International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Analysis of Predictive Models for Learner Performance using Synthetic Data and Regression Techniques

Shabnam Ara S J*[1], Tanuja R[2], Manjula S H[3]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: Timely identification of learners' performance is crucial for educators to intervene effectively before students encounter academic challenges. However, the scarcity and privacy concerns surrounding educational datasets pose significant hurdles. In this study, we investigate the efficacy of predictive models for learner performance using synthetic data and regression techniques. Our analysis focuses on a multi-source dataset from technical education, which has been expanded through synthetic data generation. Employing regression machine learning algorithms, we evaluate the prediction performance across actual, generated, and augmented datasets. Our findings indicate notable improvements with augmented datasets, achieving an R-squared coefficient of 0.8776. These results underscore the effectiveness of hybrid data approaches and advocate for the integration of synthetic data as a viable alternative, particularly in contexts where access to real data is limited. This integration holds promise for advancing educational technology and machine learning methodologies. Through comprehensive analysis of diverse data sources and the application of regression techniques on synthetic and augmented datasets, this investigation endeavors to evaluate the efficacy of predictive models concerning learner performance. Additionally, this study elucidates the potential utility of synthetic data as a viable alternative in instances where the available real dataset is limited in scale.

Keywords: *Education Data- Generators, Learners' Performance, Predictive Models, Regression Models, Technical Education*

1. Introduction

The effectiveness of educational initiatives is intricately tied to the ability to promptly and precisely evaluate student performance, a task that has become notably intricate amidst the proliferation of data within contemporary digital learning ecosystems [1]. Predictive analytics has emerged as a potent instrument, empowering educators to preemptively discern students at risk and customize interventions accordingly. Nonetheless, despite their promise, these methodologies often encounter obstacles due to the constrained accessibility and delicate nature of educational data.

Synthetic data, artfully crafted to reflect the statistical properties of real datasets, presents a groundbreaking opportunity in educational research. It not only bypasses the privacy and ethical considerations linked with real data but also offers an enriched dataset for training Machine Learning (ML) models. However, the efficacy of synthetic data and its comparative effectiveness against real data in educational settings remains underexplored.

To surmount these obstacles, our study delves into the utilization of synthetic data, an innovative strategy designed to replicate the statistical characteristics of authentic data while circumventing privacy and scarcity concerns using Gretel.ai. Integrated with advanced regression techniques like Support Vector Regression (SVR), Gradient Boosting Regression Trees (GBRT), Random Forest (RF), eXtreme Gradient Random Boost (XGB), and K- Nearest Neighbour (KNN), we propose an all-encompassing framework for forecasting learner performance. Through the generation and integration of synthetic data into our multi-source dataset from the technical education sphere, our objective is to augment the predictive model's efficacy and applicability. Additionally, we undertake a thorough examination of the comparison between actual and synthetic datasets to gauge the effectiveness and dependability of our machine learning algorithms. The incorporation of synthetic data presents considerable potential for guiding the enhancement of sophisticated pedagogical instruments, and our inquiry aims to furnish a substantial contribution to the realms of educational technology and analytics. Through the perspective offered by this study, we strive to lay the foundation for a fresh paradigm in educational data analysis—one that fosters wider implementation and ingenuity in addressing data limitations. The present study addresses the following Research Questions (RQ):

RQ1: Does the Combination of Different Feature Sets Enhance Predictive Models for Academic Performance in Real, Synthetic, and Mixed Datasets?

*[1] *Department of CSE, UVCE, Bangalore University, Bangalore.*
ORCID ID: 0000-0002-4507-5888
[2] Department of CSE, UVCE, Bangalore University, Bangalore.
ORCID ID: 0009-0008-2702-5837
[3] Department of CSE, UVCE, Bangalore University, Bangalore.
ORCID ID: 0000-0002-8177-2672
** Corresponding Author Email: Shabnam.jahagirdar@gmail.com*

RQ2: What is the comparative predictive performance of models across actual, generated, and augmented datasets for academic performance prediction in BL?

1.1 Contributions

- Comprehensive Academic Performance Forecasting: Our study pioneers a holistic method for predicting academic performance, extending beyond single-course predictions to accurately forecast outcomes for a student's entire semester.
- Cutting-edge Synthetic Data Generation: Leveraging the Tabular-ACTGAN algorithm via Gretel.ai, our research generates a substantial synthetic dataset of 5,000 entries with an 83% quality score, overcoming the limitations of small sample sizes and enhancing the reliability of our predictive models.
- Enhanced Multi-source Data-driven Regression: Our research enhances regression analysis by integrating a multi-source dataset, delving into various factors like lifestyle habits, digital engagement, and socio-economic indicators. This approach significantly improves the potential for targeted educational intervention.

2. Related Work

Data synthesis, an essential component in the realm of data science, encompasses various approaches and methods devised by researchers. One widely adopted technique employs Generative Adversarial Networks (GANs), demonstrating their efficacy in generating synthetic data that faithfully reproduces the original data distribution [2]. To address privacy concerns, differentially private GANs have been introduced [3], adding noise into generated samples to protect sensitive information.

An alternative technique involves rule-based synthesis methods, exemplified by the Data Synthesizer framework [3]. This method leverages Bayesian networks to capture statistical dependencies among attributes, generating synthetic data while preserving essential characteristics. Privacy-preserving data synthesis is tackled by the PrivBTS algorithm [4], which utilizes Bayesian network structures to create synthetic data while preserving privacy guarantees. Moreover, the utility of synthetic data is a paramount concern. The PrivBayes algorithm [5], combining sampling and tree-based partitioning, generates synthetic data that balances privacy preservation with data utility.

The authors in [6] explored the application of GANs in educational technology research. They assessed the compatibility of synthetic data with real data and investigated GANs' suitability for educational research. By employing the COPULA-GAN algorithm, they created synthetic datasets for analysis. The study involved a two-stage cluster analysis, highlighting the resemblance and interchangeability between synthetic and original datasets.

The work in [7] emphasized the importance of regression analysis in teaching students the significance of statistical analysis. They proposed a novel approach using multiple linear regression, which involves generating alternative multivariate datasets to emphasize the importance of advanced statistical analysis. Researcher in [8] introduced an improved approach that combines a Conditional Generative Adversarial Network (CGAN) with a deep-layer-based SVM to predict academic success. To overcome the limitation of having a limited number of student educational records, the team utilizes synthetic data samples created through an enhanced CGAN. The findings from the CGAN training indicate that the combination of school and home tutoring positively impacts children's performance. Notably, when compared to existing solutions in the literature, suggested advanced CGAN combined with the deep SVM exhibits superior performance, particularly in terms of sensitivity, specificity, and the area under the curve. Their study demonstrates the potential of synthetic data generated by CGAN in improving performance prediction models for technology-assisted learning platforms.

An interpretable model for predicting student performance in "Introduction to Programming" courses was developed [9]. Their model utilizes data derived from programming assignment submissions and employs a stacked ensemble model with SHAP (SHapley Additive exPlanations), a game-theory-based framework to forecast students' final exam grades. This study also discerns distinct student profiles based on their problem-solving tendencies. Learners' academic outcome prediction using data mining and learning analytics was done in [10]. They analyzed 62 papers from 2010 to 2020 and identified key predictors of learning outcomes, emphasizing the use of regression and supervised ML models. Noteworthy predictors of learning outcomes include online learning activities, term assessment grades, and the emotional state of the students during their academic journey.

ML techniques were evaluated [11] for forecasting students' final grades. They introduced a multiclass prediction model that integrated the Synthetic Minority Oversampling Technique (SMOTE) and feature selection methods, highlighting its potential to improve predictive performance. Being able to predict student performance in a timely manner empowers educators by enabling them to quickly identify underperforming students, which in turn facilitates early intervention and the implementation of essential support measures.

A guide for educators was provided [12] on the utilization of data mining methods to anticipate student performance in higher education. They categorized data mining analysis methods and proposed a systematic framework for educators. The use of synthetic educational data was investigated [13] in training academic performance prediction models. They distributed synthetic data to participants in data challenges, revealing challenges and limitations associated with prediction models in such contexts. The synthetic data was generated from a confidential dataset and distributed to participants in data challenges, facilitating the training of prediction algorithms. These participants submitted their models in Docker containers, which were then rigorously evaluated and ranked against separate synthetic datasets. Certain models that had been trained on synthetic data exhibited considerably diminished performance when applied to the non-synthetic dataset.

A systematic review of ML was conducted [14] in predicting student performance. They analysed 162 research articles and identified prevalent methodologies for prediction. The quintet of ML algorithms that reigned supreme comprised the Artificial Neural Network (ANN), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM). Furthermore, the bedrock features underpinning the prediction of students' performance included historical academic records, class performance, academic data sourced from learning management systems, and students' demographic information. A comparison of supervised data mining methods for the prediction of student exam performance was presented [15]. They highlighted the effectiveness of ANN and emphasized the importance of robust data acquisition and student engagement. Table 1 shows a summary of some key publications referred for the research.

Table 1. Summary of Related Work

Study	Methodology	Focus
GANs for Data Synthesis [4]	GANs, Rule-based methods	Data Synthesis, Privacy Preservation
GANs in Educational Technology [6]	COPULA-GAN	Synthetic Data's Compatibility
Advanced Statistical Analysis [7]	Multiple Linear Regression	Teaching Advanced Statistical Concepts
Enhanced Performance Prediction [8]	Enhanced CGAN, Deep Support Vector Machine (SVM)	Academic Success Prediction
Interpretable Models [9]	Stacked Ensemble, SHAP	Predicting Student Performance
Comprehensive Review ([10], [16])	Data Mining, Learning Analytics	Academic Performance Prediction
Improved Predictive Performance ([1],[11])	ML Techniques	Predicting Final Student Grades
Data Mining Guide [12]	Data Mining Techniques	Predicting Student Performance
Synthetic Data Challenges [13]	Synthetic Data Utilization	Challenges and Limitations
Systematic Review [14]	ML in Education	Prevalent Prediction Methodologies
Comparative Analysis [15]	Supervised Data Mining	Predicting Student Exam Performance
Predicting Dropout [17]	Deep Learning Methods	Student Dropout Prediction
Predictive Analytics in E-Learning ([18], [19], [20])	Predictive Analytics	Early Identification of At-Risk Students

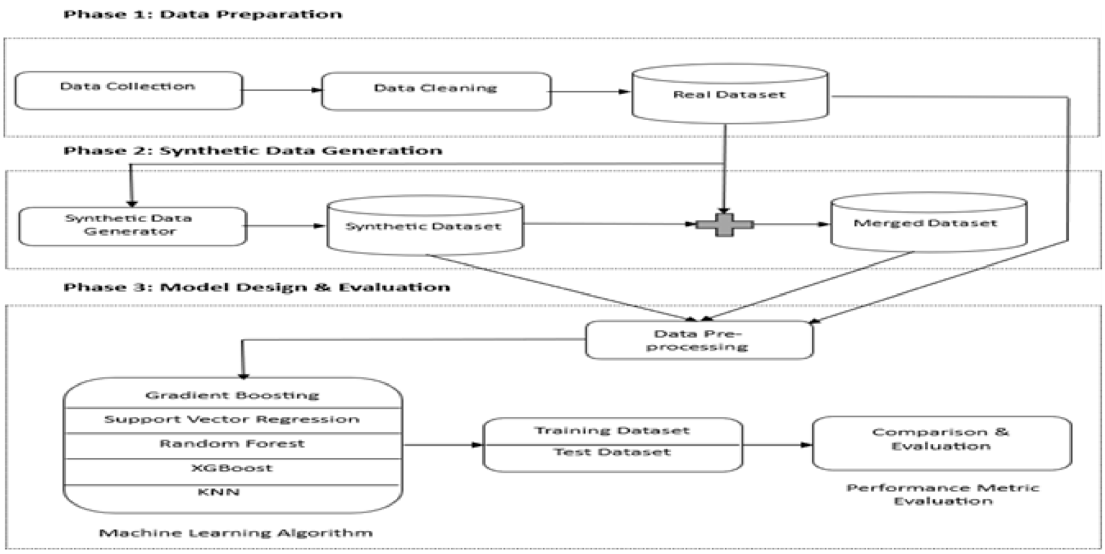

Fig. 1. Workflow Diagram

3. Methodology

This section outlines the methods employed in the study, including the data acquisition, Synthetic data generation, and rigorous model evaluation.

3.2 Dataset Description

The real-world educational dataset of 580 students is collected from Government Polytechnic of Karnataka, India comprising a diverse range of learner attributes, such as demographics, prior academic performance, and engagement metrics within online learning portals as mentioned in detail in [1]. The study utilizes a synthetic dataset of 5,000 records generated via Gretel API and a combination of real and synthetic datasets. Dataset is split in to five categories as below:

Learners' Background Data (P1): Incorporated within the learner's background data combination set are several crucial parameters like Matriculation Medium of Study, Residential Background (Rural/Urban), and Family Annual Income.

Experience with Prior Digital Learning Environment (P2): The P2 dataset included assessments of fundamental computer proficiency, online connectivity, and the user-friendliness of Learning Management Systems (LMS).

Interaction with Digital Learning Environment (P3): This includes Login Frequency Lectures Accessed Time Devoted to Viewing Online Lectures Time Allocated to Completing Online Assignments Activities Successfully Concluded Average Lecture Replay Frequency, and Average Lecture Viewing Interruptions.

Forum participation (P4): This includes Frequency of Inquiries, Peer Engagement, Instructor Interaction Group Activity Participation.

Lifestyle and Behavioral Metrics (P5): The dataset referred to as

P5 encompasses Physical Activity Frequency, Sleep Duration, Smartphone Usage for Educational Purposes, Dietary Preferences, and Library Visit Frequency.

3.3 System Overview

Fig. 1. shows an overview of the proposed system. It has been

divided into three phases:

3.3.1 Phase 1: Data Preparation

1. Data Collection: Data was collected from students through a multiple source on their experience with the Learning Management System (LMS), lifestyle, demographic information, and socioeconomic background.

2. Data Cleaning: Errors, inconsistencies, missing values, and outliers in the collected data were identified and addressed to ensure data quality and integrity.

3. Data Set: After performing data cleaning, a cleaned and prepared real data set was obtained, which included the survey responses collected from students.

3.3.2 Phase 2: Synthetic Data Generation

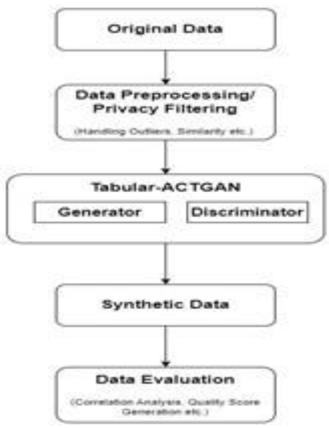

Fig. 2. Generation of Synthetic Dataset

Fig. 2. shows an overview of the generation of synthetic data. Tabular ACTGAN method is used to generate synthetic data using gretel.ai.

3.3.2.1 Synthetic Data Generator

Algorithm 1: Data Synthesis

Require: Original Data
Ensure: Synthetic Data
Initialise Parameters
1. Set the number of epochs automatically.
2. Define the generator neural network with dimensions [1024, 1024]
3. Specify the discriminator neural network with dimensions [1024, 1024].
4. Assign a learning rate of 0.0001 to the generator.
5. Set the discriminator's learning rate to 0.00033.
6. Determine the batch size automatically
Generate Synthetic Data
7. Specify the number of synthetic records to generate as 5000.
8. Apply privacy filters, including handling outliers and ensuring similarity.
Train the Model
9. Start training the model using the tabular-ACTGAN algorithm.
Evaluate the Synthetic Data
10. Calculate the number of columns used for correlations.
11. Generate a synthetic quality score report.
12. Identify mandatory columns (if any).

The parameter initialization step in Algorithm 1, plays a crucial role in configuring the training process, including the determination of training epochs, representing the iterations over the dataset for neural network training. The architecture of the generator and discriminator neural networks is specified as [1024, 1024], defining their structural design. The learning rates govern the speed at which these networks acquire knowledge from the data. Additionally, the batch size is automatically determined, representing the number of data samples utilized in each training iteration.

In step 2, the actual data generation takes place. The algorithm specifies the number of synthetic records to generate, which, in this case, is set to 5000 records. Privacy filters are applied in this step, which typically involves techniques to ensure that sensitive or personally identifiable information in the data is protected. This includes methods to handle outliers (extreme data points) and techniques to ensure that the synthetic data is similar in characteristics to the original data.

During the "Train the Model" phase, the algorithm initiates the training process for an ML model, using the tabular- ACTGAN (Anyway Conditional Tabular GAN). Tabular ACtGAN is an extension of the CTGAN (Conditional Tabular GAN) model and is used to generate synthetic tabular data that closely mimics the statistical characteristics of a provided dataset. This model is particularly useful in scenarios where data is scarce or sensitive and sharing it is restricted. The generator network employs random noise as input to generate synthetic data samples, which are then assessed by the discriminator network. The discriminator network's role is to learn how to distinguish between authentic and synthetic data samples. A distinctive feature of Tabular ACtGAN is its ability to control specific attributes or features of the generated data. The training process involves utilizing the original dataset in combination with synthetic data to instruct the model in understanding and replicating the statistical patterns inherent in the original dataset.

After the model has been trained, the generated synthetic data will be evaluated as in Fig. 3. and Fig. 4. It defines the number of columns used for correlation analysis, reporting, the maximum number of rows in the report, and other evaluation-related settings such as target variables and metrics. Additionally, a synthetic quality score report and data summary statistics are generated. The algorithm also identifies any mandatory columns, essential variables, or attributes that must be present in the synthetic data.

3.3.2.2 Merging Data Sets: The generated synthetic data was combined with the real data set, creating a merged data set that encompassed both real and synthetic data. This integration ensured a diverse and comprehensive data set for subsequent analysis.

Fig. 3. Report for generated synthetic data

Fig. 4. Data Summary Statistics

Table 2. *Accuracy Metrics*

Metric	Formula	Description
R-Squared Co-efficient (R^2)	$\dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$	Quantifies the extent to which the variation in the dependent variable can be explained by the independent variables. A measure of the regression model's goodness of fit. Higher the R^2 value stronger fit between the model and the data.
Root Mean Square Error (RMSE)	$\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Measures the average magnitude of the errors between predicted ($\hat{y}_i$) and actual (y_i) values. Smaller RMSE values indicate better model performance.
Mean Absolute Error (MAE)	$\dfrac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Computes the mean of the absolute disparities between predicted ($\hat{y}_i$) and observed (y_i) values, serving as an indicator of the model's typical prediction inaccuracy.

Phase 3: Model Design and Evaluation

1. Data Preprocessing: This phase involved the usage of feature selection techniques like filter and wrapper and handling categorical variables through encoding.

2.Appling ML algorithm: ML regression algorithms, including RF, Gradient Boosting Regression Trees (GBRT), eXtreme Gradient Boosting (XGB), K- Nearest Neighbor (KNN), and Support Vector Regression (SVR), were employed to train and model the preprocessed datasets. These algorithms aim to learn a mapping function that can predict the target variable (performance in this context) based on input features.

a. RF: RF is an ensemble learning technique that relies on decision trees. The mathematical representation RF (X) is expressed in (1).

Let:

N be the number of decision trees in the forest.

T_i represents the prediction made by i-th decision tree.

X denotes the input features.

$$RF(X) = \frac{1}{N}\sum_{i=1}^{N} T_i(X) \quad (1)$$

b. GB: It's an ensemble technique that builds an additive model by combining numerous weaker learners, typically in the form of decision trees. The mathematical representation of GB(X), prediction made by the GB model for input X is expressed in (2).

$$GB(X) = \sum_{m=1}^{M} h_m(X) \quad (2)$$

Where:

M is the number of boosting iterations.

$h_m(X)$ is the prediction of the m-th weak learner.

c. XGB: XGB works by minimizing a loss function that measures the disparity between the actual target values (Y) and the predictions generated by an ensemble of decision trees. In regression, the typical choice for this loss function is the Mean Squared Error (MSE). The objective function is to find optimal prediction function $F_m(X)$ at each boosting iteration by minimizing this objective function and is given by (3).

$$Objective(M) = \Sigma_i L(y_i, Fm-1(x_i)) + \Omega(Fm) \quad (3)$$

Where:

M be the number of boosting iterations.

h_m be the m-th weak learner.

Fm−1(X) represent the ensemble's prediction at iteration m − 1.

L(Y, Fm−1(X)) be the loss function that quantifies the difference between the

true target Y and the current prediction Fm−1(X).

Ω(Fm) is the regularization term that penalizes model complexity.

d. KNN: KNN is a non-parametric instance-based learning method. When presented with a new data point, it locates the k training examples that are most similar to it in feature space and derives a prediction for the target variable by considering the majority class among these k nearest neighbors. The prediction KNN (X) for a new data point X can be represented as in (4).

$$KNN(x) = \frac{1}{k}\sum_{i=1}^{k} y_i \quad (4)$$

Where:
k be the number of nearest neighbors.

y_i represent target values.

e. SVR: SVR is a supervised learning technique employed for regression tasks The objective in S is to find the optimal hyperplane that minimizes the prediction error while staying within a specified margin (ϵ-tube) around the target values and is given by (5).

Minimize:

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \varepsilon_i^*) \quad (5)$$

Subject to:

yi − (w. ϕ(xi)+b) ≤ϵ + ξi

(w⋅ϕ(xi)+b) −yi ≤ ε+ξi∗

ξi, ξi ∗≥0 and i=1, 2, …., N

Where:

w represents the weight vector.

b is the bias term.

ξi and ξi ∗ are slack variables that quantify the prediction error.

C is the cost parameter that balances the trade-off between minimizing error and

ensuring data points are within the margin.

ϵ specifies the margin size.

yi are the target values.

ϕ(xi) represents the feature mapping, often involving a kernel function.

3.Training and Test Data Split: The preprocessed datasets were meticulously divided 80% of the data for model training and remaining 20% for testing.

4. Comparison and Evaluation: The outcomes derived from various datasets, including real, synthetic, and merged data, underwent a comprehensive comparative assessment through the

4. Results

Table 3: Comparison of real, synthetic, and mixed dataset on single feature set

Feature set	Algorithm	Real Dataset			Synthetic Dataset			Mixed Dataset		
		RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
	RF	14.09	10.88	0.8248	13.08	10.30	0.8415	11.54	8.05	0.8739
P1	XGB	14.31	10.88	0.8250	13.08	10.30	0.8415	11.54	8.06	0.8739
	KNN	12.92	10.36	0.8363	13.44	10.72	0.8297	12.64	9.73	0.8528
	SVR	13.02	10.21	0.8313	13.08	10.24	0.8420	11.54	7.89	0.8763
	GBRT	13.96	10.92	0.8240	13.09	10.31	0.8414	11.53	8.03	0.8742

	RF	14.49	11.59	0.8140	13.11	10.33	0.8414	11.64	8.20	0.8716
P2	XGB	14.54	11.62	0.8135	13.10	10.32	0.8414	11.65	8.21	0.8716
	KNN	16.32	12.61	0.7933	14.14	11.34	0.8244	12.70	9.64	0.8491
	SVR	13.00	10.09	0.8348	13.05	10.17	0.8433	11.55	7.90	0.8761
	GBRT	14.24	11.26	0.8181	13.10	10.31	0.8417	11.61	8.10	0.8730
	RF	12.30	9.42	0.8490	13.98	11.18	0.8293	12.13	8.98	0.8607
P3	XGB	14.25	10.86	0.8272	15.54	12.34	0.8121	12.75	9.73	0.8499
	KNN	13.56	10.66	0.8306	14.16	11.26	0.8289	12.40	9.39	0.8554
	SVR	12.68	9.93	0.8371	13.05	10.24	0.8423	11.55	7.91	0.8760
	GBRT	13.05	10.07	0.8390	13.32	10.59	0.8376	11.56	8.24	0.8712
	RF	15.60	11.95	0.8110	14.42	11.52	0.8253	12.47	9.25	0.8576
P4	XGB	19.35	14.83	0.7710	14.22	11.34	0.8273	12.38	9.11	0.8597
	KNN	13.14	9.39	0.8478	14.18	11.50	0.8260	12.43	9.34	0.8564
	SVR	12.53	9.58	0.8434	13.10	10.26	0.8416	11.55	7.88	0.8763
	GBRT	14.61	10.92	0.8277	13.34	10.55	0.8380	11.59	8.12	0.8729
	RF	13.86	10.60	0.8319	14.80	11.71	0.8212	12.97	9.86	0.8485
P5	XGB	13.90	10.72	0.8274	14.67	11.69	0.8214	12.70	9.51	0.8536
	KNN	13.71	10.71	0.8253	14.39	11.48	0.8239	12.77	9.77	0.8507
	SVR	12.64	9.77	0.8374	13.04	10.21	0.8426	11.54	7.89	0.8763
	GBRT	12.90	9.74	0.8446	13.06	10.24	0.8419	11.64	8.29	0.8706

trained ML model. A diverse set of evaluation metrics as in Table 2 was systematically employed.

RQ 1: Does the Combination of Different Feature Sets Enhance Predictive Models for Academic Performance in Real, Synthetic, and Mixed Dataset?

Single feature set: In single feature sets, RMSE values spanned from 11.53 to 19.35, indicating variability in model performance across different feature sets. The highest RMSE (19.35) was observed with the XGB model for feature set "P4" on the real dataset, indicating a sensitivity to the variability in single feature sets. The lowest RMSE (11.53) was noted with the GBRT model for feature set "P1" on the mixed dataset, highlighting the strength of GB methods when demographic and background data are incorporated. Overall, the mixed dataset consistently produced superior outcomes, highlighting the value of incorporating synthetic data to bolster predictive models, as substantiated by the data in Table 3 and visually by Fig. 5.

Twin Feature set: When two parameter sets were taken together, RF and XGB consistently performed well across different datasets as shown in Table 4, with RF often having a slight edge in terms of RMSE and MAE. KNN and SVR also exhibited competitive performance, and GB stood out in some cases, particularly in the "mixed" dataset as in Fig. 6. P1_P2, the RF algorithm achieves an RMSE of 12.11 on the mixed dataset, markedly lower than 15.94 on the real dataset, indicating the added value of integrating synthetic data for a more robust predictive model. Similarly, the SVR algorithm stands out with consistent performance, particularly in feature set P1_P2, where it achieves an RMSE of 11.55 on the mixed dataset, one of the lowest across all combinations. This points to SVR's strength in handling diverse data inputs. effectively. The RMSE range for twin feature sets varies with the lowest observed for SVR in the P1_P2 combination on the mixed dataset (11.55) and the highest for XGB in the P1_P4 combination on the real dataset (19.00). The mixed dataset repeatedly results in enhanced model performance.

Triple Feature set: Across triple feature set combinations, models trained on mixed datasets consistently outperform those trained solely on real or synthetic datasets, reinforcing the proposition that a combination of different feature sets can indeed enhance predictive accuracy as shown in Fig. 7. For instance, when considering the feature set P1_P2_P3, the RF algorithm delivers the lowest RMSE (11.87) on the mixed dataset, markedly improving from 12.54 on the real dataset. This suggests the amalgamation of real and synthetic data yields a more accurate model, as highlighted by the increased R^2 (0. 8656 for the mixed dataset versus 0.8444 for the real dataset). It is noteworthy that while the GBRT model shows heightened accuracy on mixed datasets, its performance is closely rivalled by the RF model, which offers consistent RMSE improvements across most combinations. The SVR model also demonstrates robust performance, particularly in the mixed dataset context, suggesting

Table 4. Comparison of real, synthetic, and mixed dataset on twin feature set

Feature set	Algorithm	Real Dataset			Synthetic Dataset			Mixed Dataset		
		RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
P1_P2	RF	15.94	12.63	0.7986	13.87	11.04	0.831	12.11	8.80	0.8634
	XGB	17.79	14.24	0.7779	14.05	11.18	0.8291	12.10	8.80	0.8638
	KNN	14.93	12.00	0.8084	14.55	11.59	0.8218	12.42	9.28	0.8556
	SVR	13.31	10.45	0.8282	13.10	10.26	0.8416	11.55	7.89	0.8762
	GBRT	15.76	12.86	0.7959	13.23	10.46	0.8394	11.64	8.22	0.8716
P1_P3	RF	12.17	9.51	0.8495	13.78	11.04	0.8317	12.09	8.88	0.8622
	XGB	14.37	11.34	0.8221	15.28	12.31	0.8134	13.02	9.93	0.8473
	KNN	13.80	11.05	0.8243	14.21	11.28	0.8281	12.61	9.59	0.8520
	SVR	12.76	10.09	0.8333	13.13	10.30	0.8413	11.54	7.89	0.8761
	GBRT	12.99	10.31	0.8362	13.31	10.60	0.8375	11.55	8.23	0.8715
P1_P4	RF	15.61	12.09	0.8083	14.56	11.66	0.8236	12.69	9.62	0.8517
	XGB	19.00	14.31	0.7779	14.63	11.55	0.8248	12.65	9.47	0.8544
	KNN	14.32	11.05	0.8202	14.42	11.53	0.8251	12.37	9.44	0.8548
	SVR	12.61	9.62	0.8405	13.10	10.27	0.8415	11.55	7.89	0.8762
	GBRT	17.05	12.59	0.8028	13.31	10.55	0.8380	11.64	8.24	0.8712
P1_P5	RF	14.16	11.24	0.8211	14.34	11.42	0.8260	12.91	9.70	0.8509
	XGB	16.19	12.70	0.7993	14.71	11.69	0.8216	12.68	9.61	0.8523
	KNN	14.00	10.59	0.8260	13.98	11.18	0.8294	12.62	9.46	0.8541
	SVR	12.89	9.90	0.8351	13.07	10.26	0.8418	11.54	7.90	0.8761
	GBRT	14.00	11.20	0.8216	13.09	10.30	0.8416	11.62	8.29	0.8707
P2_P3	RF	12.56	10.07	0.8417	13.46	10.82	0.8352	11.86	8.71	0.8645
	XGB	14.19	10.94	0.8262	15.07	12.14	0.8154	12.82	9.74	0.8497
	KNN	13.89	10.77	0.8259	14.00	11.31	0.8277	12.49	9.46	0.8535
	SVR	12.92	10.05	0.8351	13.05	10.24	0.8422	11.54	7.91	0.8759
	GBRT	12.90	10.01	0.8417	13.32	10.59	0.8378	11.67	8.31	0.8699
P2_P4	RF	14.20	11.29	0.8229	14.87	12.09	0.8192	12.73	9.53	0.8535
	XGB	16.84	13.49	0.7884	15.11	12.10	0.8185	12.78	9.70	0.8510
	KNN	12.95	10.44	0.8342	14.71	11.83	0.8214	12.62	9.47	0.8541
	SVR	12.83	9.90	0.8379	13.10	10.28	0.8411	11.55	7.90	0.8761
	GBRT	14.18	11.06	0.8242	13.39	10.64	0.8372	11.68	8.20	0.8718

	RF	13.34	10.52	0.8322	14.35	11.45	0.8249	12.57	9.45	0.8538
	XGB	14.79	11.67	0.8116	15.07	12.08	0.8158	12.98	9.89	0.8476
P2_P5	KNN	14.02	10.60	0.8233	14.41	11.54	0.8236	12.93	9.77	0.8496
	SVR	12.71	9.78	0.8383	13.06	10.26	0.8417	11.55	7.91	0.8760
	GBRT	12.67	9.93	0.8442	13.25	10.44	0.8392	11.68	8.33	0.8698
	RF	12.28	9.58	0.8464	13.56	10.82	0.8351	11.71	8.55	0.8668
	XGB	14.22	10.77	0.8319	15.20	12.25	0.8139	12.70	9.73	0.8500
P3_P4	KNN	13.68	10.81	0.8246	14.52	11.46	0.8249	12.56	9.62	0.8534
	SVR	13.06	10.19	0.8317	13.09	10.26	0.8416	11.54	7.91	0.8759
	GBRT	12.93	9.95	0.8444	13.40	10.60	0.8375	11.56	8.23	0.8710
	RF	12.66	10.21	0.8406	13.44	10.70	0.8363	11.74	8.55	0.8672
	XGB	14.05	11.16	0.8281	15.26	12.13	0.8137	12.58	9.62	0.8523
P3_P5	KNN	13.34	10.79	0.8257	14.49	11.57	0.8230	12.66	9.59	0.8516
	SVR	12.63	9.86	0.8370	13.04	10.22	0.8424	11.54	7.91	0.8759
	GBRT	13.46	10.31	0.8376	13.30	10.51	0.8381	11.62	8.30	0.8705
	RF	13.68	11.03	0.8270	13.50	10.81	0.8350	11.97	8.82	0.8634
	XGB	15.02	12.26	0.8112	14.85	11.88	0.8197	12.84	9.77	0.8499
P4_P5	KNN	13.85	11.04	0.8219	14.31	11.49	0.8249	12.60	9.61	0.8515
	SVR	12.48	9.64	0.8405	13.08	10.30	0.8410	11.54	7.91	0.8760
	GBRT	14.14	11.14	0.8268	13.22	10.47	0.8389	11.64	8.25	0.8711

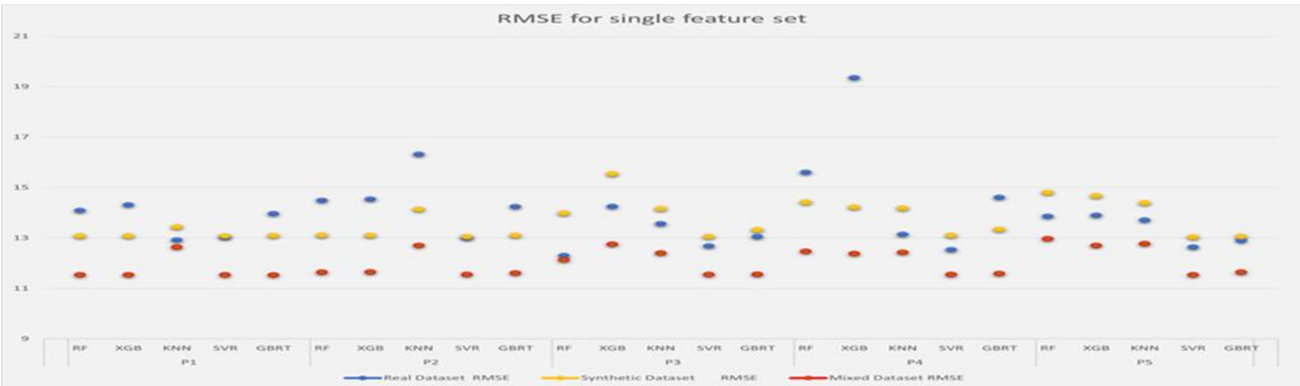

Fig. 5. RMSE comparison of single feature set on different datasets and algorithms

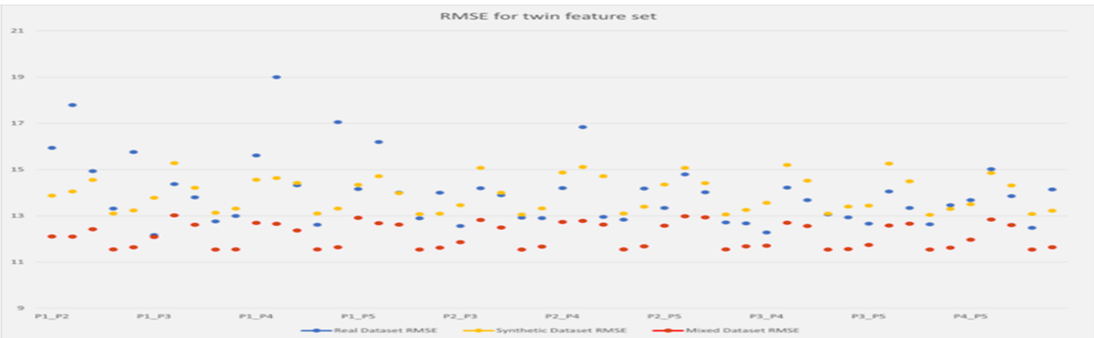

Fig. 6. RMSE comparison of twin feature set on different datasets and algorithms

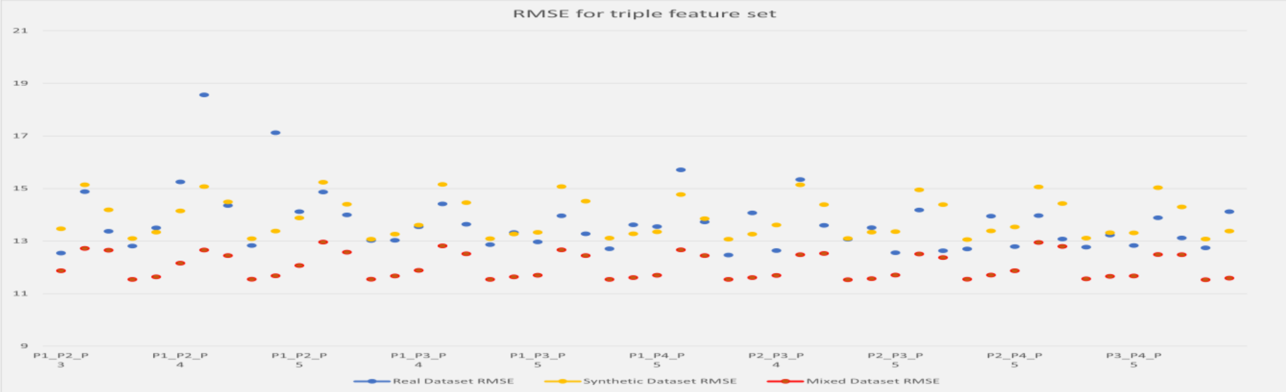

Fig. 7. RMSE comparison of triple feature set on different datasets and algorithms

its effective handling of composite data inputs. The RF model stands out as a particularly effective algorithm across various combinations, making it a strong candidate for academic performance prediction tasks in blending learning environments.

Four and Five Feature set: Table 5 and Fig. 7. indicates that the RF algorithm consistently outperforms other models across quad and five-feature sets. While the XGB model shows promise, especially in mixed datasets, it does exhibit higher RMSE values in more complex feature combinations, suggesting possible limitations in handling intricate data structures. KNN remains a viable model, with performance that closely follows the RF model, especially in mixed datasets where data diversity is inherent. SVR maintains commendable accuracy levels, although it presents slightly higher RMSE figures in mixed datasets, hinting at a trade-off between error rate and accuracy.

GBRT, while showing moderate increases in error metrics, secures the highest accuracy rates in mixed dataset conditions, reinforcing the benefits of feature diversity in predictive modeling. The range of RMSE values observed spans from 11.53 to 17.34 for the quad feature sets and from 11.54 to 15.35 for the five-feature sets. Notably, the most comprehensive feature set, "P1_P2_P3_P4_P5," when processed through the RF algorithm and applied to mixed datasets, achieved an optimal balance between complexity and accuracy, marking the lowest RMSE value of 11.54, showcasing the effectiveness of the RF model in complex modeling scenarios.

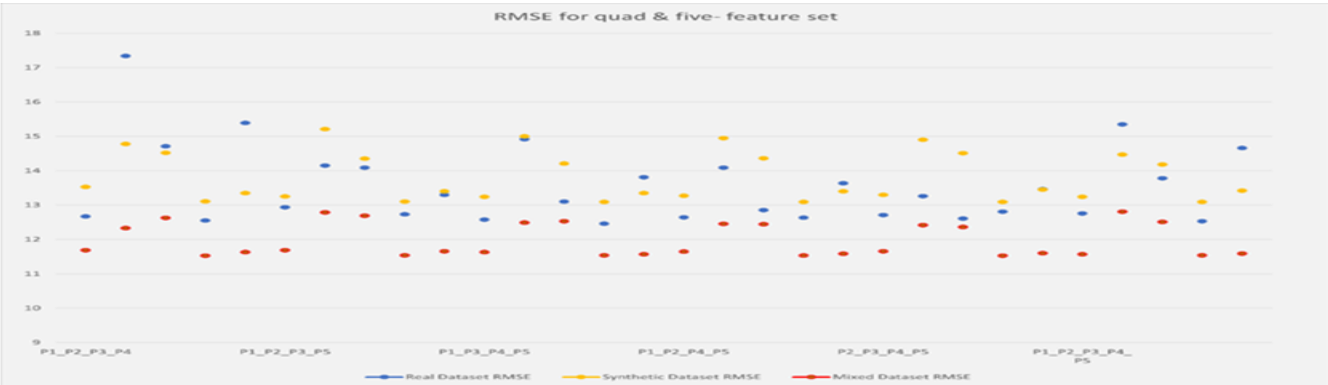

Fig. 8. RMSE comparison quad and five feature set on different datasets and algorithms

Table 5. Comparison of real, synthetic, and mixed dataset on quad and five-feature set

Feature set	Algorithm	Real Dataset			Synthetic Dataset			Mixed Dataset		
		RMSE	MAE	R2	RMSE	MAE	R2	RMSE	MAE	R2
P1_P2_P3_P4	RF	12.67	10.32	0.8373	13.53	10.76	0.8363	11.69	8.49	0.8680
	XGB	17.34	13.34	0.7973	14.78	11.91	0.8189	12.33	9.34	0.8560
	KNN	14.71	11.78	0.8112	14.52	11.62	0.8231	12.63	9.65	0.8520
	SVR	12.55	9.84	0.8378	13.11	10.30	0.8409	11.53	7.92	0.8759
	GBRT	15.39	12.08	0.8137	13.35	10.56	0.8382	11.63	8.29	0.8703
P1_P2_P3_P5	RF	12.94	10.55	0.8340	13.25	10.58	0.8381	11.69	8.46	0.8683
	XGB	14.15	11.17	0.8257	15.21	12.26	0.8130	12.79	9.64	0.8509
	KNN	14.09	10.86	0.8245	14.35	11.53	0.8240	12.69	9.67	0.8511

	SVR	12.73	9.98	0.8351	13.10	10.28	0.8413	11.54	7.93	0.8757
	GBRT	13.30	10.94	0.828	13.40	10.62	0.8369	11.66	8.33	0.8699
	RF	12.58	10.42	0.8362	13.24	10.55	0.8389	11.63	8.39	0.8696
	XGB	14.92	11.64	0.8202	15.00	12.02	0.8180	12.49	9.53	0.8529
P1_P3_P4_P5	KNN	13.10	10.38	0.8316	14.21	11.37	0.8260	12.53	9.56	0.8529
	SVR	12.46	9.70	0.8392	13.09	10.28	0.8413	11.54	7.92	0.8757
	GBRT	13.81	11.04	0.8304	13.35	10.57	0.8380	11.57	8.31	0.8704
	RF	12.65	10.4	0.8369	13.27	10.55	0.8389	11.65	8.37	0.8697
	XGB	14.09	11.13	0.8289	14.95	12.01	0.8180	12.46	9.48	0.8540
P1_P2_P4_P5	KNN	12.86	10.22	0.8351	14.36	11.44	0.8252	12.45	9.46	0.8545
	SVR	12.64	9.85	0.8373	13.09	10.28	0.8413	11.54	7.92	0.8758
	GBRT	13.64	10.87	0.8325	13.4	10.59	0.8376	11.59	8.3	0.8704
	RF	12.71	10.38	0.8375	13.30	10.55	0.8388	11.66	8.35	0.8698
	XGB	13.26	10.62	0.8376	14.90	12.00	0.8180	12.42	9.43	0.8550
P2_P3_P4_P5	KNN	12.61	10.06	0.8386	14.51	11.51	0.8243	12.36	9.35	0.8561
	SVR	12.81	9.99	0.8353	13.09	10.28	0.8412	11.53	7.92	0.8758
	GBRT	13.46	10.69	0.8345	13.45	10.61	0.8371	11.60	8.29	0.8704
	RF	12.76	10.63	0.8329	13.24	10.50	0.8397	11.57	8.33	0.8703
	XGB	15.35	12.31	0.8125	14.47	11.53	0.8246	12.81	9.69	0.8504
P1_P2_P3_P4_P5	KNN	13.78	10.88	0.8246	14.18	11.29	0.8283	12.51	9.41	0.8550
	SVR	12.53	9.77	0.8384	13.09	10.28	0.8412	11.54	7.93	0.8756
	GBRT	14.66	11.63	0.8220	13.42	10.61	0.8372	11.59	8.32	0.8701

RQ2. What is the comparative predictive performance of models across actual, generated, and augmented datasets for academic performance prediction in BL?

Fig. 9. consistently illustrates a clear trend in which the mixed dataset surpasses the synthetic dataset, and the synthetic dataset outperforms the real dataset across a range of machine learning algorithms. This trend underscores the effectiveness of combining real and synthetic data for predicting learner performance in BL environments. Specifically, the SVR model achieves the highest R^2 Co-efficient of 0.8756 when applied to the mixed dataset, indicating its superior performance. Thus, mixed dataset encompasses a wider array of scenarios and learner data variations, enhancing the algorithms' predictive capacity. The diversity and increased data volume offer richer insights, resulting in improved accuracy for all algorithms tested.

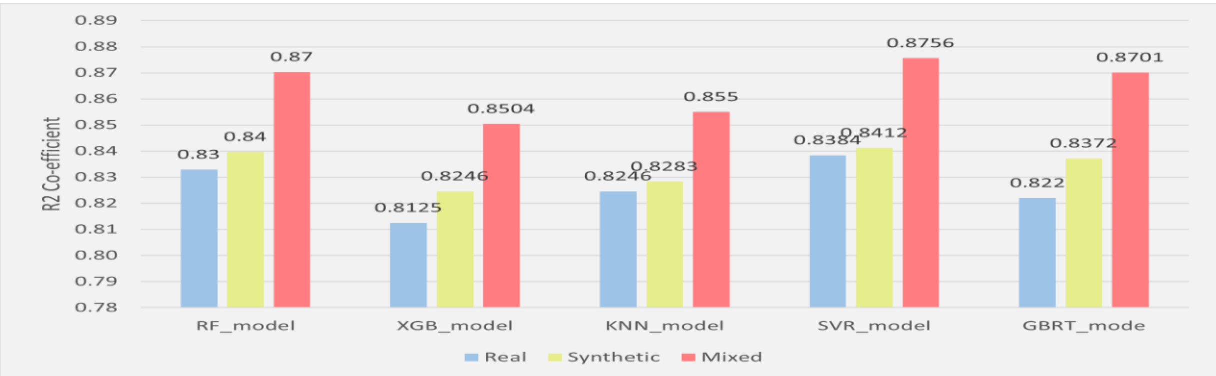

Fig. 9. Comparison of accuracy on P1_P2_P3_P4_P5 feature set on different dataset and algorithms

4. Conclusion and Future scope

This work provides a comprehensive comparative analysis between real and synthetic data, generated via the Tabular-ACTGAN-based algorithm, for synthesizing high-quality data while ensuring privacy protection for learner performance prediction in online learning portals. The findings suggest that synthetic data shows promise as a viable alternative to real data, with ML models trained on synthetic data demonstrating competitive performance. The mixed dataset showcases a notable advantage, where ML models trained on this hybrid data exhibit even more robust and accurate performance.

Future research directions include refining the techniques for generating high-quality synthetic data, exploring the transferability of models trained on mixed data to real-world scenarios, addressing biases, and ensuring fairness in synthetic data generation, along with extending our analysis to predict long-term learner performance. Additionally, the success of mixed data integration encourages further investigation into innovative data synthesis approaches, reinforcing the importance of data quality and privacy while advancing the field of ML for educational purposes.

References

[1] S. J. Shabnam Ara, R. Tanuja and S. H. Manjula, "Regression-Driven Predictive Model to Estimate Learners' Performance through Multisource Data," in International Conference on Futuristic Technologies (INCOFT), Belagavi, Karnataka, India, pp. 1-6, 2023, DOI: 10.1109/INCOFT60753.2023.10425033.

[2] I. Goodfellow et al., "Generative Adversarial Nets," arXiv preprint arXiv:1406.2661, 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[3] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 308-318.

[4] L. Wang et al., "A State-of-the-Art Review on Image Synthesis with Generative Adversarial Networks," IEEE Access, vol. 8, pp. 63514–63537, 2020.

[5] V. Bindschaedler and R. Shokri, "Synthesizing Plausible Privacy-Preserving Location Traces," in 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2016, pp. 546-563.

[6] A. Bethencourt-Aguilar, D. Castellanos-Nieves, J. J. Sosa-Alonso, and M. Area-Moreira, "Use of Generative Adversarial Networks (GANs) in Educational Technology Research," 2023.

[7] L. L. Murray and J. G. Wilson, "Generating data sets for teaching the importance of regression analysis," Decision Sciences Journal of Innovative Education, vol. 19, no. 2, pp. 157-166, 2021.

[8] S. Sarwat et al., "Predicting Students' Academic Performance with Conditional Generative Adversarial Network and Deep SVM," Sensors, vol. 22, no. 13, p. 4834, 2022.

[9] M. Hoq, P. Brusilovsky, and B. Akram, "Analysis of an Explainable Student Performance Prediction Model in an Introductory Programming Course," International Educational Data Mining Society, 2023.

[10] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," Applied Sciences, vol. 11, no. 1, p. 237, 2020.

[11] S. D. A. Bujang et al., "Multiclass prediction model for student grade prediction using machine learning," IEEE Access, vol. 9, pp. 95608-95621, 2021.

[12] E. Alyahyan and D. Dustegor, "Predicting academic success in higher education: literature review and best practices," International Journal of Educational Technology in Higher Education, vol. 17, pp. 1-21, 2020.

[13] B. Flanagan, R. Majumdar, and H. Ogata, "Fine grain synthetic educational data: challenges and limitations of collaborative learning analytics," IEEE Access, vol. 10, pp. 26230-26241, 2022.

[14] K. Alalawi, R. Athauda, and R. Chiong, "Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review," Engineering Reports, vol. e12699, 2023.

[15] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An Overview and Comparison of Supervised Data Mining Techniques for Student Exam Performance Prediction," Computers and Education, vol. 143, Article ID: 103676, 2019.

[16] S.J. Shabnam Ara, R. Tanuja, S. H. Manjula, K.R Venugopal, "A Comprehensive Survey on Usage of Learning Analytics for Enhancing Learner's Performance in Learning Portals," Journal of Educational Technology Systems, vol. 52, no. 2, pp: 245-73, 2023.

[17] F. Giannakas, C. Troussas, I. Voyiatzis, C. Sgouropoulou, "A deep learning classification framework for early prediction of team-based academic performance," Applied Soft Computing. vol. 106, pp:107355, 2021.

[18] I. EI Guabassi, Z. Bousalem, R. Marah, and A. Qazdar, "Comparative Analysis of Supervised Machine Learning Algorithms to Build a Predictive Model for Evaluating Students' Performance," 2021.

[19] L. Ismail, H. Materwala, and A. Hennebelle, "Comparative Analysis of Machine Learning Models for Students' Performance Prediction," in T. Antipova (Ed.), Advances in Digital Science, ICADS 2021, Advances in Intelligent Systems and Computing, vol. 1352, Springer, Cham, pp. 157-166, 2021.

[20] L. Zhao, K. Chen, J. Song, X. Zhu, J, Sun, B. Caulfield, and B. Mac Namee, "Academic performance prediction based on multisource, multi-

feature behavioral data," IEEE Access, vol. 9, pp. 5453-5465, 2020.

Acknowledgements: None

Author contributions

Shabnam Ara S. J: Conceptualization, Methodology, Implementation, Analysis, Writing-Original draft preparation. **Tanuja R:** Reviewed the Manuscript.

Manjula S H: Reviewed the Manuscript.

Conflicts of interest

The authors declare no conflicts of interest.