Technical Report

1. Introduction

This report outlines the development of a **Land Cover Classification Model**, created as part of the Amini Data Science Internship technical assignment. The goal was to predict different land cover types (**building, cropland, and wcover categories**) using geospatial and environmental data. A **Random Forest model** was trained to classify land cover for new locations based on historical data.

2. Approach & Methodology

I followed a structured data science workflow to ensure accurate predictions.

2.1 Data Preprocessing

- The **training dataset** (train_land_cover_assignment.csv) and **test dataset** (test_land_cover_assignment.csv) were loaded into Pandas.
- Missing values in numerical columns were **replaced with column averages**.
- Categorical data was handled as follows:
  - building and cropland labels were converted from Yes/No to **1/0**.
  - wcover values (<30%, >30%, >60%) were **one-hot encoded** into separate features.

2.2 Feature Engineering

- We used various geospatial and environmental features like **latitude, longitude, soil composition, and climate indicators**.
- The target labels were **separated** from the input features.
- The dataset was split into **80% training and 20% validation** for performance testing.

2.3 Model Selection & Training

- A **Random Forest Classifier** was chosen for its reliability with structured data and ability to capture complex relationships.
- Separate models were trained for each target variable using **200 trees (n_estimators=200)** and **balanced class weights**.
- Hyperparameters were fine-tuned using **grid search** to optimize performance.

2.4 Evaluating Model Performance

- We assessed the model using:
  - **Accuracy Score** to measure correctness.
  - **Precision, Recall, and F1-score** to analyze class-level performance.
- Final validation results:
  - **Building:** 99.94% accuracy
  - **Cropland:** 78.53% accuracy
  - **Wcover Categories:** 100% accuracy

2.5 Generating Predictions

- Instead of just classifying 0 or 1, we used **probability predictions (predict_proba)**.

The predictions were saved in submission.csv, formatted as:
subid, building, cropland, wcover_<30%, wcover_>30%, wcover_>60%
1548905, 0.01, 0.78, 0.45, 0.32, 0.92

- 1548829, 0.03, 0.65, 0.58, 0.21, 0.87

3. Key Findings

- **High Accuracy for Buildings & Wcover Categories:** The model performed exceptionally well, predicting buildings and wcover types with **near-perfect accuracy**.
- **Cropland Classification Needs Improvement:** Accuracy was lower (~78.53%), likely due to data imbalance or similarities between cropland and other land types.
- **Class Imbalance Impact:** The dataset contained significantly **more 0s than 1s**, leading the model to predict more negatives.
- **Feature Importance Analysis:** Key predictors included **latitude, soil quality, and climate indicators**.

4. Recommendations

- **Improve Cropland Predictions:** Use **SMOTE (Synthetic Minority Over-sampling Technique)** or **stratified sampling** to balance cropland data.
- **Try Alternative Models:** Test **XGBoost or LightGBM**, which may handle complex patterns better.
- **Enhance Feature Engineering:** Incorporate **seasonal data** and **satellite imagery**.
- **Optimize Hyperparameters:** Use **Bayesian Optimization** or **Random Search** for better tuning.

5. Conclusion

This project successfully built a **machine learning-based land cover classification model** with **high accuracy** in most categories. While building and wcover classifications performed well, cropland classification can be improved. Future work should focus on **handling class imbalance**, **exploring advanced models**, and **enhancing feature selection** to further refine predictions.