for sharing private data through Bayesian networks, which has been shown to produce synthetic data that is both useful and privacy-compliant. Sarwat et al. (2022) developed a method combining a Conditional GAN (CGAN) with a deep Support Vector Machine (SVM) to forecast academic achievement. To address the issue of limited student data, they generated synthetic samples using an enhanced CGAN. Their results showed that integrating school and home tutoring improved student performance, and the CGAN-SVM model outperformed existing methods, particularly in sensitivity, specificity, and AUC. This study highlights the potential of CGAN-generated synthetic data in enhancing prediction models for technology-assisted learning. While studies like Flanagan et al. (2022) provide valuable insights, more research is needed to fully understand the conditions under which synthetic data can be as effective—or more effective—than real data. The literature suggests that while synthetic data offers significant advantages in terms of privacy and data availability, it may not always capture the complexity of real-world educational data.

Table 1

Summary of Related Work

Author	Objectives	Findings
Zhao et al. (2020)	Forecasting academic performance using behavioral data	Dataset was very small (156 students). A classification accuracy of 86.6% was achieved.
Yagci et al. (2022)	Forecasting students' academic performance through ML algorithms**.**	Attained a classification accuracy of 70–75%.
Moreno-Marcos et al. (2020	To examine the factors affecting the prediction of learners' performance through learning analytics.	Lack of data on learners' offline behavior.
Garcia et al. (2022)	To enhance educational data privacy through GAN by generating synthetic educational data.	Focuses only on privacy preservation techniques.
Bethencourt et al. (2023)	Explored synthetic data compatibility and GAN's suitability for educational research.	Doesn't fully explore the synthetic dataset's potential benefits and insights.
Proposed Work	Evaluated whether integrating various feature sets improves predictive models for academic performance and compared the efficacy of real, synthetic, and hybrid data in forecasting learner performance in a BL environment.	Combining diverse feature sets improves model accuracy and robustness and Hybrid datasets outperform real and synthetic data alone in learner performance prediction.

Garcia et al. (2022) explored the application of GANs to generate synthetic educational data with the primary aim of enhancing data privacy. While the study focused on privacy preservation, it did not extensively evaluate the effect of synthetic data on the performance of models. This research highlights the critical balance between maintaining privacy and ensuring accurate performance predictions. These studies emphasize the need for integrating diverse data sources and advanced machine learning techniques to improve academic performance predictions while highlighting the potential of synthetic data to balance privacy and predictive accuracy. Table 1 summarizes the related work.

The application of synthetic data in education is still an emerging area, with studies such as Bethencourt et al. (2023) and Garcia et al. (2022) primarily focusing on its privacy benefits. However, there is limited research comparing synthetic data to real data in terms of predictive accuracy. Our study builds on this by directly evaluating synthetic data alongside real and hybrid datasets, demonstrating that a hybrid approach can significantly improve model accuracy while maintaining privacy. This highlights the untapped potential of synthetic data in educational predictive modelling.

The field of online learning faces several significant challenges that hinder the development of effective educational technologies. Current challenges in online learning research include:

Data Availability

- limited access to high-quality educational datasets restricts the effectiveness of predictive models
- real-world data is often scarce or incomplete, affecting the robustness of machine learning algorithms

Data Privacy Concerns

- growing concerns about the privacy of sensitive learner information necessitate solutions that protect data while enabling effective analysis

Integration of Diverse Data Sources

- difficulty in combining data from various sources (e.g., demographic, behavioral) to create a comprehensive view of learner performance

Our research addresses these gaps through several key approaches. First, it uses hybrid datasets that integrate real and synthetic data, enhancing the robustness of predictions while maintaining data privacy. Second, by employing GANs for synthetic data generation, the study offers a solution to the limitations posed by the availability of real-world data. Additionally, the research provides empirical evidence for enhanced predictive performance by comparing real, synthetic, and hybrid datasets, demonstrating their respective impacts on predictive accuracy in blended learning environments. Finally, the insights derived from this research can inform the development of more effective and privacy-compliant educational technologies, thereby contributing to the ongoing discourse in online learning research.

Method

This section provides a detailed overview of the approach used in the study, covering the system architecture, model development, evaluation process, and hyperparameter tuning techniques.

System Overview

The proposed system architecture, as illustrated in Figure 1, is structured into three distinct phases: data pre-processing, synthetic data generation, and model training and evaluation.

Phase 1: Data Pre-processing

Data Cleaning: In this step, errors, missing values, and outliers in the dataset were detected and rectified to maintain data quality and integrity. To address missing values, some were either removed or imputed. Out of the initial 338 student records, 8 records were excluded due to significant missing values, and 5 records were corrected using mean imputation. This process resulted in a final dataset consisting of 330 student records.

Figure 1

System Architecture

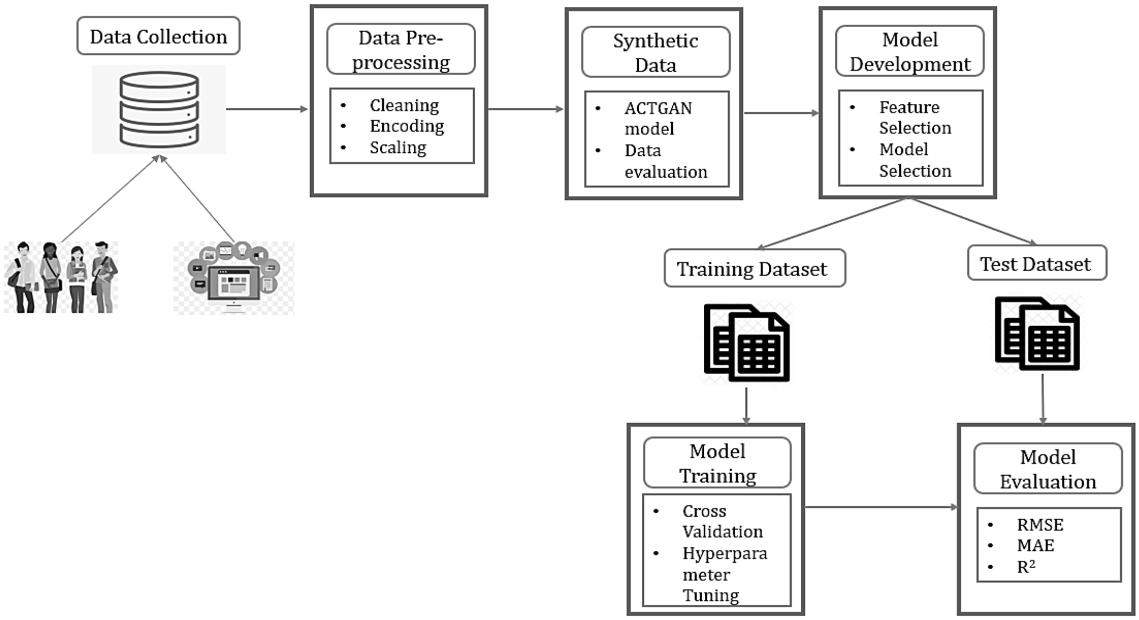


Table 2

Data Encoding Methods Employed

Encoding Type	Specification	Used on
Binary Encoding	Applied to binary data with only two categories where the order does not matter.	5 features e.g., gender, medium of study, region
Ordinal Encoding	Used for ordinal data where the categories have a meaningful order or ranking, though the intervals between them may not be equal.	20 features e.g., login frequency, repetition of videos

1. Dataset Encoding: To facilitate the application of machine learning models, categorical variables were converted into numerical values through encoding techniques as in Table 2.

2. Dataset Scaling: Feature scaling was applied to adjust the numerical feature values to a consistent range, ensuring that all features have equal influence on the model

by removing variations in scale. We normalized the data using the formula below, scaling it between 0 and 1.

$$o' = \frac{o - o_{min}}{o_{max} - o_{min}}$$

where o_{min} and o_{max} is minimum and maximum value of feature, and o' and o is scaled and original value of feature

Phase 2: Synthetic Data Generation

In this phase, depicted in Figure 2, synthetic data is produced using the ACTGAN model, a tabular GAN approach, via the Gretel.ai tool. A dataset consisting of 330 student records was used to generate 5,000 synthetic records.

Figure 2

Generation of Synthetic Dataset

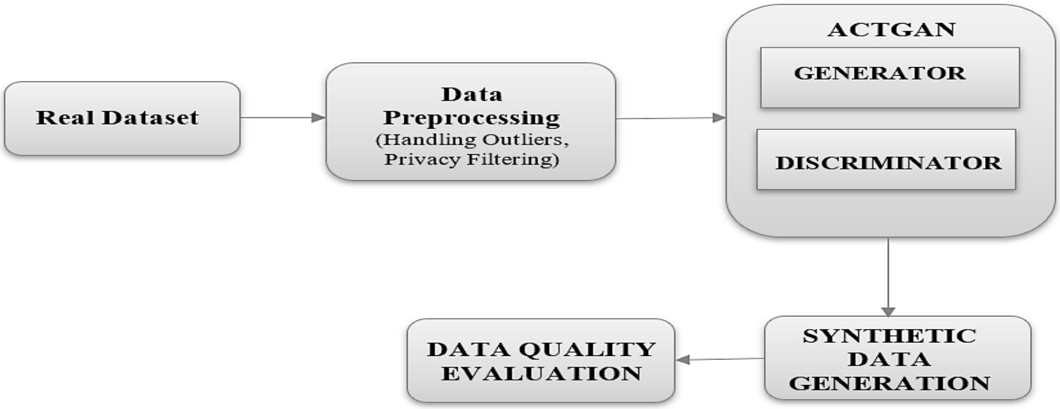

Real Dataset
Data Preprocessing (Handling Outliers, Privacy Filtering)
ACTGAN
GENERATOR
DISCRIMINATOR
SYNTHETIC DATA GENERATION
DATA QUALITY EVALUATION

Table 3 details the configuration used for generating synthetic data with the ACTGAN model on the Gretel.ai platform. ACTGAN, a specialized type of Generative Adversarial Network (GAN), is employed here to produce synthetic tabular datasets. The Adam optimizer is chosen for managing the model's training process, renowned for its ability to adapt learning rates for each parameter. The generator, tasked with producing the synthetic data, operates with a learning rate of 0.0001, facilitating gradual and consistent adjustments during training. Meanwhile, the discriminator, which differentiates between real and synthetic data, is configured with a slightly higher learning rate of 0.00033. This setup helps balance the adversarial relationship between the generator and discriminator, leading to improved data generation quality throughout the 50 training epochs.

Table 3

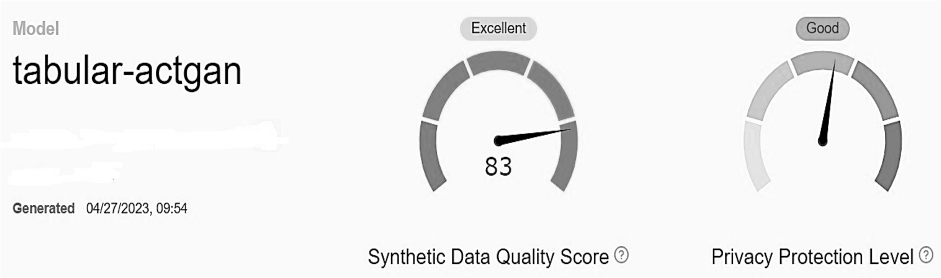
Synthetic Data Configuration

System	Gretel.ai
Model	ACTGAN

Generator Learning Rate	0.0001
Epochs	50
Optimizer	Adam
Discriminator Learning Rate	0.00033

Figure 3

Synthetic Data Quality Score



Model

tabular-actgan

[redacted]

Generated 04/27/2023, 09:54

Figure 3 depicts the synthetic data quality score, which reached 83 percent, indicating high data quality and also a robust privacy protection is achieved for the generated dataset. The score is calculated using a weighted combination of multiple quality metrics, such as field distribution stability, field correlation stability, and deep structure stability. The synthetic data quality score indicates how well the synthetic data replicates the statistical characteristics of the original dataset. As such, it serves as a score, suggesting that scientific conclusions drawn from the synthetic data would likely align with those derived from the original data. However, in scenarios where statistical accuracy is less critical, such as in testing or demonstration environments, a lower quality score may still be acceptable.

Figure 4

Synthetic Data Quality Summary Statistics

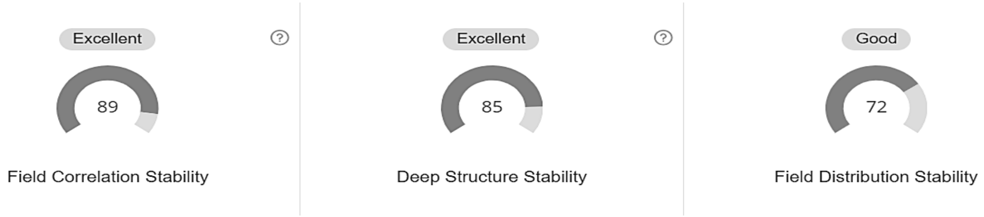


Figure 4 provides an overview of the quality statistics for synthetic data, showing between each pair of fields in the original training data is compared with the corresponding pairs in the synthetic data. The absolute differences between these correlations are averaged across all field pairs, with a lower average indicating a higher stability score. To visualize these correlations, heatmaps are provided for both the training and synthetic data, along with a heatmap illustrating the differences. Preserving field correlations is crucial when the synthetic data is intended for statistical analysis or ML.

Additionally, deep structure stability has been measured at 85%. This is assessed using Principal Component Analysis (PCA), applied separately to both the original and generated datasets. The similarity of the principal components between the two datasets determines the synthetic quality score, with a closer alignment indicating better quality. As PCA is widely used for dimensionality reduction and visualization in machine learning, this metric offers immediate feedback on the synthetic data's utility.

Lastly, the field distribution stability score is 72%, reflecting how well the synthetic data replicates the distributions of the original data fields. This is measured using the Jensen-Shannon Distance for each numeric or categorical field, where a lower distance score corresponds to higher stability. Bar charts or histograms are used to compare these distributions. Maintaining field distribution integrity is essential, especially depending on the data's intended application.

The synthetic data generated was merged with the original real dataset, resulting in a unified dataset of 5,330 records. This integration provided a more varied and complete dataset for additional analysis. We named this the Hybrid Dataset.

Table 4

Feature Set

Feature Category	Attribute Name
P1: Learners' Background Data	Age
	Gender
	Matriculation Medium of Study
	Region Residing
	Family Annual Income
	Parental Occupation
P2: Previous Experience with Digital Learning Environment	Basic Computer Skills
	Internet Accessibility
	Ease of Use
P3: Digital Learning Environment Interaction	Login Frequency
	Percentage of Online Lectures Viewed
	Time Spent Viewing Online Lectures
	Percentage of Activities Completed
	Average Number of Repetition of Lectures
	Average Number of Pauses While Watching Lectures
	Number of Practice Tests Taken

	Number of Supplementary Study Materials Downloaded
P4: Forum Interaction	Frequency of Questions Asked in Forum
	Frequency of Interaction with Peers
	Frequency of Interaction with Instructor
P5: Lifestyle Behavioral Data	Frequency of Physical Activity
	Duration of Sleep
	Smartphone Usage for Study Purpose
	Smartphone Usage
	Diet

Phase 3: Model Development

Feature set description. This study consists of 330 students records and encompasses a variety of learner characteristics, including demographic details, previous academic performance, engagement metrics within online learning platforms, and lifestyle habits, collected from the Government Polytechnic of Karnataka, India. The dataset is categorized in five features as P1, P2, P3, P4, and P5. Table 4 shows the feature set used for model development.

Feature selection. In the research, feature selection was conducted using both filter and wrapper methods to enhance the predictive precision of the ML models. The filter method was used to evaluate the relevance of features based on statistical measures, identifying the most significant features individually. This approach was particularly useful in the single feature selection phase, where each feature's contribution to the model's performance was assessed independently. Building on this, the wrapper method was employed to evaluate combinations of features by training models on various feature subsets. This method was crucial in the twin, triple, quad, and five feature set selections, where multiple features were grouped together and their combined impact on model performance was analyzed. The wrapper method allowed for a more comprehensive assessment, considering the interactions between features that might not be evident when features are evaluated in isolation.

Model selection. In this section, we detail the various models that were employed and compared in the study to evaluate their effectiveness in predicting student performance. The ML models as in Table 5 were selected based on their relevance and prior success in similar educational data mining tasks.

Table 5

Mathematical Representations and Descriptions of Machine Learning Models

Model	Mathematical Representation	Description
GBRT	$GB(X) = \sum_{m=1}^{M} h_m(X)$	GB builds an additive model by summing predictions from M weak learners. Each $h_m(X)$ represents the prediction from the m-th weak learner.

RF	$RF(X) = \frac{1}{N}\sum_{i=1}^{N} T_i(X)$	RF is an ensemble technique that combines predictions from N decision trees. Each $T_i(X)$ represents the prediction from the *i*-th DT, and *RF(X)* is the average of these predictions.
XGB	$Objective(N) \;=\; \Sigma_{\text{i}}\, L(y_{\text{i}}, Fn - 1(x_{\text{i}})) \;+\; \Omega(Fn)$	XGB minimizes an objective function combining a loss term *L* (e.g., Mean Squared Error) and a regularization term *Ω* to find the optimal model parameters across *N* boosting iterations.
KNN	$KNN(x) = \frac{1}{k}\sum_{i=1}^{k} y_i$	KNN predicts by averaging the target values y_i of the k nearest neighbors to the data point *X*.
SVR	Minimize: $\frac{1}{2}\lVert\omega\rVert^2 + C\sum_{i=1}^{N}(\xi_i + \varepsilon_i^*)$	SVR seeks to find a hyperplane that reduces the prediction error while keeping the errors within a predefined margin *(ϵ)*. The model balances error minimization and margin constraints using parameters *ω* (weights), *b* (bias), and *C* (cost).

Hyperparameter tuning. The hyperparameters for various machine learning models were carefully tuned to optimize their performance. Hyperparameters were initially selected based on prior research and initial estimates. To optimize these, we employed both RandomizedSearchCV and GridSearchCV. We started with RandomizedSearchCV to quickly explore a wide range of hyperparameters by sampling a set number of combinations from predefined distributions, allowing for efficient coverage of the search space. After narrowing down to a promising range, we used GridSearchCV for a more exhaustive and systematic evaluation, testing all possible combinations within the defined grid to fine-tune the hyperparameters for optimal model performance. Details of the hyperparameter configurations used are provided in Table 6.

Evaluation Metrics

To assess the effectiveness of our models, we used a range of established metrics that provide a comprehensive assessment of prediction accuracy and reliability. Table 7 presents these metrics in detail, highlighting the model's strengths and areas for improvement. A lower RMSE and MAE indicate a model with higher accuracy. These metrics, collectively, enabled a robust evaluation of model performance, ensuring that both the accuracy and consistency of predictions were thoroughly assessed.

Table 6

Hyperparameter Configuration

Model	Hyperparameter	Value
GBRT	Learning Rate	0.1
	Number of Estimators	200
	Maximum Depth	3

RF	Number of Trees	100
	Maximum Depth	10
	Criterion	Gini Impurity
XGB	Learning Rate	0.1
	Maximum Depth	6
	Subsample	0.8
SVR	Kernel	Radial Basis Function
	C (Regularization Parameter)	1
	Epsilon	0.1
KNN	Number of Neighbors	5
	Distance Metric	Euclidean

Table 7

Accuracy Metrics

Metric	**Formula**	**Description**
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert p_i - \hat{p}_i\rvert$	Calculates the average absolute deviations between predicted values ($\hat{p}_i$) and actual values (p_i), offering an assessment of the model's mean prediction error.
Accuracy measured using R-Squared Co-efficient (R^2)	$\frac{\sum_{i=1}^{n}(p_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(p_i - \overline{p_i})^2}$	It quantifies how well the predicted values ($\overline{p_i}$) from the model align with the actual values (p_i).
Root Mean Square Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2}$	Quantifies the average extend of the discrepancies between predicted values ($\hat{p}_i$) and actual values (p_i).

Implementation Details

The experiments were carried out using Python 3.10.12 on a computing system equipped with 16 GB of RAM and 8 GB of GPU. For the implementation, the NumPy, Pandas, and Scikit-learn libraries were used for numerical computations, data manipulation, and machine learning algorithms, respectively. The dataset, consisting of both pre-processed real and synthetic data, was stored in CSV files. To prepare the data for model training and evaluation, the train_test_split function was employed to partition the dataset, using 80% of the data for training and keeping the remaining 20% for evaluating the model's performance.

Results

RQ 1: Does integrating various feature sets improve the accuracy of predictive models for academic performance?

To address RQ1, we conducted experiments on real datasets using various combinations of feature sets to evaluate their impact on predictive models for academic performance. We first tested each feature set individually, then explored combinations of two feature sets, three feature sets, and four feature sets. Finally, we assessed the performance when all five feature sets were used together. These experiments were carried out using RF, XGB, KNN, SVR, and GBRT.

Single Feature Set

Figure 5 presents the RMSE values obtained using individual feature sets on the real dataset. The XGB model recorded the highest RMSE of 19.35 for the "P4" feature set, highlighting its sensitivity to fluctuations within single feature sets. Conversely, the lowest RMSE of 12.3 was achieved by the RF model for the feature set "P3" demonstrating the effectiveness of RF methods when Digital Learning Environment Interaction features were used. SVR was consistent and showed higher accuracy across all single parameter combinations. Overall, the RMSE results for single feature sets show that while certain feature sets offer a reasonable degree of predictive capability, none of the models performed optimally when relying on a single feature set alone.

Twin Feature Set

When combining two feature sets, SVR consistently delivered strong performance across all pairs, as depicted in Figure 6. The highest performance was achieved by the RF model with the P1 P3 feature set combination, yielding an RMSE of 12.47 showing strength when P1 and P3 features taken together. In contrast, the XGB model performed the worst with the P1_P4 combination, recording an RMSE of 19, though P4 individually performed better. The addition of a second feature set generally enhanced predictive accuracy, leading to lower RMSE values across most models.

Triple Feature Set

With the integration of a third feature set, a noticeable improvement in model performance was observed as depicted in Figure 7, particularly in reducing RMSE values. The SVR model achieved the best performance, recording the lowest RMSE of 12.47 for the combination of feature sets P1, P4, and P5. In contrast, the XGB model displayed the highest RMSE of 18.56 for the feature sets P1, P2, and P4. The consistent performance of the SVR model across different feature set combinations underscores the value of incorporating additional data dimensions, which enhances the model's predictive accuracy.

Figure 5

Over Single Feature Set

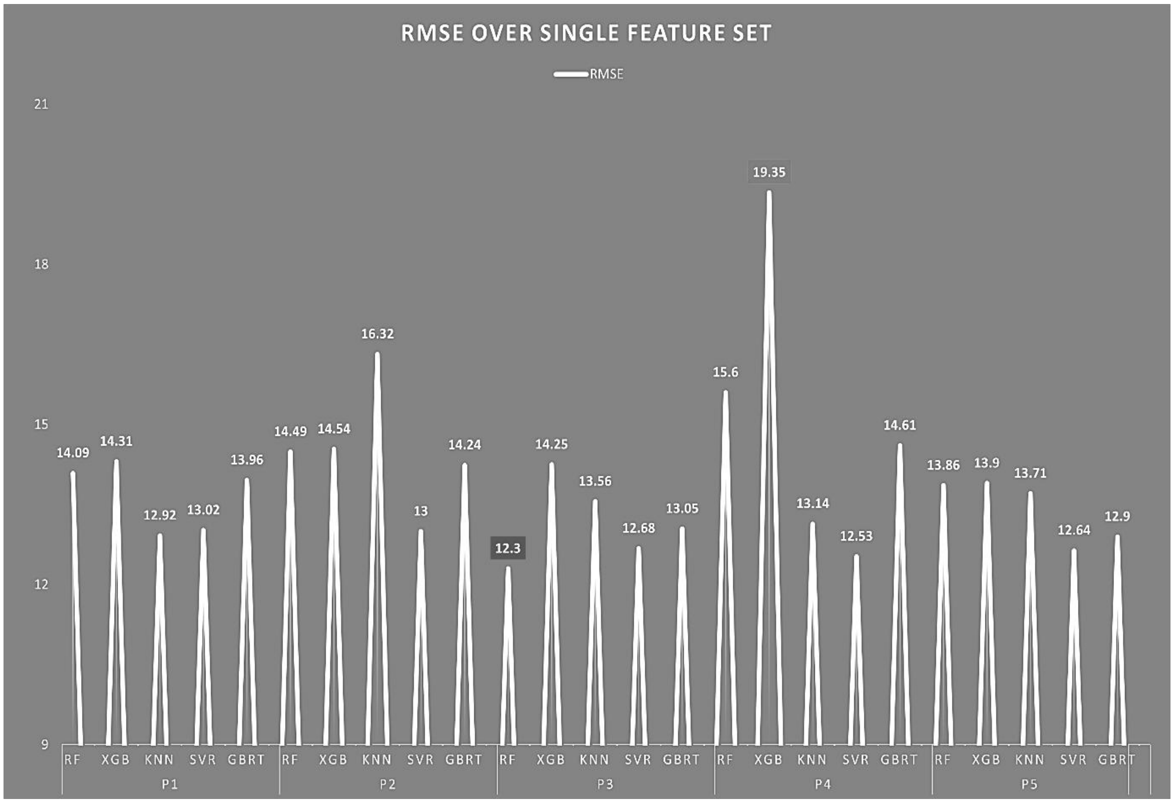


Figure 6

RMSE Over Twin Feature Set

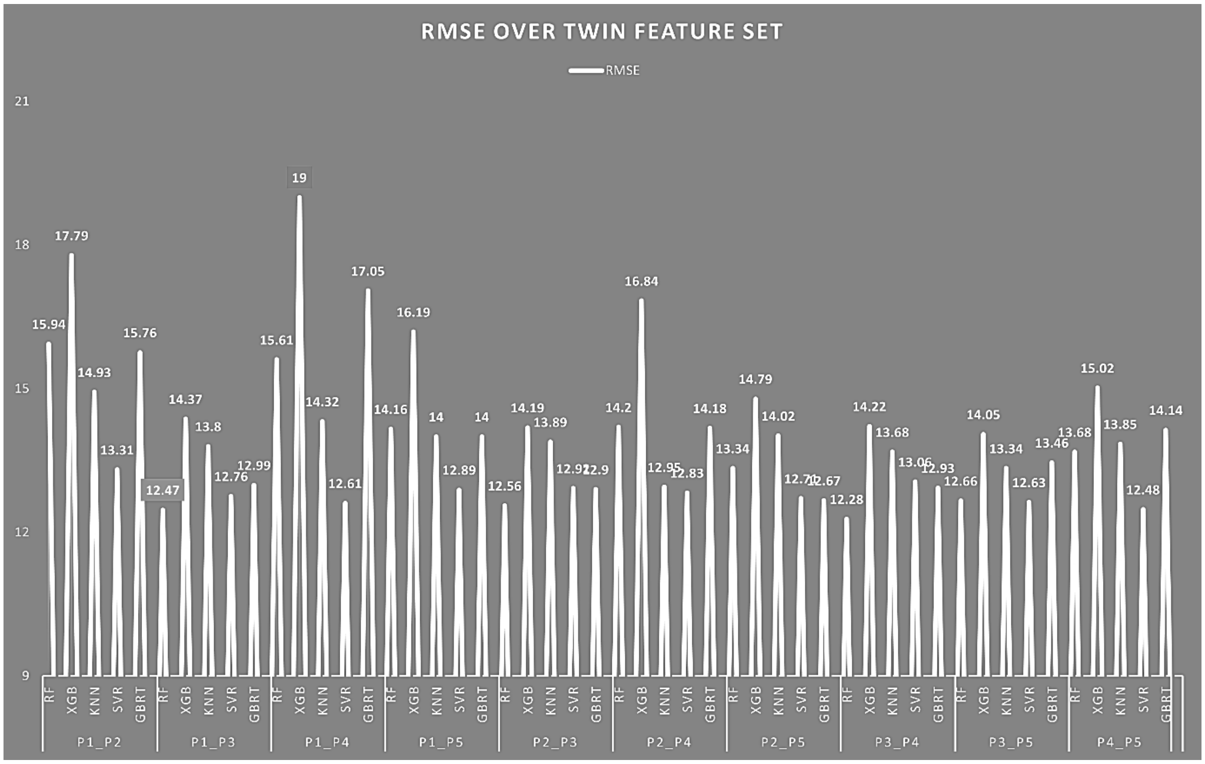


Figure 7

RMSE Over Triple Feature Set

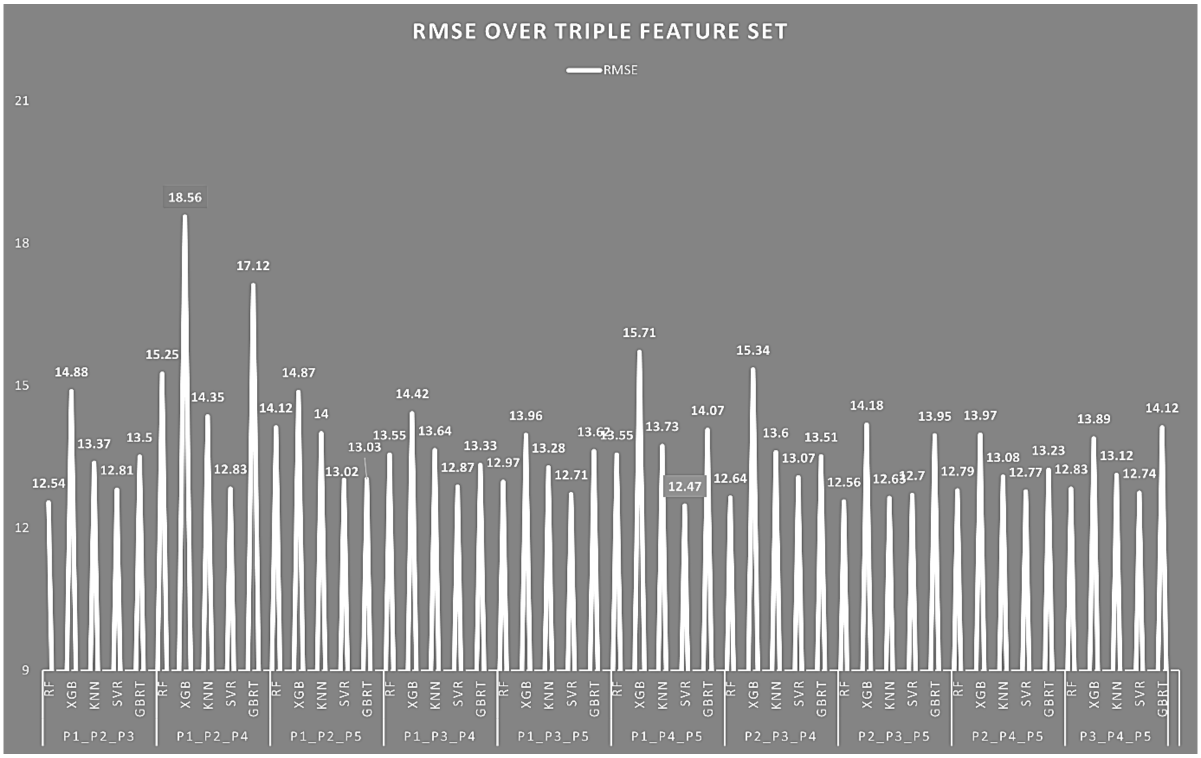
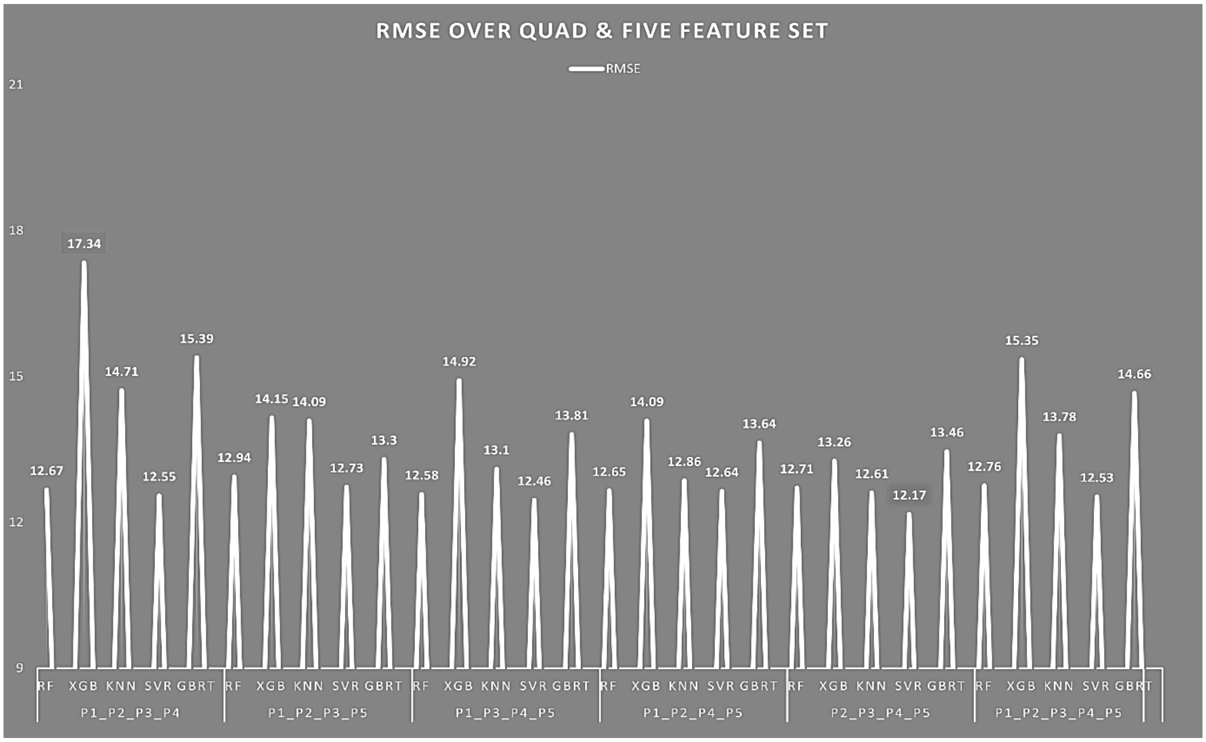


Figure 8

RMSE Over Quad and Five Feature Set



Four and Five Feature Set

When combining four or five feature sets as depicted in Figure 8, the SVR model continued to demonstrate strong performance, achieving the lowest RMSE of 12.17 for the feature combination P2_P3_P4_P5. On the other hand, the XGB model recorded the highest RMSE of 17.34 for the combination of P1_P2_P3_P4. The SVR model consistently performed well across all feature set combinations, highlighting its robustness in handling diverse data dimensions.

To summarize the results, we have Figure 9 and Figure 10 depicting average RMSE and average MAE. Average errors decrease significantly for the quad feature set, indicating that this is the optimal feature set for this particular analysis. This reduction in error rates as the number of features increases suggests that a more comprehensive feature set allows the models to capture underlying patterns more effectively, leading to improved predictive accuracy. These results highlight the importance of features that represents learners' appropriately in enhancing model performance, emphasizing that including a well-balanced set of relevant features can significantly improve the learners' ability to generalize from the data.

Figure 9

Average RMSE Variation Over Feature Set

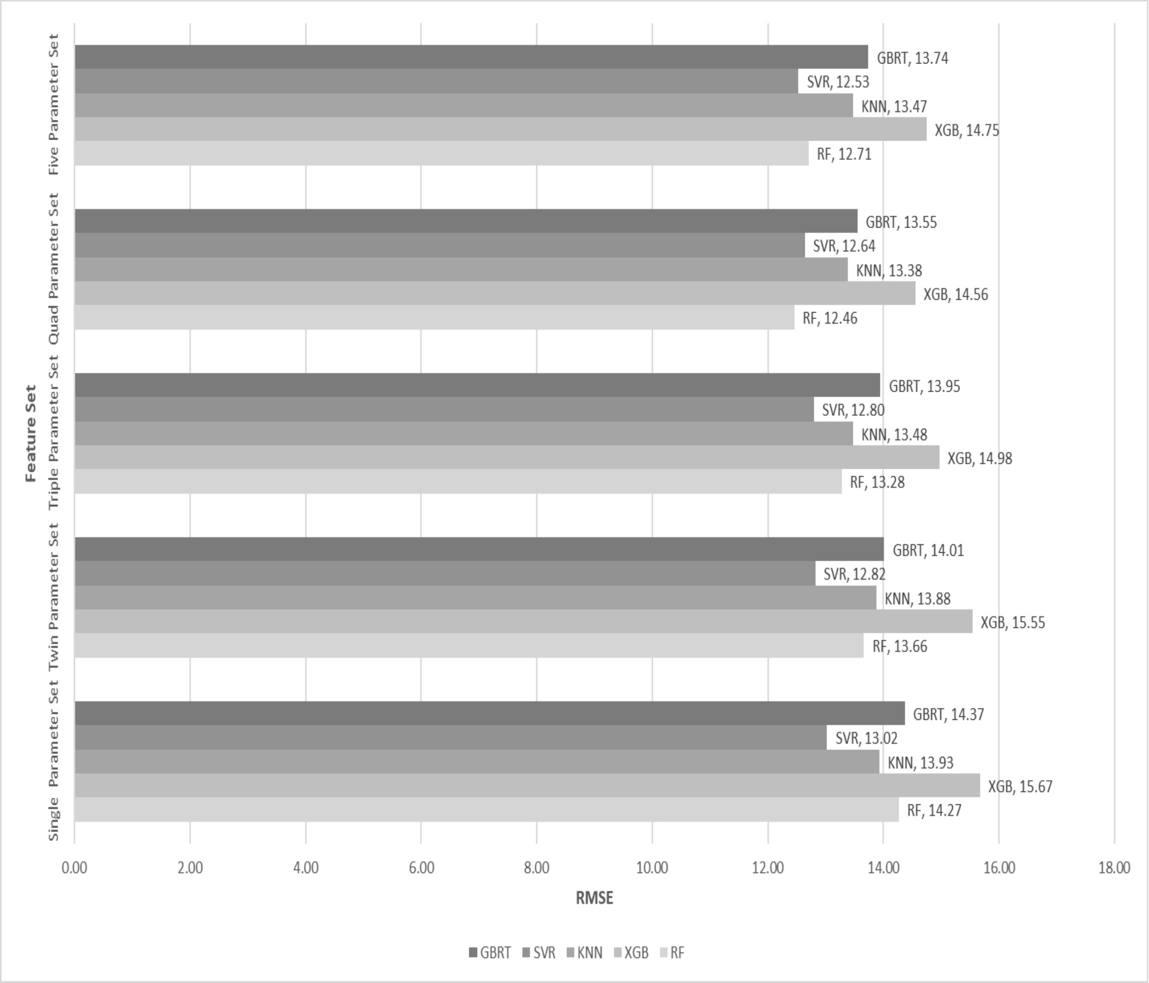
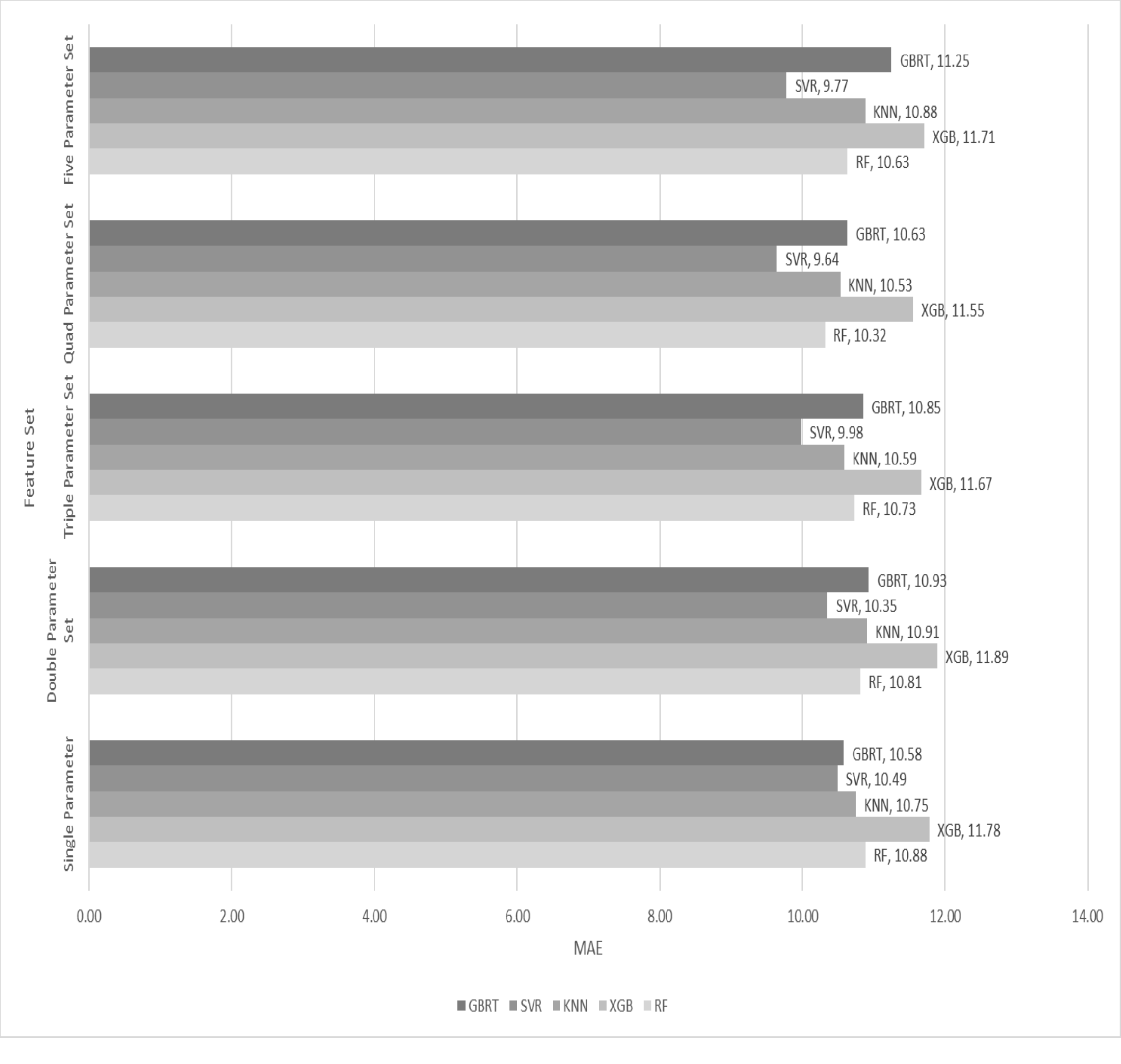


Figure 10

Average MAE Variation Over Feature Set

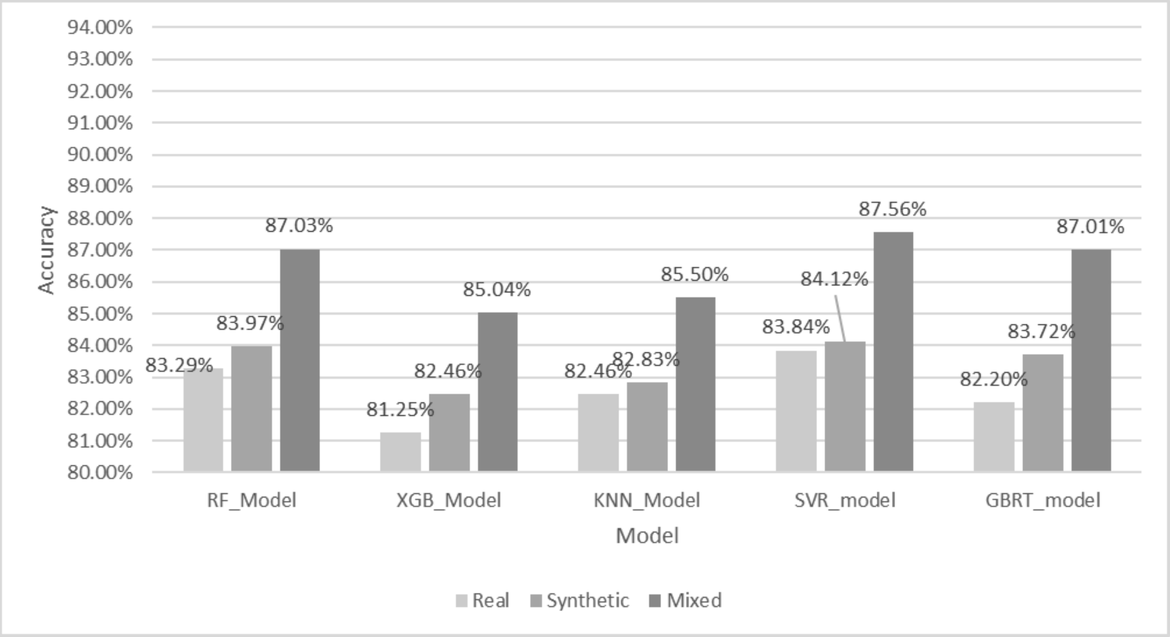


RQ 2: How do real, synthetic, and hybrid data compare in their effectiveness for predicting learner performance in BL environments?

Figure 11 clearly illustrates that the hybrid dataset consistently surpasses both synthetic and real datasets across various machine learning models. This result emphasizes the significant advantage of integrating real and synthetic data to predict learner performance in BL environments. Among the traditional models, the SVR model achieved the highest accuracy of 87.56% using the hybrid dataset, highlighting its superior performance. This suggests that the hybrid dataset's broader scope and diverse learner data enhance the predictive accuracy of the algorithms. Additionally, the application of feature engineering techniques, such as wrapper and filter methods, further improves the accuracy and reliability of academic performance. predictions.

Figure 11

Accuracy Comparison Across Real, Synthetic and Hybrid Dataset



Discussion

This study investigates the impact of different feature sets on predicting student performance and evaluates the effectiveness of real, synthetic, and hybrid data in forecasting learner outcomes. The experimental results demonstrate improved performance due to several key factors:

1. **Feature Optimization**: The use of filter and wrapper methods significantly enhanced both the accuracy and efficiency of the model. Additionally, the inclusion of synthetic and hybrid data during feature selection contributed to a more comprehensive and robust feature set.
2. **Dataset Enrichment**: Integrating multi-source student data with synthetic data substantially increased the dataset's depth and diversity. The hybrid dataset, combining real and synthetic data, proved particularly valuable, as it balanced data variety and volume, leading to richer insights and improved predictive performance.

The key findings of this study highlight that a diverse range of learner data—including background information, digital engagement, and lifestyle behaviors—can effectively predict student performance. The hybrid dataset consistently outperformed both real and synthetic datasets due to its greater volume and diversity, which enhanced the performance of all tested algorithms. This indicates that synthetic data can effectively supplement real data and that combining multiple data sources significantly improves predictive accuracy. It is crucial to acknowledge the limitations of this study. The research is restricted to Karnataka, which may restrict the generalizability of the results to other regions. Additionally, while synthetic data was used, it may not completely reflect the intricacies of the original dataset, and the computational resources available might affect the method's scalability.

Conclusions

This study offers a thorough analysis of actual and generated data using tabular GAN method, for learners' academic score prediction in BL environment. The findings suggest that synthetic data can be an effective substitute for actual data for student score prediction in academics and synthetic data demonstrates comparable performance. Notably, models trained on a hybrid dataset demonstrated even greater accuracy and robustness. Further exploration into innovative data synthesis methods and their impact on long-term performance prediction will be crucial. The success of integrated hybrid data underscores the significance of enhancing data quality and ensuring privacy while enhancing ML applications in education.

Practical Implications

The findings of this study offer important insights for educational institutions seeking to improve student outcomes performance prediction models. By leveraging a hybrid dataset that integrates real and synthetic data, educational systems can achieve more accurate and reliable predictions. This approach not only improves the personalization of learning experiences but also allows institutions to better identify at-risk students early on. Moreover, the successful integration of synthetic data reduces reliance on large volumes of real data, alleviating privacy concerns and allowing for more flexible data collection practices. The enhanced feature selection process, which incorporates both synthetic and hybrid data, can be applied across different educational contexts, improving the adaptability and scalability of predictive models. As a result, institutions can develop more robust data-driven strategies to support student success and optimize educational outcomes.

Declarations

Availability of Data and Materials

The dataset generated and analyzed during the current study is not publicly available due to privacy and confidentiality but are available from the corresponding author on reasonable request.

Competing Interests

The authors declare that they have no competing interests.

Funding
The research was not funded by any organization.

References

Alalawi, K., Athauda, R., & Chiong, R. (2023). Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review. *Engineering Reports*, p.e12699.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*, 1–21.

Bethencourt-Aguilar, A., Castellanos-Nieves, D., Sosa-Alonso, J. J., & Area-Moreira, M. (2023). Use of generative adversarial networks (GANs) in educational technology research. *Journal of New Approaches in Educational Research, 12*(1), 153–170.

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N.A.M. (2021). Multiclass prediction model for student grade prediction using machine learning. *IEEE Access*, *9*, 95608–95621.

Flanagan, B., Majumdar, R., & Ogata, H. (2022). Fine grain synthetic educational data: challenges and limitations of collaborative learning analytics. *IEEE Access*, 10, 26230–26241.

Garcia, M., Smith, J., & Lee, H. (2022). Enhancing educational data privacy through generative adversarial networks. *Journal of Educational Technology*, *26*(3), 123–136.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144.

Moreno-Marcos, P. M., Pong, T. C., Munoz-Merino, P. J., and Kloos, C. D. (2020). Analysis of the factors influencing learners' performance prediction of multisource, multifeature behavioral data with learning analytics. *IEEE Access, 8,* 5264–5282.

Murray, L. L., & Wilson, J. G. (2021). Generating data sets for teaching the importance of regression analysis. *Decision Sciences Journal of Innovative Education*, *19*(2), 157–166.

Yagci, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 11. https://doi.org/10.1186/s40561-022-00199-2.

Sarwat, S., Ullah, N., Sadiq, S., Saleem, R., Umer, M., Eshmawi, A.A., Mohamed, A., & Ashraf, I. (2022). Predicting students' academic performance with conditional generative adversarial network and deep SVM. *Sensors, 22*(13), 4834.

Shabnam Ara, S. J., & Tanuja, R. (2023). Investigating the influential factors of learner performance in online education using a learning analytics approach. In *2023 3rd*

International Conference on Intelligent Technologies (CONIT), pp. 1–11. https://doi.org/10.1109/CONIT59222.2023.10205849

Shabnam Ara, S. J., & Tanuja, R. (2024). Exploring key parameters influencing student performance in a blended learning environment using learning analytics. *Journal of Education and e-Learning Research*, *11*(1), 77–89. https://doi.org/10.20448/jeelr.v11i1.5330

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, *143*, 103676.

Wang, L., Chen, W., Yang, W., Bi, F., & Yu, F. R. (2020). A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, *8*, 63514–63537.

Zhao, L., Chen, K., Song, J., Zhu, X., Sun, J., Caulfield, B., & Mac Namee, B. (2020). Academic performance prediction based on multisource, multifeature behavioral data. *IEEE Access*, *9*, 5453–5465.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., & Xiao, X. (2017). Privbayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems (TODS)*, *42*(4), 1–41.