Short Report: Machine Learning Pipeline for Spectral Data Prediction

1. Preprocessing Steps & Rationale

- Data Cleaning: Removed missing and duplicate values to ensure consistency.
- Feature Scaling: Standardized all numerical features using StandardScaler.
- Dimensionality Reduction: Applied PCA to retain 95% variance, reducing computational complexity.
- Train-Test Split: Data was split into 80% training and 20% testing for model evaluation.

2. Model Selection, Training, & Evaluation

- Models Used:
  - RandomForestRegressor
  - XGBRegressor
- Grid Search was used to optimize hyperparameters for XGBoost.
- Evaluation Metrics:
  - MAE, RMSE, R^2 were used to compare model performance.

3. Key Findings & Improvements

- Both RandomForest and XGBoost provided reasonable accuracy.
- Future work: Experiment with more advanced models and hyperparameter tuning.