**Report: Implementation of Feedforward Neural Network for Generating Synthetic Spatial Data**

**1. Introduction**

This project includes a feedforward neural network (FNN) to generate synthetic spatial data for origin and destination coordinates:

* **D\_ORIXCOOR**: X-coordinate of origin.
* **D\_ORIYCOOR**: Y-coordinate of origin.
* **D\_DESTXCOOR**: X-coordinate of destination.
* **D\_DESTYCOOR**: Y-coordinate of destination.

**2. Data Preprocessing**

The preprocessing pipeline ensures the dataset is properly scaled and ready for training:

1. **Feature Selection**:
   * Four numerical columns (D\_ORIXCOOR, D\_ORIYCOOR, D\_DESTXCOOR, D\_DESTYCOOR) are selected for analysis.
2. **Quantile Transformation**:
   * A QuantileTransformer adjusts the original data to a uniform distribution. This step reduces the effect of extreme values and ensures stable learning during training.
   * Transformed data is used for both features (X) and targets (y), ensuring a consistent mapping between input and output distributions.
3. **Resulting Data**:
   * Transformed data aligns with uniform distribution for enhanced training stability.

**3. Neural Network Architecture**

The FNN is designed to process transformed spatial data and produce synthetic outputs that match the original dataset's frequency distribution:

* **Input Layer**:
  + Accepts four features (D\_ORIXCOOR, D\_ORIYCOOR, D\_DESTXCOOR, D\_DESTYCOOR).
* **Hidden Layers**:
  + Two dense layers with 256 and 128 neurons, respectively, using ReLU activation.
  + Dropout layers (30%) added for regularization to reduce overfitting.
* **Output Layer**:
  + A linear layer with four neurons to match the dimensionality of the input.
* **Optimization**:
  + **Optimizer**: Adam for efficient gradient-based learning.
  + **Loss Function**: Mean Squared Error (MSE) to minimize reconstruction error.
  + **Metrics**: Mean Absolute Error (MAE) to measure average prediction deviation.

**4. Model Training**

The network was trained for 200 epochs with a batch size of 16, using 80% of the data for training and 20% for validation. Key training configurations:

* **Training Loss (MSE)**: Stabilized at a low value, indicating effective learning.
* **Validation Loss (MSE)**: Closely tracked the training loss, suggesting minimal overfitting.

The model demonstrated convergence during training, achieving consistent performance on validation data.

**5. Synthetic Data Generation**

After training, the network was used to generate synthetic spatial data:

1. **Input Generation**:
   * Randomly sampled uniform inputs mimic the transformed distribution of the original data.
2. **Prediction**:
   * The trained network produces outputs corresponding to the synthetic spatial coordinates.
3. **Inverse Transformation**:
   * The QuantileTransformer reverses the transformation applied during preprocessing, ensuring the synthetic data matches the original frequency distribution.

Synthetic data columns:

* **D\_ORIXCOOR**, **D\_ORIYCOOR**, **D\_DESTXCOOR**, **D\_DESTYCOOR**.

**6. Dynamic Filtering and Frequency Distribution Comparison**

To validate the quality of synthetic data:

1. **Dynamic Filtering**:
   * Original and synthetic data are filtered based on the 5th and 95th percentiles of the original data for each column.
   * This ensures a focus on the most relevant ranges while excluding extreme outliers.
2. **Frequency Distribution Comparison**:
   * Histograms of original and synthetic data are compared for each column.
   * Plots show close alignment between the original and synthetic distributions, demonstrating the network’s ability to reproduce realistic data patterns.

Example comparison for **D\_ORIXCOOR**:

* Original and synthetic distributions align well, confirming the preservation of frequency characteristics.

**7. Results and Observations**

* **Training**:
  + The network successfully learned the relationships in the spatial data, achieving low training and validation losses.
* **Synthetic Data**:
  + Generated data closely resembles the original dataset, as validated by distribution comparisons.
* **Dynamic Filtering**:
  + Ensures the analysis focuses on relevant data ranges, enhancing interpretability.

**8. Conclusion**

Key achievements include:

1. Training and validating a robust neural network with low loss values.
2. Generating synthetic data that matches the original dataset’s frequency distribution.
3. Dynamic filtering and frequency comparison validated the model’s effectiveness.

**Future Directions**:

1. Extend the model to include additional spatial and non-spatial features.
2. Experiment with alternative architectures, such as variational autoencoders (VAEs) for enhanced generation.
3. Optimize hyperparameters to improve performance further.