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 . In addition, the model outputs class occurrence probabilities.

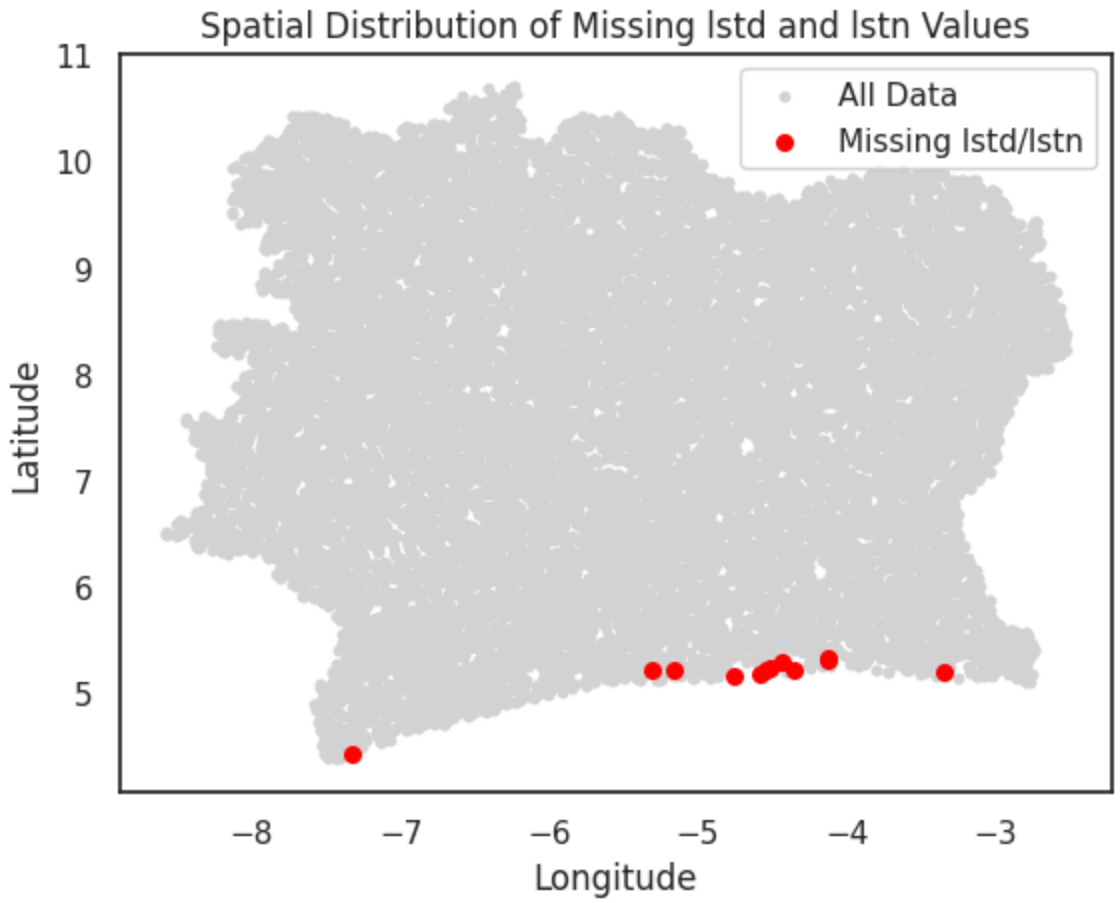
Methodology and Approach

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 target variables were provided as three separate indicators: building, cropland, and wcover, which were merged into a unified target, and then label encoded for modeling.

Data Preprocessing and Exploratory Data Analysis

Assessed missing values using a heatmap , which revealed that most features had complete data and some missing values in a few columns . Most columns had complete data, with only minor missingness in a few variables (e.g., bio1, cec20, ph20, snd20) that were dropped.
The lstd(Average day-time land surface temp. (deg. C , 2001-2020) and lstn (Average night-time land surface temp. (deg. C, 2001-2020) also had missing values. Upon investigation of their Spatial Distribution, it showed that the points that were with missing data were all near the southern/southwestern coast, with one row farther west. This could have been caused by persistent cloud cover, coaster influence, some other technical issues.

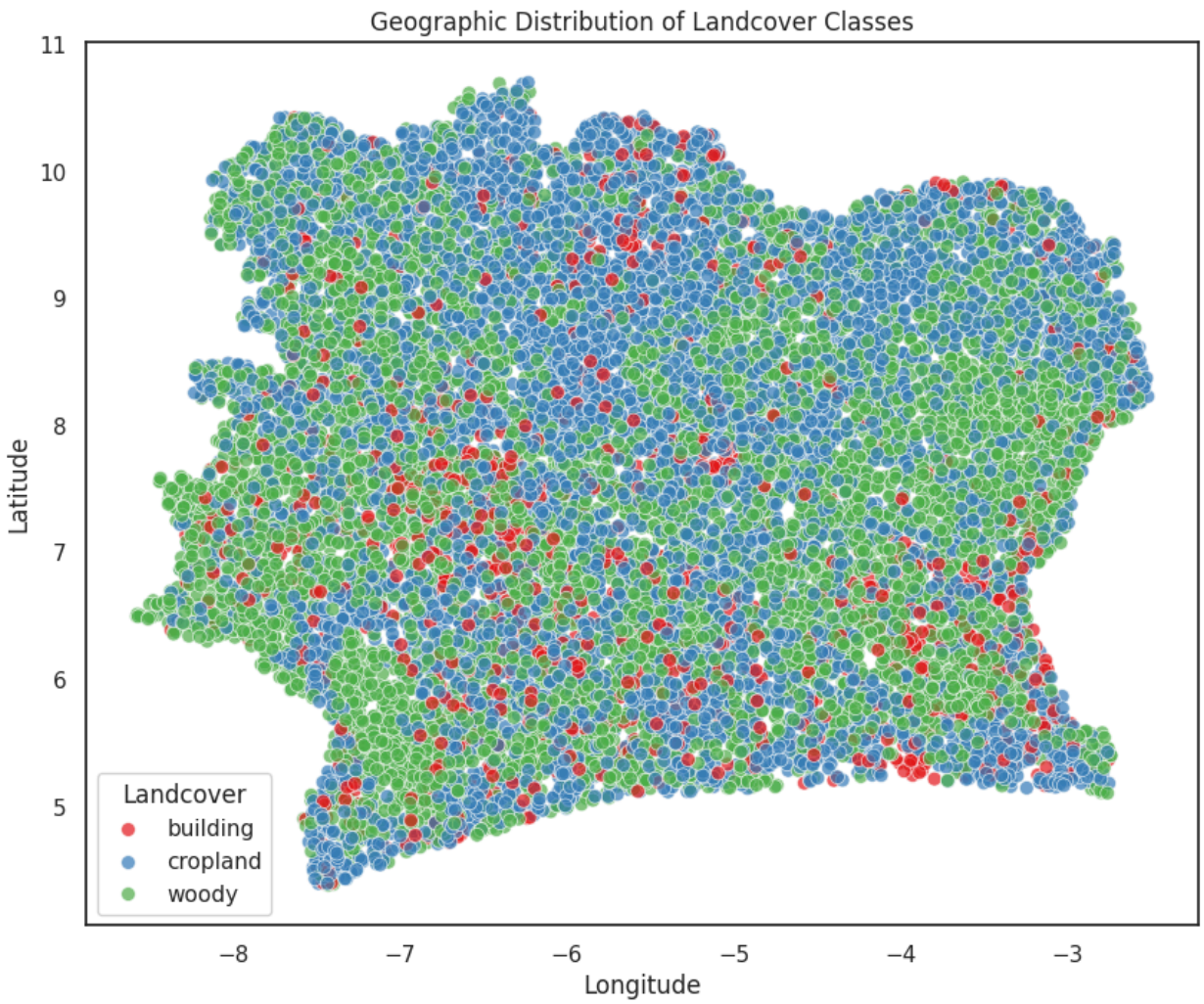


For this case, the missing data was imputed using KNNImputer, which included latitude (lat), longitude (lon), and elevation (mdem). The reason to use KNN Imputer was so that it could look for the closest points and average their known LST values, and therefore this would help capture coastal vs inland differences better.

Exploratory Data Analysis (EDA)

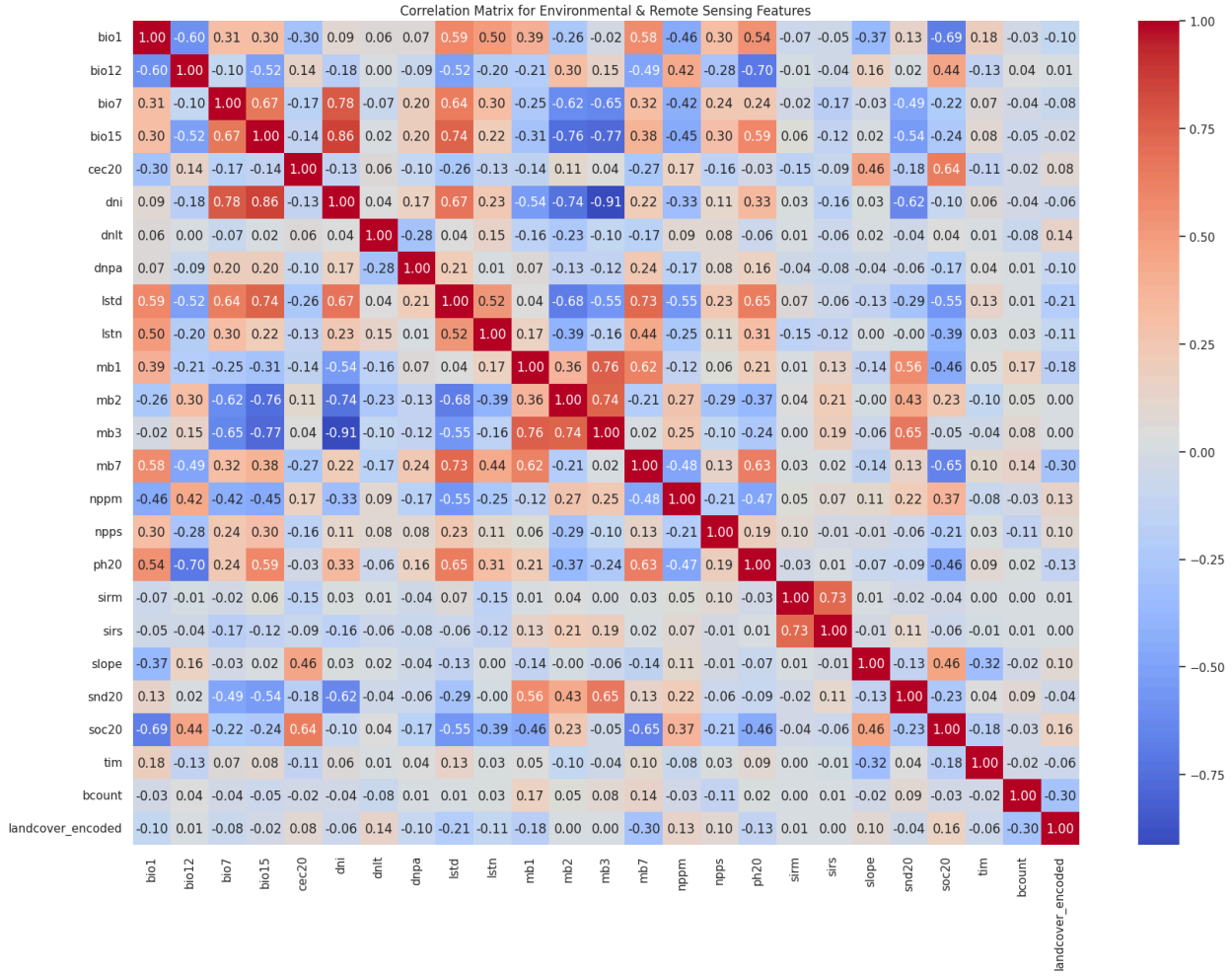
Spatial Distribution:

A scatter plot of latitude vs. longitude, color-coded by landcover, revealed that building-related pixels are relatively sparse and clustered (urban centers), while cropland and woody areas are more widespread.
This visualization validated the imbalanced distribution (buildings were the minority class) and helped identify spatial clusters.



Environmental and Remote Sensing Influences:

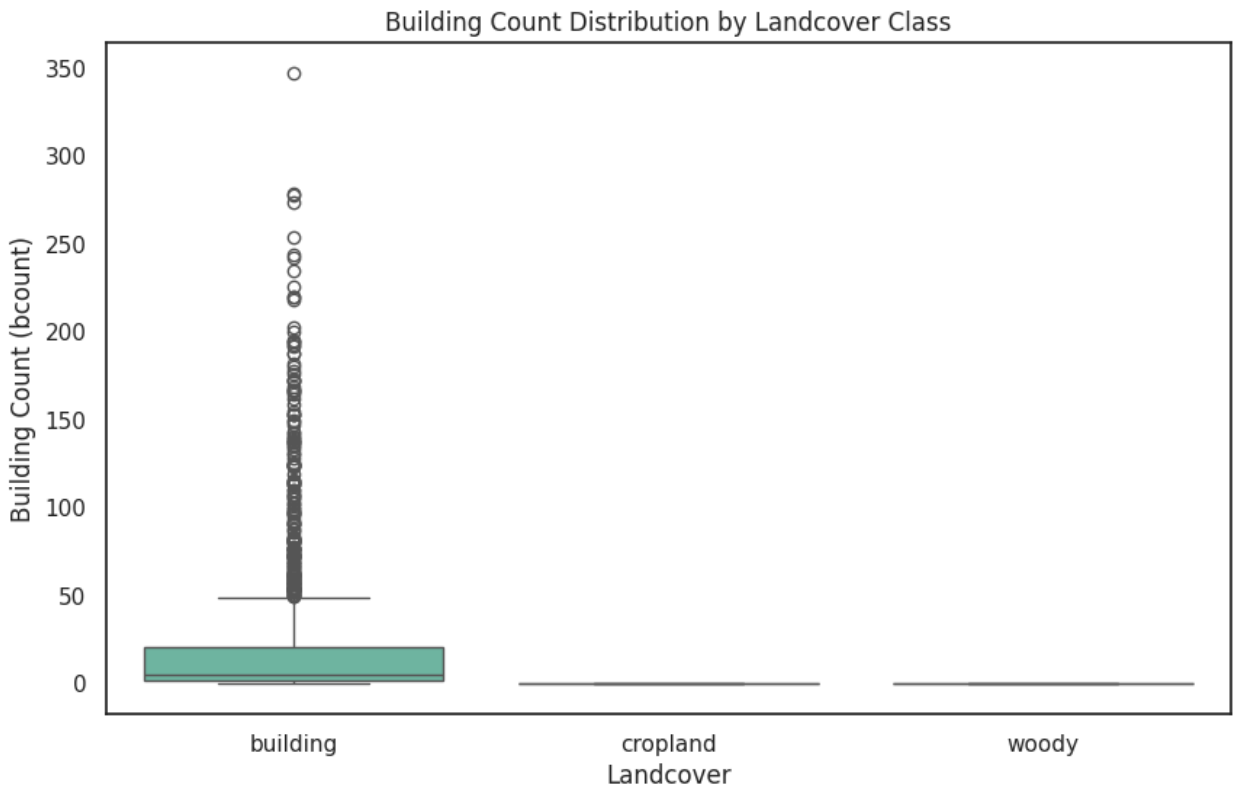
Which environmental and remote sensing features correlate with the land cover classes?
An extended correlation matrix showed that features such as bcount (Which we assimed to be building count) have a strong (negative) correlation, while others (e.g., dnlt, nppm) show moderate positive correlations, indicating they help distinguish natural from built-up areas.



Impact of Urbanization:

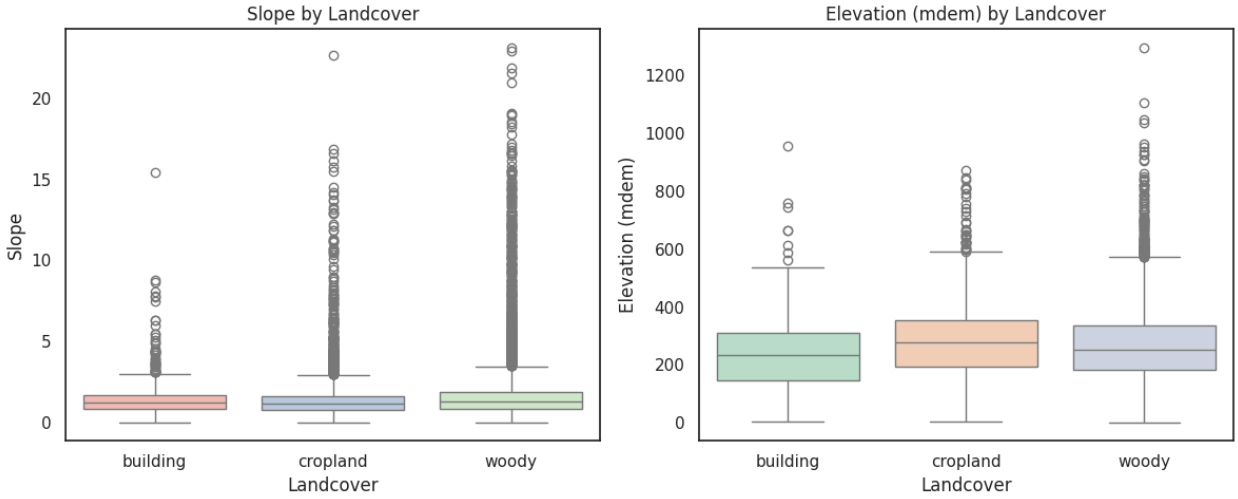
How does building count (bcount) influence the classification?

Boxplots demonstrated that observations labeled as “building” have significantly higher bcount values than cropland or woody areas, emphasizing urban density as a key predictor.The fact that building count only appeared in the building category confirmed our assumption that bcount, represents building count.



Role of Topography and Soil Characteristics:

What is the relationship between topographic features (e.g., slope, elevation) and land cover? The analysis of slope and elevation (mdem) via boxplots showed that buildings tend to occur in flatter, lower-lying regions, while woody areas are associated with steeper slopes and higher elevations.



Utility of Undocumented Features:

Do features without formal documentation (e.g., dnlt, nppm, sirs) add predictive value?

Undocumented features (e.g., dnlt, nppm, and sirs) were retained. Although sirs showed almost no linear correlation with the target, its unique distribution (narrow range) might capture subtle environmental nuances when combined with other variables.

Model Training and Evaluation

Data Balancing and Scaling:

SMOTE was applied to the training data to balance the imbalanced class distribution, particularly boosting the minority "building" 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	0.742333	0.812591	0.810363	0.811472	0.877504
1	Gradient Boosting	0.725577	0.800402	0.798160	0.799275	0.871230
2	K-Nearest Neighbors	0.539045	0.497209	0.551667	0.490493	0.720055
3	Decision Tree	0.654758	0.746705	0.746749	0.746719	0.767122
4	XGBoost	0.723996	0.799265	0.797066	0.798160	0.870535

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:

Accuracy: ~74.2%
Macro F1 Score: ~81.1%
ROC-AUC: ~87.8% (multi-class, one-vs-rest)
These metrics ensured balanced performance across classes.
Feature Importance:
Analysis showed that bcount is the dominant predictor (≈32.6% 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

- Dominant Role of bcount:

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.

- Environmental Gradients Matter:

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.

- Effective Handling of Imbalance:

SMOTE balanced the classes successfully, leading to improved performance on the minority "building" class.

- Value of Undocumented Features:

Despite lacking formal descriptions, features like dnlt and nppm were shown to add valuable signals to the model.

Recommendations

- Adopt the Tuned Random Forest Model:

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

- Incorporate Domain Expertise:

Some assumptions were made concerning the undocumented features. Validate assumptions regarding the features (e.g., bcount as building count) through domain experts.