**Report: Implementation of Feedforward Neural Network for P-SEXE, P-GRAGE, and DHREDE**

**1. Introduction**

This project implements a feedforward neural network (FNN) to model relationships among demographic and trip-related variables. The dataset focuses on the following attributes:

* **P-SEXE**: Gender, encoded as binary or categorical data.
* **P-GRAGE**: Age group categories, representing individuals in predefined age brackets.
* **DHREDE**: Hours of trip, representing the duration of trips taken by individuals.

The goals of the project are to:

1. Train and optimize a neural network to predict the target variable (**P-GRAGE**).
2. Evaluate model performance using training, validation, and test sets.
3. Generate synthetic data for further analysis.

**2. Data Preprocessing**

The preprocessing steps ensure the dataset is ready for neural network modeling. The pipeline includes:

1. **Feature Selection**: Selecting three key columns (P-SEXE, P-GRAGE, DHREDE) from the dataset.
2. **Normalization**: Scaling continuous features to a range of 0 to 1 using MinMaxScaler. This step ensures consistent input magnitudes, which is critical for gradient-based optimization.
3. **Encoding**: Binary or categorical variables, such as P-SEXE, are encoded appropriately.
4. **Dataset Splitting**:
   * Features (X) include P-SEXE and DHREDE.
   * Target (y) is P-GRAGE.
   * The dataset is divided into training (80%) and testing (20%) subsets, with validation data derived from the training set.

After preprocessing, the data is organized as follows:

* **Training set**: Used to train the model.
* **Validation set**: Used to monitor performance during training.
* **Test set**: Used to evaluate final model performance.

**3. Neural Network Architecture**

The feedforward neural network is designed to predict P-GRAGE based on the input features (P-SEXE and DHREDE). The architecture includes:

* **Input Layer**: Accepts normalized features.
* **Hidden Layers**:
  + Configurable layers (64-128 neurons) with ReLU activation for non-linearity.
  + Dropout layers (30%) to prevent overfitting by randomly deactivating neurons during training.
* **Output Layer**:
  + A single neuron with linear activation, used for regression tasks.
* **Optimization**:
  + **Optimizer**: Adam, a robust gradient-based optimizer.
  + **Loss Function**: Mean Squared Error (MSE) to minimize prediction errors.
  + **Evaluation Metrics**: Mean Absolute Error (MAE) to measure the average prediction deviation.

**4. Model Training**

The model was initially trained for 50 epochs with a batch size of 32. Training and validation losses were logged to monitor performance over epochs.

Key Results:

* **Final Training Loss**: ~0.0754 (MSE)
* **Final Validation Loss**: ~0.0760 (MSE)

The results indicate effective convergence during training, with minimal overfitting.

**5. Model Retraining**

To further enhance performance, the model was retrained with the following modifications:

1. **Expanded Architecture**:
   * Additional hidden layers (128, 64, 32 neurons) to increase learning capacity.
   * Dropout layers retained for regularization.
2. **Extended Epochs**: The retraining process was run for 100 epochs to ensure the model had sufficient time to learn complex patterns.

Retraining Results:

* **Final Training Loss**: ~0.0743 (MSE)
* **Final Validation Loss**: ~0.0742 (MSE)

Retraining demonstrated improved convergence and lower losses, highlighting the benefit of the updated architecture.

**6. Testing Phase**

The retrained model was evaluated on the test set to assess its generalization capabilities. The results are as follows:

* **Test Loss**: 0.0732 (MSE)
* **Test MAE**: 0.2279

These values confirm that the model performed well on unseen data, with minimal deviation between predicted and actual values.

**7. Synthetic Data Generation**

The trained model was used to generate synthetic data for exploratory analysis. The process involved:

1. **Input Generation**: Randomly creating normalized values for features (P-SEXE, DHREDE).
2. **Prediction**: Using the model to predict P-GRAGE values for the generated inputs.

**Post-Processing**:

* Denormalized DHREDE values to reflect realistic trip hours.
* Converted P-SEXE and P-GRAGE to discrete values consistent with the original dataset.

**8. Conclusion**

This project successfully implemented a feedforward neural network for demographic and trip-related data. Key outcomes include:

* A robust model with low test loss (0.0732 MSE) and MAE (0.2279).
* Effective retraining, resulting in improved validation and test performance.
* Generation of realistic synthetic data for further exploration.

**Future Directions**:

1. Incorporate additional features to enhance predictive power.
2. Experiment with advanced architectures, such as ensemble models or recurrent networks.
3. Perform hyperparameter optimization to refine model performance.