# Spectral Data Analysis and Prediction Report

## 1. Preprocessing Steps and Rationale

### Data Cleaning:

- Missing Values: Checked for missing values and found none.
- Duplicate Records: No duplicate records were present.
- Feature Selection: Dropped the `hsi_id` column, as it was unnecessary for modeling.

### Feature Scaling:

- Standardization: Applied `StandardScaler` to normalize spectral reflectance values, ensuring all features have a mean of 0 and variance of 1.

## 2. Insights from Dimensionality Reduction

- Principal Component Analysis (PCA) was applied to reduce high-dimensional spectral features while preserving variance.
- PCA transformed data before model training, reducing redundancy among correlated features.
- Heatmap analysis revealed strong correlations among spectral bands, justifying dimensionality reduction.

## 3. Model Selection, Training, and Evaluation

## Models Trained:

I have trained 3 models: Random Forest Regressor, XGBoost Regressor and CNN model and have compared the results of all three models.

1. **Random Forest Regressor**
    - Trained and evaluated using standard regression metrics.
    - Performed hyperparameter tuning to optimize model.
2. **XGBoost Regressor**
    - Trained with grid search to find optimal hyperparameters.
    - Demonstrated competitive performance in feature-rich environments.
3. **Convolutional Neural Network (CNN)**
    - Reshaped data for CNN input.
    - Designed a CNN model and trained it for spectral regression.
    - Evaluated the improved CNN model performance.

### Performance Evaluation:

- Used RMSE, MAE, and R-squared to compare model accuracy.

```
Comparison of Random Forest, XGBoost, MLP, and Improved CNN:
Model                MAE             RMSE            R² Score
Random Forest        1782.6115       3709.7070       0.9508
Tuned XGBoost        1612.2167       3383.5868       0.9590
Improved CNN         1847.9323       4077.3849       0.9405
```

- XGBoost performed best in terms of prediction accuracy after tuning the model.

## 4. Key Findings and Suggestions for Improvement

- **Findings:**

**Mean Absolute Error (MAE)**

- XGBoost (1612.22) performs the best, followed by Random Forest (1782.61), and then the Improved CNN (1847.93).

**Root Mean Squared Error (RMSE)**

- XGBoost again performs the best (3383.59), followed by Random Forest (3709.71) and CNN (4077.38).

**$R^2$ Score**:

- XGBoost has the highest $R^2$ (0.9590), showing better explanatory power compared to Random Forest (0.9508) and CNN (0.9405).

- **Suggestions for Improvement:**
  - Further optimize CNN architecture for better spectral feature learning.
  - Explore additional feature engineering techniques.
  - Consider ensembling different models for improved robustness.