**Machine Learning Model Evaluation Report**

1. **Preprocessing Steps and Rationale**

Data Cleaning:

* Dropped the hsi\_id column as it is irrelevant.
* Imputed missing values using the median of each feature to ensure data completeness.
* Removed outliers using the Interquartile Range (IQR) method to improve model robustness.

Data Splitting:

* Split data into training (80%) and testing (20%) sets using train\_test\_split with random\_state=42 for reproducibility.

Feature Scaling:

* Standardized features using StandardScaler to normalize feature distributions for better model performance.

Feature Selection:

* Used SelectKBest with mutual\_info\_regression to select the top 50 most relevant features, improving efficiency and reducing noise.

Dimensionality Reduction:

* Applied Principal Component Analysis (PCA) to retain 95% of the variance, leading to a reduced feature set for improved computational efficiency and potential overfitting prevention.

**2. Insights from Dimensionality Reduction**

* PCA reduced the feature space while retaining essential variance.
* Improved model interpretability and performance by removing redundant information.
* Helps speed up model training and avoids potential multicollinearity issues.

**3. Model Selection, Training, and Evaluation**

Models Used:

* Random Forest Regressor (RF) – Tuned using RandomizedSearchCV.
* Gradient Boosting Regressor (GB) – Tuned using RandomizedSearchCV.
* LightGBM Regressor (LGB) – Tuned using RandomizedSearchCV.
* Multi-layer Perceptron (MLP) – Deep learning model trained for 100 epochs.
* Convolutional Neural Network (CNN) – 1D CNN trained for 100 epochs.
* Graph Neural Network (GNN-like Dense model) – Dense model trained for 100 epochs.

Model Evaluation Metrics:

* R² Score – Measures how well the model explains variance in the target variable.
* Mean Absolute Error (MAE) – Measures absolute errors in predictions.
* Mean Squared Error (MSE) – Penalizes larger errors more heavily.

**4. Key Findings and Suggestions for Improvement**

**Findings:**

* **Random Forest and LightGBM** performed well in terms of R² Score and MAE.
* **MLP and CNN models** required substantial training time but showed promising performance.
* **Dimensionality reduction** improved model efficiency without significant loss of information.
* **Hyperparameter tuning** through RandomizedSearchCV helped optimize tree-based models.

**Suggestions for Improvement:**

* **Further Feature Engineering:** Investigate domain-specific transformations and interactions.
* **Time-Series Considerations:** If applicable, use LSTM or ARIMA models for temporal patterns.
* **Ensemble Learning:** Combine models for improved generalization.
* **Data Augmentation:** If more data is available, improve deep learning model performance.
* **Regularization Techniques:** Reduce overfitting in deep learning models by fine-tuning dropout layers.