Land Cover Classification Technical Report

This project aimed to develop a predictive model for land cover classification in a region experiencing rapid land-use changes. The primary objective was to classify each observation into one of three land cover categories: Buildings, Cropland, and Woody Vegetation Cover .

Methodology and Approach

Data Understanding

The dataset contained 15,811 training observations and a corresponding test dataset. Key columns included geographic coordinates (lat, lon), environmental features (e.g., bio1, bio12, bio7, bio15), remote-sensing variables (e.g., mb1, mb2, mb3, mb7), soil properties (bd20, cec20, ph20, soc20), and additional features without documentation (e.g., bcount, dnlt, nppm, sirs).
The original target variables were provided as three separate indicators: building, cropland, and wcover. These were merged into a unified target variable (landcover) because my domain assumption is that each location should have one predominant land cover type. Treating it as a single-label (mutually exclusive) problem. They were unified based on the following rules:

- If building is "Yes": assign Buildings.
- Else if cropland is "Yes": assign Cropland.
- Else if wcover meets the specified threshold (e.g., ">60%" for dense woody cover or ">30%" for transitional woody): assign Woody Vegetation Cover (with ambiguous cases reclassified accordingly).
- The cases that did not meet any of the conditions, were originally labeled as “Other”, were refined based on NDVI and plantation data, and any remaining “Other” cases were dropped.

Data Preprocessing and Exploratory Data Analysis

Missing Data Handling

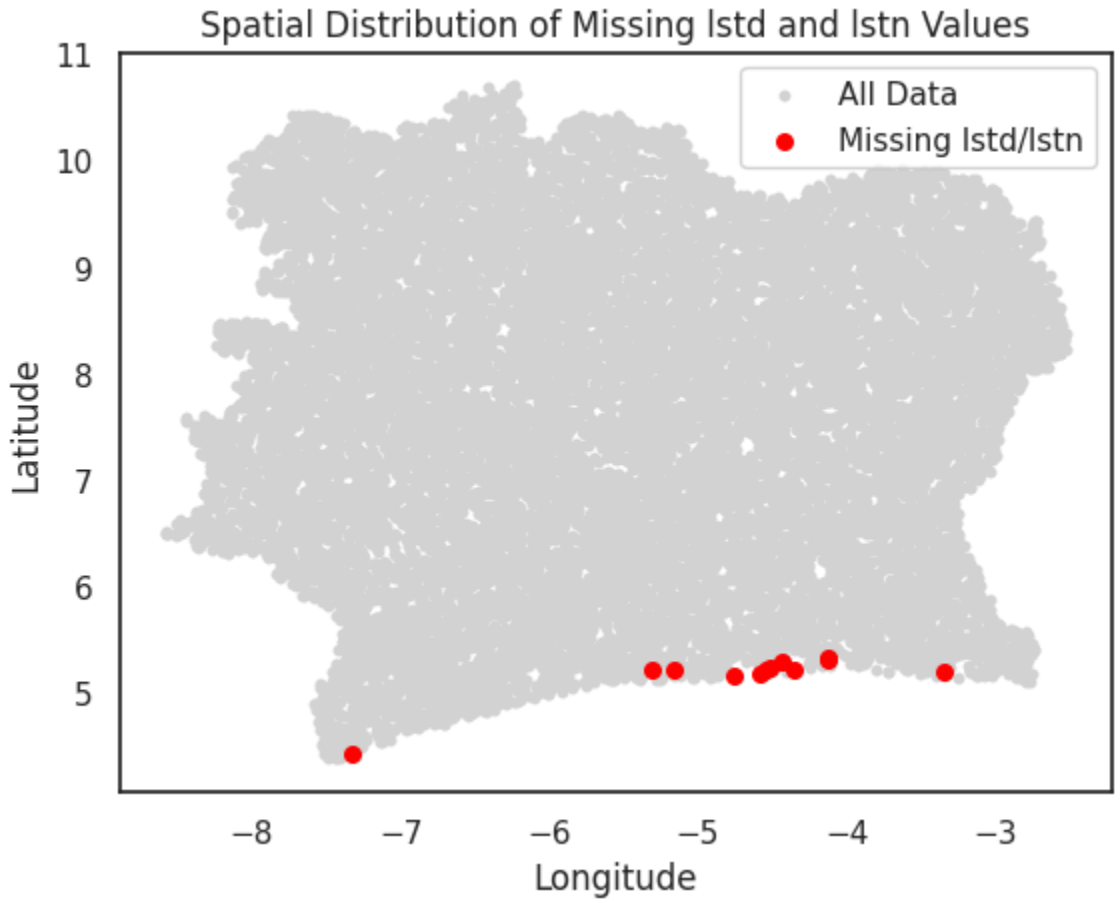
1. Soil Properties:

43 rows with missing soil variables (bd20, cec20, ph20, snd20, soc20) were dropped (≈0.3% of data).

2. LST Variables:

The day-time (lstd) and night-time (lstn) land surface temperatures had 12 missing values. Their spatial distribution showed these missing points clustered near the southern/southwestern coast, likely due to persistent cloud cover or sensor issues. We used the KNN Imputer (using predictors

such as lat, lon, and elevation (mdem)) to impute these values—capturing coastal versus inland differences.



3. Target bcount in Test Set:

The bcount feature (which I assumed to be building count) was present in training but completely missing in the test set. Given that training data is highly skewed (91.77% zeros), and Considering the low median value and high zero proportion, we opted for median imputation (training set median) for the test set.

Undocumented Feature Analysis

- bcount:Shows a strong negative correlation with the target. Its distribution is highly skewed with most values zero. Retained it for its strong signal re
- dnlt, nppm, and sirs:These features showed weak to moderate correlations with the target. Although their individual correlations are not high, they might add predictive value when combined with other features.

Model Training and Evaluation

Data Balancing and Scaling

- SMOTE was applied to the training set to address class imbalance, particularly increasing the representation of the minority "Buildings" class.
- All numeric features were standardized using StandardScaler.

Model Training and Hyperparameter Tuning:

Multiple classifiers were evaluated, and Random Forest emerged as the top performer.

	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	Random Forest	0.796396	0.832330	0.833761	0.832900	0.910096
1	Gradient Boosting	0.768890	0.813332	0.821957	0.815766	0.898368
2	K-Nearest Neighbors	0.564654	0.507762	0.570659	0.502690	0.732731
3	Decision Tree	0.726209	0.777897	0.783590	0.779670	0.803702
4	XGBoost	0.791021	0.827606	0.822935	0.825206	0.901637

Hyperparameter tuning via GridSearchCV optimized Random Forest parameters to:

n_estimators = 300, max_depth = None, min_samples_split = 2, min_samples_leaf = 1.
Evaluation Metrics:
On the evaluation set, the tuned Random Forest achieved:

Macro F1 Score: 83.5%
ROC-AUC: 91.2%
These metrics ensured balanced performance across classes.
Feature Importance:
Analysis showed that bcount is the dominant predictor (34.5% importance), with other features (including remote sensing indices and environmental variables) contributing additional but lower signals.

Final Predictions and Model Saving:
The final model was applied to the test dataset (after identical preprocessing) to generate predicted probabilities for each class. The submission file was created with columns: subid, building_prob, cropland_prob, and wcover_prob. The model was saved using joblib for future inference.

Critical Findings

- The bcount column (which upon googling, and based on the analysis I took it to represent building count) feature is the single most important variable, indicating that urban density is a strong indicator for land cover classification in this region.

- Moderate correlations of features like log-transformed dnlt and nppm suggest that natural environmental signals play a key role in distinguishing woody vegetation from cropland and urban areas.
- Despite lacking formal descriptions, features like bcount, dnlt and nppm were shown to add valuable signals to the model.

Recommendations

- Use the optimized Random Forest classifier for final predictions, as it offers the best and tuned performance and interpretable feature importances.

- Some assumptions were made concerning the undocumented features, and how the unified column was formed. Validate assumptions regarding the features (e.g., bcount as building count) through domain experts.