VAE-Based Synthetic Data Generator

Overview

This project implements a Flask web application that allows users to generate synthetic tabular data using a Variational Autoencoder (VAE) model. The model is trained on datasets capturing student performance and behavioral habits. The synthetic data mirrors the structure and statistical characteristics of the original data, making it useful for privacy-conscious data sharing and machine learning model training.

System Workflow

1. **User Input:**
   - A web form allows users to specify the number of synthetic rows to generate.
2. **Data Preprocessing:**
   - A pre-trained scaler and encoder (vae_preprocessor.joblib) process numerical and categorical columns.
3. **Synthetic Data Generation:**
   - Random latent vectors are sampled and passed through the VAE decoder to generate synthetic data.
4. **Output:**
   - The generated data is reverse-transformed into the original feature space and offered as a downloadable CSV.

System Flowchart

```
flowchart TD
  A[User Submits Row Count via Web Form] --> B[Server Receives Request]
  B --> C[Preprocess Input Data using Pre-trained Scaler/Encoder]
  C --> D[Sample Latent Vectors from Normal Distribution]
  D --> E[Decode Latent Vectors using VAE Decoder]
  E --> F[Inverse Transform Synthetic Data to Original Scale]
  F --> G[Save as CSV and Serve to User for Download]
```

What is a Variational Autoencoder (VAE)?

A Variational Autoencoder (VAE) is a generative neural network that learns to model the underlying distribution of input data. By training on real-world data, it can generate new, realistic samples that mimic the original data's patterns.

Key Components of a VAE

Component	Description
Encoder	Maps input data to latent parameters: mean (μ) and log-variance (σ^2)
Reparameterize	Samples latent vector z using: $z = \mu + \sigma * \varepsilon$, where $\varepsilon \sim N(0,1)$
Decoder	Maps the latent vector z back to the data space, reconstructing input
Loss Function	Combines reconstruction error (e.g., MSE) and KL divergence for regularity

---

Mathematical Formula

$z = \mu + \sigma * \varepsilon \quad (\text{where } \varepsilon \sim N(0,1))$

$\hat{x} = \text{Decoder}(z)$

$\text{Loss} = \text{Reconstruction\_Loss} + \text{KL\_Divergence}$

Benefits of Using VAE for Synthetic Data Generation

- Learns complex and hidden data distributions
- Produces realistic yet anonymized samples
- Suitable for mixed data types with preprocessing
- Balances randomness and structure for better generalization

---

References

- Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes*. arXiv:1312.6114
- PyTorch VAE Tutorial
- GitHub: PyTorch VAE Examples
