# Generative AI for Synthetic Data Creation: Building Mastery-Focused Educational Datasets

Tapiwa Amion Chinodakufa, Dakota State University
Advisor: Dr. Khandaker Mamun Ahmed, Dakota State University

## Introduction & Problem Statement

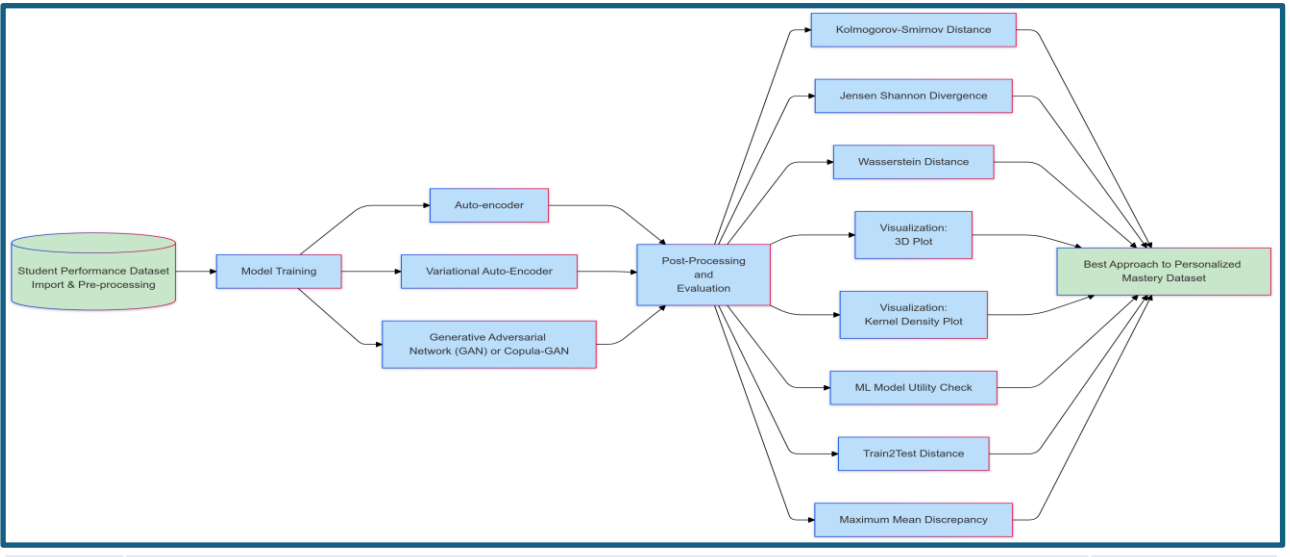
- There is no easily-available dataset for mastery-focused education, where mastery replaces grades while accurately reflecting student performance.
- Student data is restricted due to privacy & security concerns. One K-12 app was recently discovered selling unmasked data on millions of students
- Synthetic datasets may solve this by providing utility, privacy preservation, scalability, customization, variability, and resistance to reverse-engineering
- Techniques used included autoencoders, variational auto-encoders (VAE), generative adversarial networks (GAN) , and copulas combined with GANs

## Objectives

- Determine the ideal method for tabular synthetic data generation
- Create a large dataset from a small mastery-focused education dataset
- Evaluate synthetic data generation using two visualizations and six metrics

## Methodology

### Workflow Diagram




| Category | Metrics to Evaluate Synthetic Data Generation | Key Focus |
|---|---|---|
| Distribution Fidelity | **KS** Distance - measures max difference between the empirical distributions<br>**JS** Divergence – measures similarity between the probability distributions | Individual variability |
| Practical Utility | **ML** Model Utility – compares models trained on real vs synthetic<br>**T2D** ((Train-to-Test Distance) – compares models trained on synth & tested on real vs training and testing on real data | Model performance and applicability |
| Overall Divergence | **Wasserstein** Distance or Earth Mover's Distance - measures the "effort" required to morph one distribution into another, based on distance<br>**MMD** (Maximum Mean Discrepancy) – Difference between distributions in a mathematical 'sandbox' called a reproducing kernel Hilbert space | Holistic distribution differences |

## Results

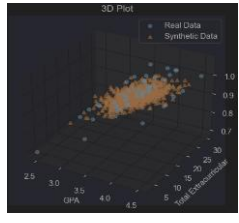
| Autoencoder | Variational Autoencoder | Copula-GAN |
|---|---|---|
| Kolmogorov-Smirnov Distance<br>KS Statistic: 0.08 | Kolmogorov-Smirnov Distance<br>KS Statistic: 0.08 | Kolmogorov-Smirnov Distance<br>KS Statistic: 0.07 |
| Jensen-Shannon Divergence 0.83 | Jensen-Shannon Divergence 0.23 | Jensen-Shannon Divergence 0.11 |
| Wasserstein Distance 1.48 | Wasserstein Distance 1.60 | Wasserstein Distance 2.24 |
| KDE Plot Visualization | KDE Plot Visualization | KDE Plot Visualization |
| ML Model Utility<br>Mean Squared Error on synthetic data: 3.27<br>R squared on real data: 0.98 | ML Model Utility<br>Mean Squared Error on synthetic data: 5.44<br>R squared on real data: 0.97 | ML Model Utility<br>Mean Squared Error on synthetic data: 56.00<br>R squared on real data: 0.59 |
| Train2Test Distance 7.23 | Train2Test Distance 3.21 | Train2Test Distance 0.14 |
| Mean 68, StDev 15 | Mean 67, StDev 13 | Mean 67, StDev 12 |
| MMD 0.06 | MMD 0.12 | MMD 0.20 |

VAE

| Mastery Education Dataset from VAE |
|---|
| Kolmogorov-Smirnov Distance<br>KS Statistic: 0.06 |
| Jensen-Shannon Divergence 0.12 |
| Wasserstein Distance 0.08 |
| KDE Plot Visualization |
| ML Model Utility<br>Mean Squared Error on synthetic data: 0.06<br>R squared on real data: 0.87 |
| Train2Test Distance 25 |
| - |
| MMD 0.18 |

## Discussion

- VAE outperformed, as it did not have any red flags. Autoencoder low MMD but also high T2D while Copula-GAN had high MSE
- VAEs are good at capturing the underlying distribution, handling mixed data, training stably, and avoiding mode collapse.
- Copula-GANS are close to VAEs on performance but ordinary GANs fail due to discriminator difficulty, mode collapse, and a high need for hyperparameter tuning. Conditional Tabular GAN (CTGAN) was not used but has potential.
- Autoencoders are good at replicating the data itself and sometimes do not generate a wider variety of data
- Maximum Mean Discrepancy, MMD, was lowest for the Autoencoder and highest for the Copula-GAN. Lower is better
- All models performed well on KS Distance, JS Divergence, Wasserstein Distance, and Mean-Standard Dev comparison
- For the Mastery Education Dataset, a 3D plot (right) shows that the real dataset had more outliers than the synthetic dataset, even though the original had 52 rows, and the synthetic had 1,000 rows
- Note: The mean for original Cleaned Student Performance dataset was 68 and standard deviation was 14




## Conclusion

- Based on multiple metrics, the synthetic data generated was a decent representative of the underlying data, while preserving privacy and reducing the risk of reverse engineering. It satisfies strict conditions for 'differential privacy'.
- The new Personalized Mastery-Focused Student-Centered Learning dataset is a new addition to this field of EdTech
- This dataset can now be used for:
  - **Data Augmentation:** Using synthetic data to augment small datasets, improving ML model performance.
  - **Education Research:** Facilitating research by providing a rich dataset without exposing sensitive information.
  - **Software Testing:** Testing educational software systems under various scenarios.
  - **Enhancing Analytics:** Providing a rich dataset for exploratory data analysis and predictive modeling.
  - **Promoting Innovation:** Facilitate the development of educational technologies and interventions, such as transitions to Personalized Mastery-Focused Student-Centered Learning.

## Formulas

*KS Distance*

$$D_{n,m} = \max_x |F_n(x) - F_m(x)|$$

*JS Divergence*

$$D_{JS}(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)$$

*Wasserstein Distance*

$$W(p,q) = \left(\inf_{\gamma \in \Gamma(p,q)} E_{(x,y)\sim\gamma}[d(x,y)]\right)$$

*Maximum Mean Discrepancy*

$$\mathrm{MMD}^2[F,X,Y] = \frac{1}{m^2}\sum_{i,j} k(x_i,x_j) + \frac{1}{n^2}\sum_{i,j} k(y_i,y_j) - \frac{2}{mn}\sum_{i,j} k(x_i,y_j)$$