Forest Cover Type Classification Report

1. Introduction

This report documents the process and results of the Forest Cover Type Classification task. The objective is to predict forest cover type based on cartographic and environmental features using the Covertype dataset (UCI Machine Learning Repository).

2. Dataset Overview

Dataset: Covertype (UCI)
Number of samples: 581,012
Number of features: 54
Target variable: Cover_Type (multi-class, 7 categories)

Class distribution:
1 → 283,301
0 → 211,840
2 → 35,754
6 → 20,510
5 → 17,367
4 → 9,493
3 → 2,747

3. Methodology

The workflow for the classification task included the following steps:
1. Data loading and preprocessing
2. Splitting into training and testing sets (80/20)
3. Model training with Random Forest and XGBoost
4. Hyperparameter tuning using GridSearchCV
5. Model evaluation with classification report
6. Visualization of results (confusion matrix and feature importance)

4. Model Training & Evaluation

Random Forest Results

Best parameters: {'n_estimators': 100, 'min_samples_split': 2, 'max_depth': None}

Classification Report:
precision recall f1-score support

	precision	recall	f1-score	support
0	0.96	0.94	0.95	42368
1	0.95	0.97	0.96	56661
2	0.94	0.96	0.95	7151
3	0.92	0.86	0.89	549
4	0.95	0.77	0.85	1899
5	0.93	0.89	0.91	3473
6	0.97	0.95	0.96	4102
accuracy			0.95	116203
macro avg	0.95	0.91	0.92	116203
weighted avg	0.95	0.95	0.95	116203

XGBoost Results

Best parameters: {'n_estimators': 100, 'max_depth': 6, 'learning_rate': 0.1}

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.78	0.80	42368
1	0.81	0.86	0.84	56661
2	0.79	0.86	0.82	7151
3	0.83	0.83	0.83	549
4	0.88	0.31	0.46	1899
5	0.74	0.49	0.59	3473
6	0.90	0.80	0.85	4102
accuracy			0.81	116203
macro avg	0.82	0.71	0.74	116203
weighted avg	0.81	0.81	0.81	116203

5. Analysis

Random Forest achieved superior performance with ~95% accuracy compared to XGBoost (~81%). It handled class imbalance better, producing higher precision and recall across minority classes.

XGBoost underperformed, particularly on minority classes (e.g., class 4 and 5), due to class imbalance sensitivity.

6. Conclusion

The experiment demonstrates that Random Forest is more effective than XGBoost for the Covertype dataset. Further improvements to XGBoost could be made by applying class weights, resampling techniques, or hyperparameter tuning. For practical applications where accuracy is paramount, Random Forest should be preferred.