MACHINE LEARNING APPROACH FOR LANDCOVER CLASSIFICATION

Introduction

This project focuses on developing a robust predictive model to predict land cover types based on geospatial data. The model classifies land cover into three key categories:

- Buildings
- Cropland
- Woody Vegetation Cover (>60%)

This document details the methodology used in data preprocessing, feature engineering, model training, evaluation, and final prediction.

Datasets

- Training Dataset
- Test Dataset
- Sample Submission to provide a reference format for final output.

Before training, datasets were explored by:

- Checking the structure and column names using print(training_df.columns).
- Identifying missing values using training_df.isnull().sum().
- Analyzing class distributions to check for imbalances in land cover types.

Data Preprocessing

1. Handling missing data - Missing values were filled with the median of the respective column. Non-informative columns with excessive missing values were also dropped.
2. Feature Selection - Irrelevant columns (subid, building, cropland, wcover) were removed from the training set. This ensured that only the independent features are used for training, while target variables (labels) are stored separately.
3. Feature Scaling – Standardization was done using StandardScaler() to transform the feature to uniform scale centering data between 0 and 1.
4. Label Encoding - to convert categorical labels to numerical format to be read by model.

Splitting of Training and Validation Data

Data was split into 80% training set and 20% validation set so as to evaluate model performance before making predictions on the test dataset. This prevents overfitting, ensures fair evaluation, maintains reproducibility – random_state=42 ensures the split is consistent every time.

Model Training

- Random Forest Classifier was chosen because it’s a powerful ensemble learning method that performs well for tabular data.
- Since we had three target labels (Buildings, Cropland, Woody Cover), a separate RandomForestClassifier was trained for each category. n_estimators=100 ensures multiple decision trees are trained for robustness.

Model Evaluation

- To assess performance, Validation Set was used to make predictions.
- For each land cover type, classification report was generated as follows.

Findings:

Building Classification

- *Class 0 (Not a building)*: Perfect precision (1.00) and recall (1.00), meaning the model correctly identifies all non-building areas.
- *Class 1 (Building)*: Also has perfect precision (1.00), but recall is slightly lower (0.99), indicating a very minor number of buildings were misclassified as non-buildings.
- *Overall Accuracy*: 100% accuracy, indicating the model performs exceptionally well for this class.

Cropland Classification

- *Class 0 (Not cropland):* High precision (0.81) and recall (0.90), meaning the model is very good at identifying non-cropland areas.
- Class 1 (Cropland): Precision is slightly lower (0.74), and recall is 0.57, indicating that some cropland areas were misclassified as non-cropland.
- *Overall Accuracy*: 79%, meaning there is room for improvement, especially in identifying cropland areas.

Woody Vegetation Cover Classification

- *Class 0 (No woody cover):* Precision is 0.58 and recall is 0.62, showing moderate performance.
- *Class 1 (Moderate woody cover):* Precision is 0.41, with a recall of 0.19, indicating that many moderate woody cover areas are being misclassified.
- *Class 2 (High woody cover):* Precision is 0.63, and recall is 0.82, showing that high woody cover areas are well identified.
- *Overall Accuracy:* 59%, meaning further improvements are needed.

Recommendation: Use additional features (e.g., spectral indices, elevation data) to improve cropland and wcover classification.