Predicting Student Performance Using Synthetic Big Data and XGBoost

Anonymous Author(s)

Abstract

Predicting student academic performance is a vital task in educational data mining, supporting early intervention and targeted resource allocation. However, real-world educational datasets often suffer from data scarcity and privacy limitations, making large-scale analysis difficult. To address these challenges, we created a synthetically generated student performance dataset scaled to 25 million records using a Gaussian Copula-based synthesizer. This enabled simulation of a big data environment while preserving confidentiality and protecting personally identifiable information (PII). We trained a classification model using XGBoost to predict student GPA categories from a combination of demographic, socioeconomic, and academic features. Dimensionality reduction techniques—including Principal Component Analysis for numerical features and Chi-Squared tests for categorical selection—were applied to improve efficiency and reduce noise. All modeling was conducted within Google Colab using DuckDB to manage large-scale data access and maintain performance within memory-constrained environments. Evaluation across both synthetic and original test sets showed that the model trained on synthetic data achieved comparable performance, confirming that key predictive patterns and feature relationships were retained. This suggests that high-fidelity synthetic data can be a viable alternative for privacy-preserving machine learning in education. Our work highlights the potential for scalable, ethical, and reproducible pipelines that integrate synthetic data generation with state-of-the-art classification models in big data settings.

Keywords

Synthetic Data, XGBoost, Educational Data Mining, Student Performance Prediction, Big Data, Privacy Preservation, Feature Selection, Dimensionality Reduction

1 Overview

Educational institutions increasingly leveraged data-driven methods to predict and improve student success. *Educational Data Mining* (EDM) applied machine learning (ML) and data mining techniques to student records to uncover hidden patterns and predict academic results [20]. Previous studies had shown that analyzing characteristics such as demographics, socioeconomic status, and past grades could moderately predict student performance [20].

However, assembling a large-scale real-world student dataset posed challenges due to privacy constraints and data scarcity. High-quality educational data often contained sensitive personal information (e.g., grades, background), raising ethical and legal concerns about student privacy [13]. In addition, collecting millions of student records from various sources was time-consuming and costly.

To address these issues, we utilized a synthetic student performance data set obtained from Kaggle (contributed by the user 'NeuralSorcerer'). Synthetic data were artificially generated records that preserved the statistical patterns of real data without exposing real individuals. Recent research highlighted that synthetic data could effectively overcome data scarcity, privacy concerns, and high collection costs in ML applications [17]. In the education domain specifically, synthetic data generation was seen as a promising privacy-preserving approach, allowing researchers to share and analyze student data while protecting confidentiality [17].

In this project, we used a synthetic high-school performance dataset to build predictive models of student GPA. We further simulated a big data scenario by scaling the dataset to ∼25 million records. The data were loaded into a relational database and accessed via Python in Google Colab, allowing us to handle the volume by querying in chunks. This approach provided a testbed for big data analytics techniques in education, letting us evaluate model performance and scalability on a massive dataset without real student records.

Our primary goal was to classify student academic outcomes (e.g., high or low GPA) based on their input characteristics. We focused on the XGBoost classifier for this task, given its strong track record in tabular data competitions and applications [6]. The following sections discuss relevant literature (Section 2), describe the dataset and pre-processing steps (Section 3), and detail our methodology including the justification for synthetic data and the choice of XGBoost (Section 4).

2 Related Work

Student Performance Prediction: Predicting academic success has been widely studied in EDM. Diverse machine learning techniques have been applied to student datasets, including decision trees, random forests, support vector machines (SVM), and logistic regression [20]. These models use student attributes (e.g. attendance, prior grades, socio-demographics) to forecast outcomes such as course grades or GPA. For instance, earlier work by Cortez and Silva (2008) modeled secondary school grades using demographic and grade features, demonstrating that data-driven models can identify at-risk students early [7]. Recent surveys confirm that EDM has become an effective tool to predict academic achievement and inform interventions [20]. Typical classification accuracies reported range from about 0.5 to 0.8 (50–80%), depending on the algorithms and features used [20]. Moreover, ensemble methods like gradient boosting have shown promise; Fernandes et al. (2019) used Gradient Boosting Machines on demographic and prior performance features to predict student grades, finding that prior grades and attendance were the most important predictors of final achievement [10]. These studies underscore the potential of ML models to improve educational outcomes, but they generally rely on relatively small datasets (hundreds or thousands of students).

Synthetic Data in Education: Due to legitimate privacy concerns, there is a growing interest in using synthetic educational data. Researchers have begun to explore techniques such as generative models and privacy-preserving frameworks to create artificial

student records that mirror real data distributions [16]. Liu et al. (2025) introduce a framework combining synthetic data generation with differential privacy to safely share student data for research, demonstrating that such data can maintain utility for analysis while preventing privacy leaks [17].

Synthetic data offer the advantage of scale and diversity without infringing on student privacy regulations (e.g. FERPA or GDPR). Our work builds on this concept by using an openly available synthetic dataset from Kaggle. By scaling it up to 25 million samples, we emulate a scenario of "big data" in education. This approach aligns with observations in the literature that synthetic data can address the lack of large real datasets [16]. It also follows best practices suggested in recent overviews of synthetic data research, which emphasize generating realistic, diverse artificial datasets to enable model training at scale [16].

3 Analysis

Dataset Description: The project used a synthetic Student Performance dataset (initially containing 10 million data-points) created by NeuralSorcerer on Kaggle. Each record corresponded to a high school student and included 21 attributes spanning demographic, socio-economic, and academic factors. Examples of features were: gender, age, parental education levels, socio-economic indicators (e.g. family income proxy), school-related information (type of school, class attendance, extracurricular activities), and past academic performance metrics. The target variable was the student's Grade Point Average (GPA) or a categorical label derived from it (e.g. classifying students into performance bands such as high/medium/low GPA). For our classification task, we discretized the continuous GPA values into four ordinal categories based on specific numeric thresholds: Low (0.00–2.49), Mid-Low (2.50–2.99), Mid-High (3.00–3.49), and High (3.50–4.00). These categories reflect a gradient of academic performance levels and were defined to enable interpretable multi-class classification.

To simulate a *big data* environment, we use the Synthetic Data Vault [8] to generate 50 million rows of synthetic data from an original dataset of approximately 10 million rows using a Gaussian Copula synthesizer. This enables us to prototype scalable pipelines, benchmark model performance, and stress-test our database infrastructure under large-scale conditions. We use DuckDB as our in-process analytical database, interfaced via the Ibis framework [12], allowing for efficient querying, batch insertion, and seamless integration with our Python-based pipeline.

As shown in Figure 1, the data is normalized into seven separate tables to reduce redundancy and improve modularity. Categorical attributes such as Gender, Race, and Parental Education are separated into dimension tables to enforce consistency and simplify encoding and updates. School and Extracurricular data are modeled as distinct tables due to their one-to-many or many-to-many relationships with students. Academic performance metrics are stored in a dedicated table to allow for more efficient slicing, aggregation, and downstream predictive modeling.

Figure 2, showcases the flow of data from data generation to model creation. As stated earlier, data is first generated using SDV, which is processed using Ibis. This is then saved as a CSV file, which is both loaded into Docker via initDB scripts, as well as loaded into

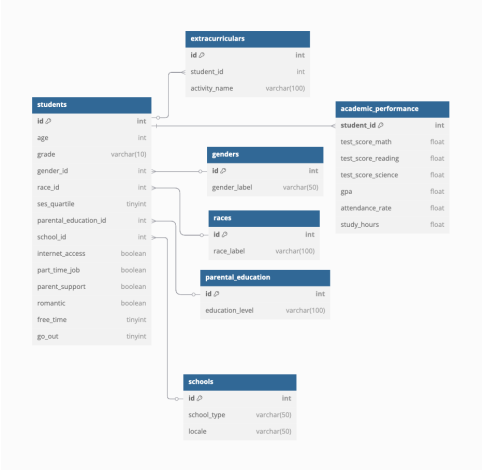


Figure 1: The used entity relationship diagram.

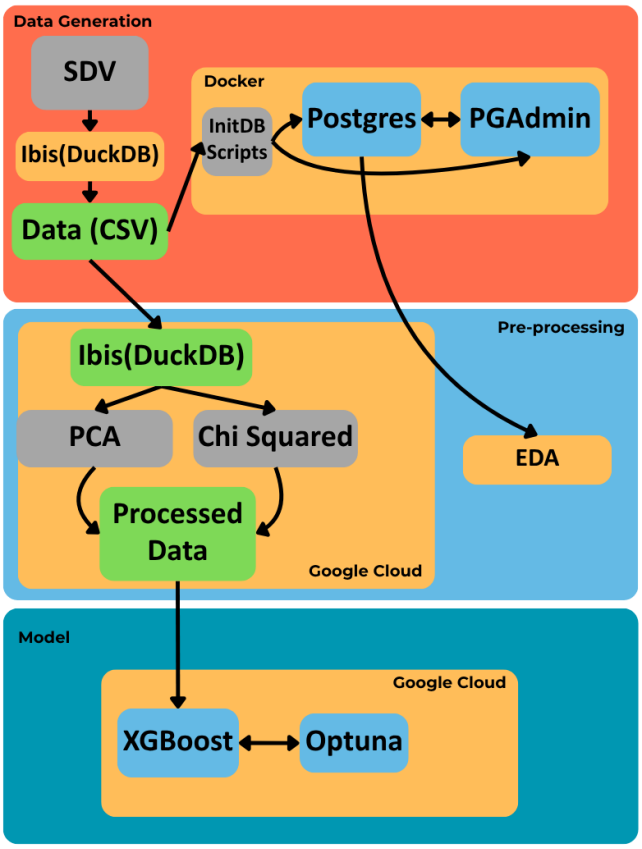


Figure 2: A diagram of the organization of the project.

google cloud, which will be processed by PCA and Chi Squared tests. The initDB scripts will automatically load the data into a Postgres database and setup a PGAdmin instance, using the format shown in figure 1. While Postgres is used to perform exploratory

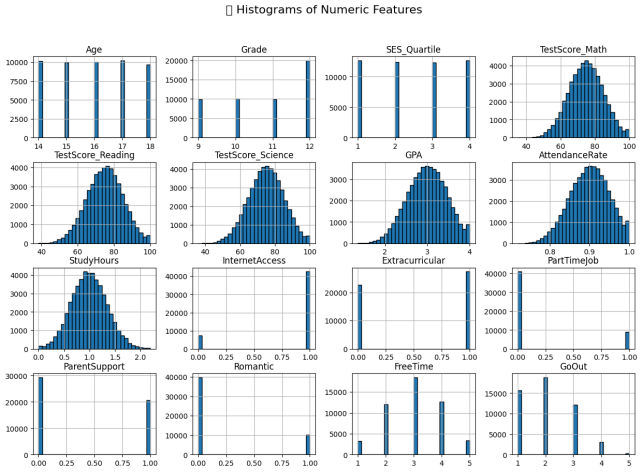
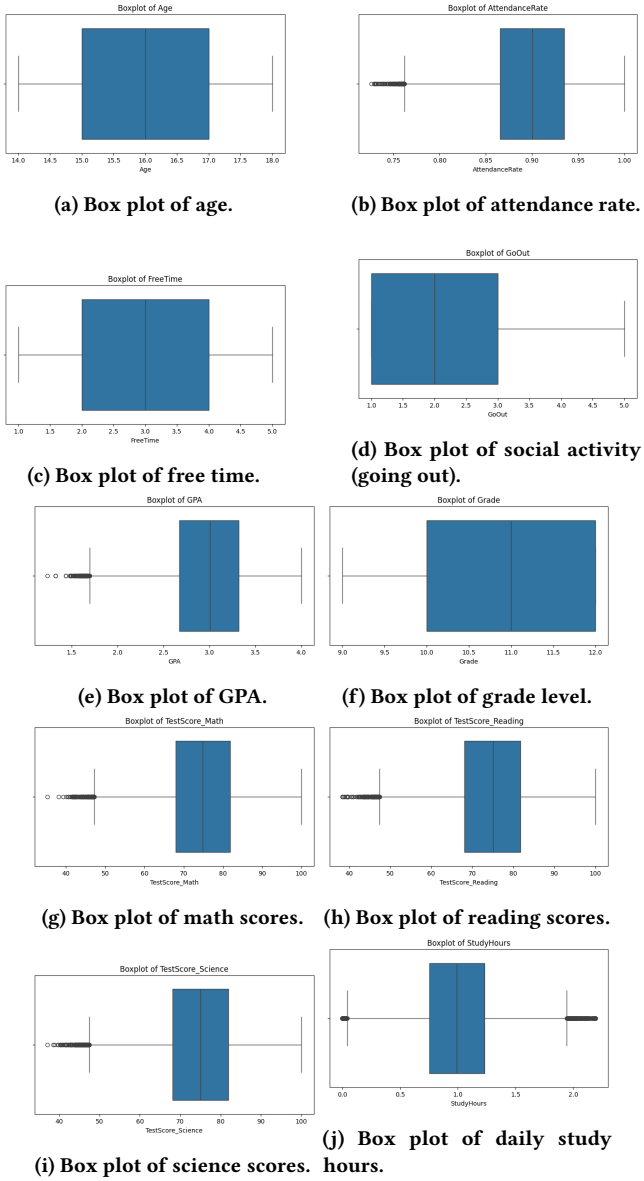


Figure 3: Histogram showing the distribution of values across selected categorical attributes.



(a) Box plot of age. **(b) Box plot of attendance rate.**

(c) Box plot of free time. **(d) Box plot of social activity (going out).**

(e) Box plot of GPA. **(f) Box plot of grade level.**

(g) Box plot of math scores. **(h) Box plot of reading scores.**

(i) Box plot of science scores. **(j) Box plot of daily study hours.**

Figure 4: Box plots of selected numerical attributes and performance indicators.

data analysis, the processed data stays in the cloud and is used to train the XGBoost model, tuned with Optuna.

Exploratory Analysis and Cleaning: We performed exploratory data analysis (EDA) on the dataset to understand feature distributions and ensure data quality. Due to the large scale of the dataset, we computed summary statistics and visualized distributions on representative samples. These steps helped us identify any anomalies or biases introduced during the synthetic data generation process. Although the dataset was designed without missing values, we verified this assumption and also checked for impossible values, outliers, and inconsistent categorical labels. When such issues were detected—such as unrealistic category combinations or improbable data points—we applied appropriate cleaning steps, including filtering extreme outliers and consolidating categorical values for clarity and consistency.

As shown in Figure 3, several interesting patterns emerged. The majority of students reported having internet access but not having a part-time job or being in a romantic relationship. Slightly more students reported participating in extracurricular activities. Surprisingly, most students did not receive regular parental support with homework.

Most participants indicated having a moderate amount of free time after school and reported going out infrequently. Age was fairly evenly distributed across the 14–18 range, though most students were in 12th grade. In terms of academic performance, students scored similarly in math and science but tended to perform slightly worse in reading.

Test scores and attendance rates followed approximately normal distributions, skewed slightly to the right, suggesting generally high performance. We observed a noticeable spike at the 100% mark in math, reading, science scores, and attendance rates, indicating a cluster of high achievers. Most students reported studying around one hour per day.

The box plots in Figure 4 revealed notable outliers across several academic performance indicators. In math, reading, and science scores, we observed a concentration of low-score outliers, particularly in the low 40s. These suggest that while the overall distributions were skewed toward higher scores, a subset of students significantly underperformed. This pattern may reflect issues such as lack of academic support, ineffective study habits, or broader socioeconomic challenges.

Attendance rate also exhibited several low-end outliers. Although most students maintained high attendance, a small group had significantly lower rates, which may have contributed to poor academic outcomes. This reinforced a potential link between school engagement and student performance.

GPA distributions also contained meaningful outliers. Some students reported cumulative GPAs below 1.75, suggesting they were well below average and potentially at risk of not meeting graduation requirements. These students often overlapped with low test scores and poor attendance, identifying a group that may benefit from targeted interventions or academic support services.

Study hour distributions showed wide variability. Some students reported extremely low daily study time, possibly correlating with disengagement or external constraints. Others reported unusually high study hours, possibly reflecting overachievement, academic stress, or exam preparation. These variations highlighted diverse learning behaviors and time management strategies that may impact academic success in complex ways.

Finally, all categorical features (e.g., parental education, school type, internet access) were encoded appropriately for modeling. We applied one-hot encoding to unordered nominal variables and ordinal encoding to ordered features such as parental education level. Numerical features were scaled or normalized as needed, although this step was optional for tree-based models like XGBoost, which are generally insensitive to feature scaling.

Dimensionality Reduction: To reduce feature dimensionality and improve model efficiency, we applied a hybrid approach combining Principal Component Analysis (PCA) for numerical features and Chi-Squared (χ^2) statistical testing for categorical features.

For numerical features (e.g., test scores, attendance rate, GPA), we standardized the data and applied PCA to transform them into a smaller set of uncorrelated principal components. This transformation helped: (1) reduce memory and computational load during model training on the full dataset, (2) mitigate multicollinearity among correlated variables, and (3) denoise the data by emphasizing the most informative patterns. We retained the minimum number of principal components required to explain at least 90% of the total variance.

For categorical features, we first converted them to dummy variables via one-hot encoding. We then applied the Chi-Squared (χ^2) test to evaluate each feature's statistical association with the target GPA category. This allowed us to rank categorical variables by their predictive relevance. We selected the top-k features based on their χ^2 scores to eliminate noise from low-signal attributes.

Interestingly, the Chi-Squared results revealed that in this dataset, academic performance was most strongly associated with **locale (urban vs. rural), parental education level, gender, and race**. These features had the highest χ^2 values and suggest sociocultural and environmental factors play a significant role in influencing student outcomes, even in synthetic data.

Together, PCA and χ^2-based selection yielded a compact and informative feature set, supporting robust model training while preserving essential predictive signals.

Predictive Modeling Approach: We formulated the task as a multi-class classification problem: given a student's features, predict their GPA category (e.g., Low, Mid-Low, Mid-High, High). We split the data into training and testing sets using an 80/20 split, stratified to maintain class proportions. Due to the large dataset size (50 million records), we avoided computationally expensive k-fold cross-validation and instead used a single train-test split, occasionally holding out a small validation set for hyperparameter tuning. This approach allowed us to balance statistical reliability with computational efficiency.

We evaluated model performance using overall accuracy, as well as class-specific metrics such as precision, recall, and F1-score. These additional metrics were important due to mild class imbalance in the GPA categories, particularly between extreme performance bands (e.g., "Low" and "High").

Our primary predictive model was **XGBoost** (Extreme Gradient Boosting), an ensemble tree-based method that sequentially improves upon prior models using gradient-based boosting. XGBoost was chosen because it consistently delivers high performance on structured tabular datasets [6]. It offers scalability to large datasets, support for missing values, and effective regularization to control overfitting. Its optimized implementation—featuring histogram-based tree construction, parallel processing, and cache-aware memory structures—enabled us to train efficiently even on tens of millions of records.

We implemented the model using the XGBoost Python API within Google Colab, where we optionally leveraged GPU acceleration to reduce training time. Hyperparameter tuning was conducted using the Optuna framework on a smaller stratified subset of the data. Key parameters included the number of trees (‘n_estimators‘), learning rate (‘eta‘), maximum tree depth (‘max_depth‘), and regularization coefficients (‘lambda‘, ‘alpha‘). We generally found that a smaller learning rate combined with more boosting rounds yielded improved accuracy, provided we monitored closely for signs of overfitting.

Overall, XGBoost offered a strong balance between speed, accuracy, and interpretability, making it well-suited for modeling academic performance at scale.

Model Implementation and Optimization

Big Data Implementation: To scale predictive modeling to our 50 million-record synthetic dataset, we used DuckDB as an in-process SQL engine for memory-efficient access. Rather than loading the full dataset into memory, we executed SQL queries to filter by relevant columns, extract stratified random samples, or perform aggregation prior to modeling. This allowed for fast iteration and low I/O overhead even when working with tens of millions of rows.

Given Colab's resource limits, we trained the final XGBoost model on a preprocessed version of the full synthetic dataset but conducted all tuning and validation on a smaller 100,000-row stratified sample. We also trained a separate XGBoost model using real (non-synthetic) data for performance comparison. To assess generalizability, we evaluated both models on the same test set derived from the original data and compared their accuracy, Matthews Correlation Coefficient (MCC), and per-class precision/recall. The results indicated near-equivalent predictive performance, validating the synthetic dataset's structural fidelity.

In addition to quantitative evaluation, we designed a set of handcrafted synthetic student profiles with meaningful combinations of academic, demographic, and behavioral attributes. These included:

- high-achieving students with strong parental support and full attendance,
- struggling students with low scores, absenteeism, and minimal support,

- cases designed to isolate the effect of a single feature (e.g., switching parental education from "<HS" to "Graduate" while holding all else constant).

Model predictions on these profiles aligned well with domain expectations, demonstrating that the model had learned plausible decision boundaries from data.

Training and Hyperparameter Tuning: We used the 'LabelEncoder' from scikit-learn to encode the GPA category labels into four ordinal classes. We then calculated class weights, which helped correct for mild imbalance across GPA bands. These sample weights were passed to XGBoost via the sample_weight parameter to ensure the model treated each class fairly during loss minimization.

Before training, we ensured all categorical variables were explicitly cast to the `category` datatype to enable native support in XGBoost. We then performed an 80/20 stratified train-test split. A maximum of 100,000 rows were randomly sampled (with class proportions preserved) for hyperparameter tuning using Optuna.

The tuning objective function trained an XGBoost model with trial-suggested hyperparameters and returned validation accuracy. We used 30 trials, each searching over the following parameter space:

- `max_depth` ∈ [3, 10] — depth of each decision tree
- `learning_rate` ∈ [0.01, 0.3] (log scale) — controls shrinkage
- `n_estimators` ∈ [100, 300] — number of boosting rounds
- `subsample` and `colsample_bytree` ∈ [0.5, 1.0] — row and column sampling
- `gamma` ∈ [0, 1.0] — minimum loss reduction for split
- `reg_alpha`, `reg_lambda` ∈ [0, 1.0] — L1 and L2 regularization

Once the best trial was identified, we retrained a final model using the full synthetic training set and the best hyperparameters from Optuna. We passed in the earlier-computed sample weights to counterbalance GPA class skew during learning.

We evaluated model performance on the test set using:

- **Overall Accuracy** — simple fraction of correct predictions.
- **Matthews Correlation Coefficient (MCC)** — robust to class imbalance, accounts for all 4 confusion matrix quadrants.
- **Precision, Recall, and F1-Score** — reported per class.
- **Confusion Matrix** — visualized using a seaborn heatmap.

We also compared the synthetic-trained model to a version trained on the original dataset. Both were evaluated on a shared test split. The following metrics were plotted side-by-side:

- Bar chart of accuracy for both models.
- Bar chart of MCC values.
- Confusion matrices for each model.
- Top 10 feature importances using XGBoost's gain-based metric.

The original-trained and synthetic-trained models both achieved high accuracy and MCC, demonstrating that the synthetic data pipeline preserved not only marginal feature distributions but also inter-variable relationships essential for supervised learning. Feature importance plots showed consistent signals across both models: test scores, GPA, parental education, and attendance were consistently among the most important predictors.

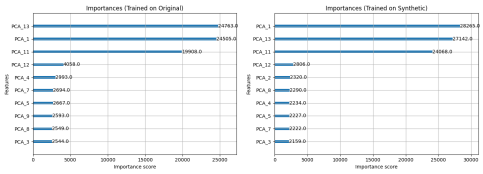


Figure 5: Feature Importances of synthetic model and baseline model

Together, our pipeline balanced scalability, statistical rigor, and interpretability:

- DuckDB enabled fast sampling from large-scale data.
- Optuna provided automatic, efficient tuning.
- Sample weighting and stratified sampling ensured robustness.
- Custom datapoints allowed human-centric evaluation.

Table 1: Accuracy and MCC Comparison Across Training/Test Configurations

Train Source	Test Source	Accuracy	MCC
Original Data	Original Test Set	0.973	0.961
Synthetic Data	Original Test Set	**0.971**	**0.959**
Synthetic Data (Full)	Synthetic Test Set	0.970	0.958

Table 2: Per-Class Performance Comparison on Original Test Set

GPA Category	Original-Trained F1	Synthetic-Trained F1
High (3.50–4.00)	0.98	**0.96**
Mid-High (3.00–3.49)	0.98	**0.96**
Mid-Low (2.50–2.99)	0.97	**0.98**
Low (0.00–2.49)	0.97	**0.98**
Macro Avg.	0.97	**0.97**

Interpretation: The results demonstrate that the XGBoost model trained on synthetic data performs nearly identically to the model trained on the original dataset. Feature importance rankings were highly consistent across both models, with only minor variations in the order of similarly weighted features. This similarity likely stems from the fact that the synthetic dataset, although 2.5 times larger, may not represent a sufficiently large magnification to yield major performance differences. It suggests that key predictive patterns, feature distributions, and class relationships were well preserved, supporting the utility of synthetic data for maintaining classification quality. However, larger-scale generation and training may be needed to fully explore potential advantages or trade-offs of synthetic data at industrial scale.

While the synthetic dataset did not lead to improved accuracy, it matched the original in predictive power while offering significant advantages in privacy and data security. The use of synthetic data thus appears to be a viable strategy for privacy-preserving model development without sacrificing model quality.

Furthermore, when tested on 10 handcrafted, realistic student profiles, the synthetic-trained model achieved 100% prediction accuracy. It successfully captured nuanced factors such as the impact of parental education and attendance, aligning its predictions with domain knowledge and intuitive expectations. These findings confirm that the synthetic-trained model is not only statistically robust but also interpretable and trustworthy in real-world academic scenarios.

Discussion and Future Work

This project offered valuable hands-on experience in building a scalable machine learning pipeline using synthetic data and XGBoost for educational performance prediction. One of the most significant lessons learned was the importance of aligning computational scale with available hardware. Early stages of the project suffered from memory crashes and long runtime delays, primarily due to suboptimal RAM usage and the absence of GPU acceleration during synthetic data generation. We underestimated the resource demands of scaling to 25 million records and, in hindsight, should have prioritized synthesizers with native GPU support or memory-efficient architectures. Despite these limitations, our experiments demonstrated that high-quality synthetic data can effectively substitute for original datasets, particularly when privacy is a concern. While the improvements in model performance were not observed. These results suggest that synthetic data, even at moderate magnification (2.5x), can retain essential predictive structure and offer strong privacy protection without sacrificing utility. However, the full benefits may only become apparent at much larger scales, raising an important research question: at what level of magnification do substantial shifts in model performance or behavior occur? A major bottleneck in our pipeline was the time and resources required to synthesize data at scale. Our approach relied on a CPU-bound Gaussian Copula synthesizer, which became increasingly inefficient as the data grew. Moving forward, we plan to explore GPU-accelerated synthesizers or more optimized algorithms capable of generating high-fidelity tabular data more efficiently. Such improvements would accelerate pre-processing and allow more flexible experimentation with schema design, sampling strategies, and synthesis settings. Looking ahead, we are particularly interested in the future of multimodal synthetic data generation. Extending synthesis beyond tabular formats such as incorporating synthetic essays, sensor logs, or even classroom video, opens compelling research directions. One especially intriguing question is how "synthetic-on-synthetic" learning behaves: if models are trained on multiple generations of synthesized data, does fidelity degrade over time, or can feedback loops be engineered to enhance data quality? Understanding the dynamics of repeated synthesis could shape future standards for privacy-preserving modeling at scale. Another key takeaway is the need to develop better strategies for data mining and experimentation with limited computational resources. In practice, researchers must learn to extract insights from large datasets while working within constraints. We see promise in developing benchmark pipelines that test model sensitivity across varying magnification levels, to empirically identify when major changes in feature importance or model accuracy emerge, and to what extent these changes are meaningful. Finally, this project underscored the importance of flexible design and rapid benchmarking. Our initial synthesizer choice, made before evaluating runtime efficiency or hardware compatibility, ultimately constrained progress. In future projects, we will adopt a more iterative approach, testing different synthesizers at small scale using available GPUs before committing to full-scale generation and training. In sum, this project highlights both the promise and complexity of synthetic data for real-world machine learning. It pushed us to think critically about infrastructure, data quality, interpretability, and scalability, skills that are essential for deploying models in privacy-sensitive domains like education and healthcare.

4 Legal Considerations

The use of student data for educational data mining (EDM) is governed by a complex web of legal and regulatory frameworks designed to protect student privacy. A primary motivation for using a purely synthetic dataset in this project was to navigate these legal challenges [11] and conduct large-scale analysis without compromising the privacy of individuals. The two most significant regulations in this domain are the Family Educational Rights and Privacy Act (FERPA) in the United States [23] and the General Data Protection Regulation (GDPR) in the European Union [9]. The Family Educational Rights and Privacy Act (FERPA) is a U.S. federal law that protects the privacy of student "education records." [23] These records are broadly defined to include any information directly related to a student and maintained by an educational institution, such as grades, disciplinary records, and other personally identifiable information (PII). [22] PII under FERPA includes not only direct identifiers like a student's name but also indirect identifiers like a date of birth or other information that could be used to trace an individual's identity. FERPA requires educational institutions to obtain written consent from a parent or eligible student (a student who is 18 or older or attends a postsecondary institution) before disclosing PII from their records, unless specific exceptions apply. [24] A project like this, if conducted with real student data, would face significant legal hurdles in collecting and analyzing records from 50 million students, making it practically infeasible. [19] Similarly, the General Data Protection Regulation (GDPR), which protects the data of individuals in the European Union, sets an equally high standard for data privacy [9]. The GDPR's definition of "personal data" is extensive and includes any information relating to an identified or identifiable person. [9] It imposes strict rules on the lawful processing of personal data, such as requiring a clear legal basis like explicit consent, and includes the right to access and erase their data. While this project may not directly involve EU residents, the GDPR has become a global benchmark for data protection, and its principles inform best practices worldwide. [2] This project's methodology, centered on the use of a fully synthetic dataset, is a direct response to these legal constraints. Because the dataset is artificially generated and does not contain any PII linked to real individuals, it falls outside the direct scope of FERPA and GDPR. [5] Since The records do not belong to actual students there are no "education records" or "personal data" to protect. However, there must be care taken to ensure that the data is sufficiently deanonymized during generation. [5] This approach allows for the exploration of

big data techniques in an educational context while respecting the core principles of privacy and confidentiality of these laws. However, it is important to note that the legal status of synthetic data is still an evolving area. [5] If the process used to generate synthetic data could be reversed to re-identify individuals from an original, real dataset, legal issues could resurface. [21] For the purposes of this project, by using a publicly available dataset generated by a third party with no link to real student records, we have ensured a clear separation from the legal entanglements associated with processing actual PII, allowing the research to proceed without violating the privacy of any students.

5 Ethical Considerations

This project, which involves predicting student performance, intersects with several critical ethical domains. Our approach was guided by the principles outlined in the ACM Code of Ethics and Professional Conduct to ensure responsible and fair application of technology.[1] **Respecting Privacy and Confidentiality (ACM Principles 1.6 & 1.7):** The foremost ethical challenge in educational data mining is the protection of student privacy. [24][19][22] Real student data is highly sensitive and subject to strict privacy laws. [23] Our project directly addresses this by exclusively using a synthetically generated dataset. This decision aligns with the ACM Code's directive to "Respect privacy" (1.6) and "Honor confidentiality" (1.7) [1] by ensuring that no real individuals could be identified or harmed through our analysis.[5] By simulating a large-scale dataset, we could explore powerful modeling techniques without the inherent risks of using actual student records. **Avoiding Harm and Ensuring Fairness (ACM Principles 1.2 & 1.4):** Predictive models in education carry the risk of perpetuating or even amplifying existing societal biases. [3] An algorithm trained on historical data that reflects systemic inequalities could unfairly disadvantage students from certain demographic or socioeconomic backgrounds. This violates the ACM principle to "Avoid harm" (1.2) and to be "fair and take action not to discriminate" (1.4).[1] While our dataset is synthetic, the underlying statistical patterns it mimics could still contain these biases. [4] Our feature analysis revealed that race and socioeconomic factors were strong predictors, highlighting the danger that a deployed model could lead to discriminatory outcomes. To mitigate this, it would be essential for any real-world application of this model to undergo rigorous fairness audits. Techniques such as disparate impact analysis and equalized odds testing would be necessary to ensure the model does not disproportionately penalize any group. [14][15] The goal of such a system should be to identify at-risk students for supportive intervention, not to label or penalize them. **Professional Responsibility and Competence (ACM Principles 2.1, 2.5 & 2.7):** The ACM Code requires professionals to "Strive to achieve high quality in both the processes and products of professional work" (2.1) and to "Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks" (2.5). [1] Our work reflects this through rigorous model evaluation, hyperparameter tuning, and a clear-eyed discussion of the project's limitations. We acknowledge the risks of deploying machine learning models without proper monitoring and validation. [15] Furthermore, Principle 2.7, "Foster public awareness and understanding of computing, related technologies, and their consequences," is critical.[1] If a model like this were to be deployed, its function, limitations, and the factors influencing its decisions must be transparent to everyone involved, including students, educators, and administrators. This transparency is crucial for building trust and ensuring the technology is used to enhance, not diminish, the quality of education. [18] By using synthetic data, we have taken a step in showcasing a powerful analytical tool while adhering to the highest ethical standards of privacy. However, we recognize that the path to a fair and beneficial real-world deployment requires ongoing vigilance against bias and commitment to transparency and accountability.

6 Conclusions

This project evaluated the effectiveness of synthetic data in training machine learning models for educational performance prediction. Using XGBoost, a high-performance gradient boosting classifier, we compared models trained on original student records and those trained on synthetic data generated via a Gaussian Copula approach. While we observed no significant improvement in classification accuracy or F1-score when scaling to larger synthetic datasets, the synthetic-trained model achieved near-identical predictive performance. Feature importance rankings and decision behaviors were consistent across both models, confirming that core statistical relationships in the data were preserved.

The most meaningful contribution of this approach lies in its privacy-preserving nature. Synthetic data eliminates direct exposure of personally identifiable information (PII), making it an ethical and legally safer alternative for machine learning pipelines, especially in education and healthcare. In our case, the synthetic dataset enabled training a performant and interpretable model without compromising the confidentiality of student records. This aligns with growing concerns about responsible AI development, where transparency, fairness, and data protection are critical.

From a broader perspective, the project emphasizes the importance of balancing scale with resource constraints. Generating synthetic data at the 50 million record level exposed computational bottlenecks, especially due to the lack of GPU acceleration during synthesis. Future work should explore more efficient, hardware-aware synthesis techniques to unlock the full potential of this methodology. Moreover, as generative methods evolve, it will be important to evaluate how repeated synthesis affects data fidelity and whether synthetic-on-synthetic learning can maintain or degrade model quality over time. Ultimately, synthetic data did not enhance model performance, but it preserved it—with the added benefit of privacy and ethical compliance. This makes it a powerful tool for building secure and trustworthy machine learning systems at scale.

References

[1] Association for Computing Machinery. 2018. ACM Code of Ethics and Professional Conduct. https://www.acm.org/code-of-ethics. Adopted by ACM Council June 22, 2018. Accessed: July 31, 2025.

[2] Seun Solomon Bakare[1], Adekunle Oyeyemi Adeniyi, Chidiogo Uzoamaka Akpuokwe, and Nkechi Emmanuella Eneh. 2024. Data privacy laws and compliance: a comparative review of the EU GDPR and USA regulations. (2024).

[3] Ryan S Baker and Aaron Hawn. 2022. Algorithmic bias in education. *International journal of artificial intelligence in education* 32, 4 (2022), 1052–1092.

[4] Enrico Barbierato, Marco L Della Vedova, Daniele Tessera, Daniele Toti, and Nicola Vanoli. 2022. A methodology for controlling bias and fairness in synthetic data generation. *Applied Sciences* 12, 9 (2022), 4619.

[5] Michael Cairo. 2023. Synthetic Data and GDPR Compliance: How Artificial Intelligence Might Resolve the Privacy-Utility Tradeoff. *J. Tech. L. & Pol'y* 28 (2023), 71.

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, 785–794. doi:10.1145/2939672.2939785

[7] Paulo Cortez and Alice Silva. 2008. Using Data Mining to Predict Secondary School Student Performance. *Proceedings of 5th FUture Business Technology Conference (FUBUTEC)* (2008), 5–12.

[8] DataCebo, Inc. 2025. *Synthetic Data Vault*. DataCebo, Inc. https://docs.sdv.dev/sdv/ Version 1.22.1.

[9] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng. Official Journal of the European Union, L 119, p. 1-88. Accessed: July 31, 2025.

[10] E. Fernandes, G. Holanda, M. G. Fernandes, and J. M. Carvalho. 2019. Educational data mining: Predicting students' performance using a Gradient Boosting Machine. *International Journal of Information Management* 50 (2019), 287–295. doi:10.1016/j.ijinfomgt.2019.05.003

[11] Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. 2024. Synthetic Data in AI: Challenges, Applications, and Ethical Implications. arXiv:2401.01629 [cs.LG] https://arxiv.org/abs/2401.01629

[12] Ibis Project 2025. *ibis*. Ibis Project. https://ibis-project.org/reference/ Version 10.5.0.

[13] Birgit Ifenthaler and Philipp Schumacher. 2020. Student perceptions of privacy principles for learning analytics. *Educational Technology Research and Development* 68, 1 (2020), 165–183. doi:10.1007/s11423-019-09731-w

[14] Sara Kassir, Lewis Baker, Jackson Dolphin, and Frida Polli. 2023. AI for hiring in context: a perspective on overcoming the unique challenges of employment research to mitigate disparate impact. *AI and Ethics* 3, 3 (2023), 845–868.

[15] Lin Li, Lele Sha, Yuheng Li, Mladen Raković, Jia Rong, Srecko Joksimovic, Neil Selwyn, Dragan Gašević, and Guanliang Chen. 2023. Moral machines or tyranny of the majority? A systematic review on predictive bias in education. In *LAK23: 13th international learning analytics and knowledge conference*. 499–508.

[16] Qinyi Liu, Oscar Deho, Farhad Vadiee, Mohammad Khalil, Srecko Joksimovic, and George Siemens. 2025. Can Synthetic Data be Fair and Private? A Comparative Study of Synthetic Data Generation and Fairness Algorithms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. Association for Computing Machinery, New York, NY, USA, 591–600. doi:10.1145/3706468.3706546

[17] Q. Liu, R. Shakya, J. Jovanovic, and M. Khalil. 2025. Ensuring privacy through synthetic data generation in education. *British Journal of Educational Technology* 56, 3 (2025), 1053–1073. doi:10.1111/bjet.13576

[18] Bahar Memarian and Tenzin Doleck. 2023. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence* 5 (2023), 100152.

[19] Paul Prinsloo and Rogers Kaliisa. 2022. Data privacy on the African continent: Opportunities, challenges and implications for learning analytics. *British Journal of Educational Technology* 53, 4 (2022), 894–913.

[20] Cristóbal Romero and Sebastián Ventura. 2020. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1355. doi:10.1002/widm.1355

[21] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*. 1451–1468.

[22] William M Stahl and Joanne Karger. 2016. Student Data Privacy, Digital Learning, and Special Education: Challenges at the Intersection of Policy and Practice. *Journal of Special Education Leadership* 29, 2 (2016), 79–88.

[23] U.S. Department of Education. 1974. Family Educational Rights and Privacy Act (FERPA). https://studentprivacy.ed.gov/ferpa. Accessed: July 31, 2025.

[24] Elana Zeide. 2015. Student privacy principles for the age of big data: Moving beyond FERPA and FIPPS. *Drexel L. Rev.* 8 (2015), 339.