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.

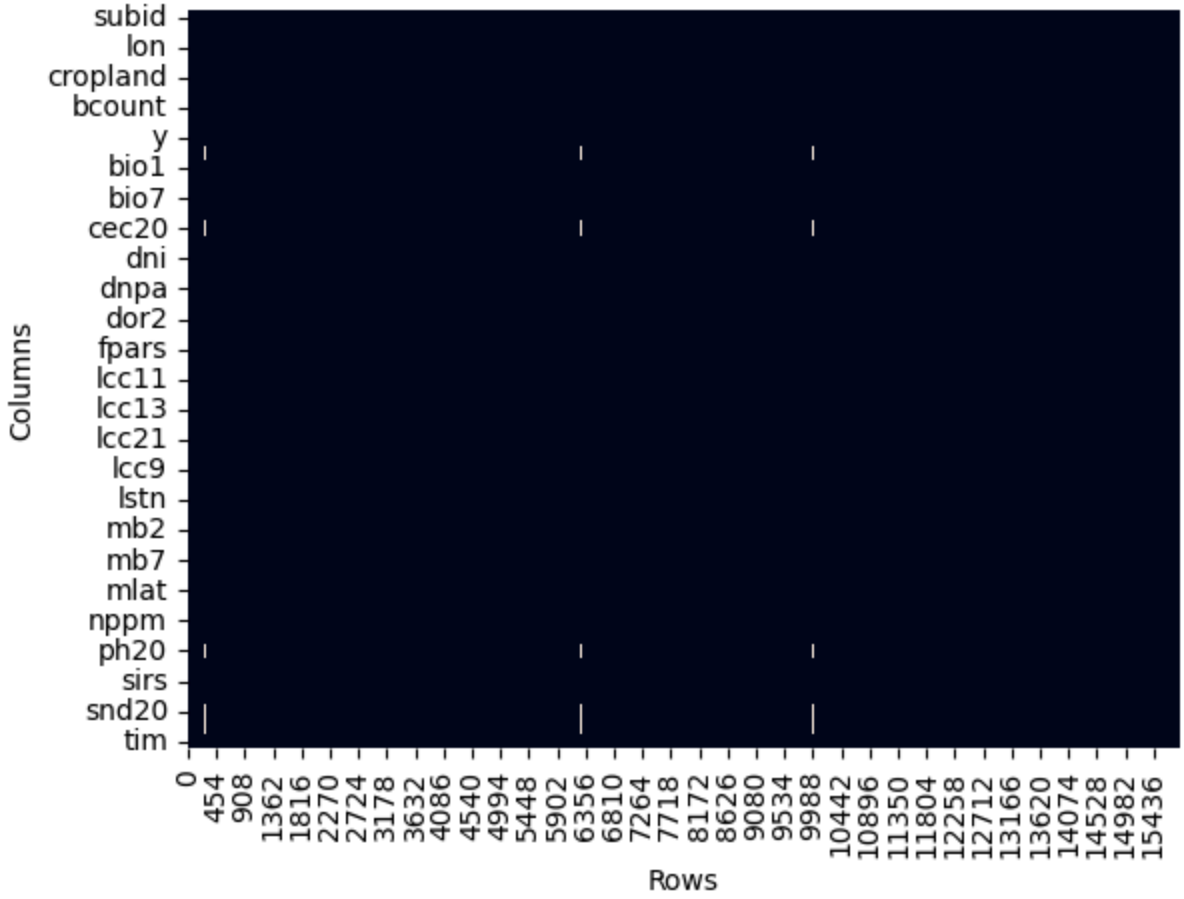
Methodology and Approach

Data Understanding

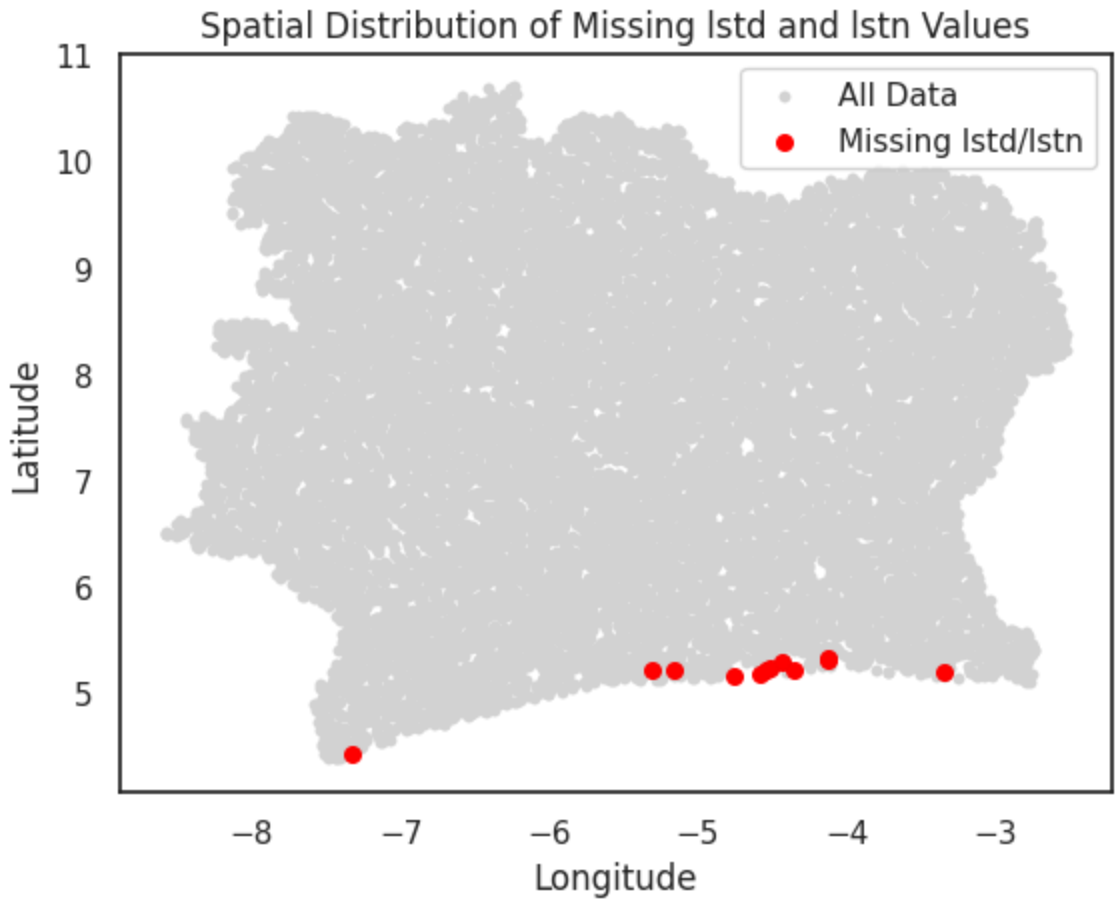
The dataset contained 15,811 training observations and a corresponding test dataset. Key columns included geographic coordinates (lat, lon), environmental features (e.g., bio1, bio12, bio7, bio15), remote-sensing variables (e.g., mb1, mb2, mb3, mb7), soil properties (bd20, cec20, ph20, soc20), and additional 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 (see Figure 1: Missing Values Heatmap), which revealed that most features had complete data and some missing values in a few columns . bio1,cec20,ph20,snd20 had missing values which showed some pattern, in that the missing values was occured in the same rows. However, upon further exploration, it showed no correlation. The missing values, were only in few columns, so I went on to delete them



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.

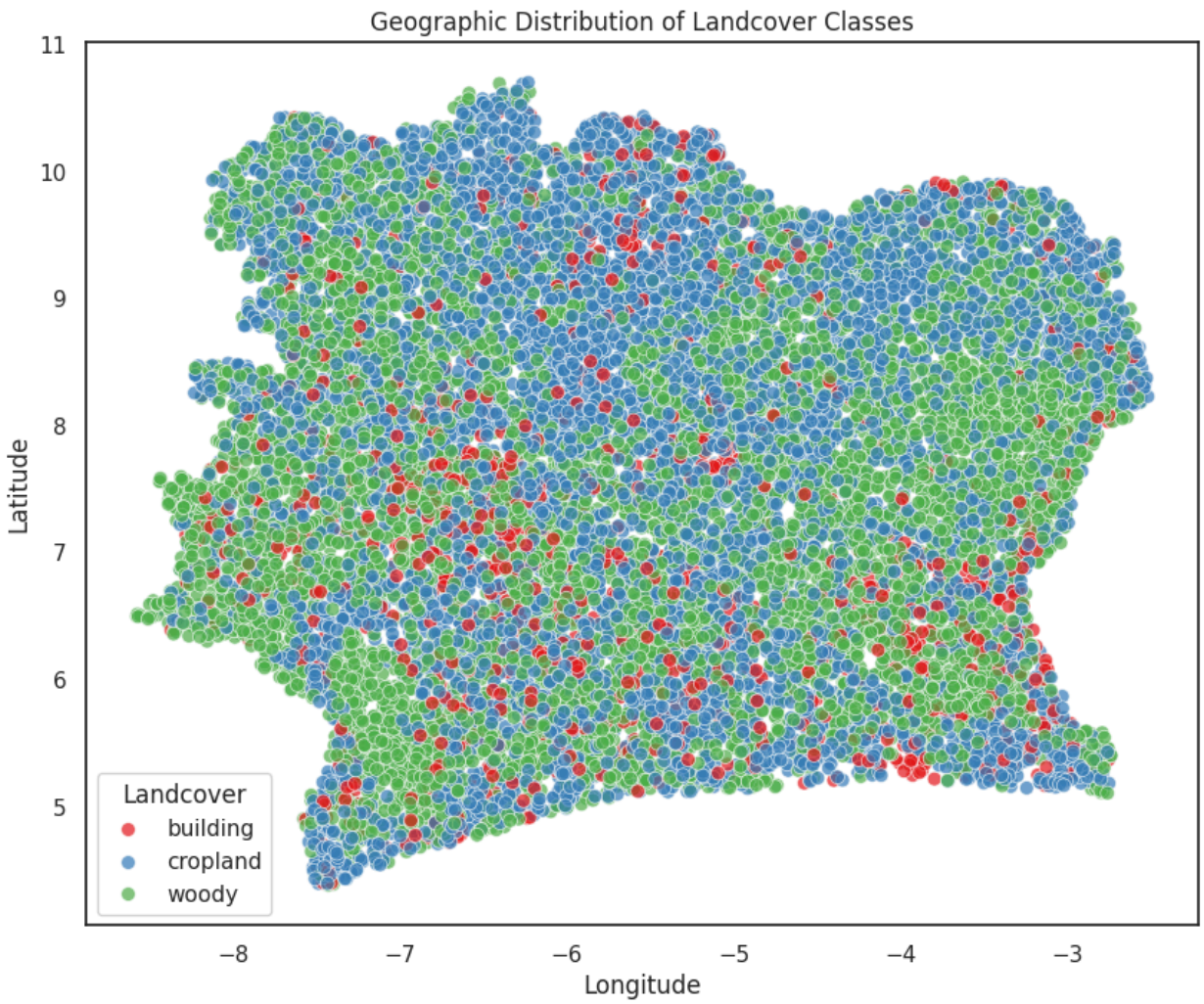


For this case, the missing data was imputed using KNNImputer, which included latitude (lat), longitude (lon), and elevation (mdem). The reason to use KNN Imputer was so that it could look for the closest points and average their known LST values, and therefore this would help capture coastal vs inland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

Exploratory Data Analysis (EDA)

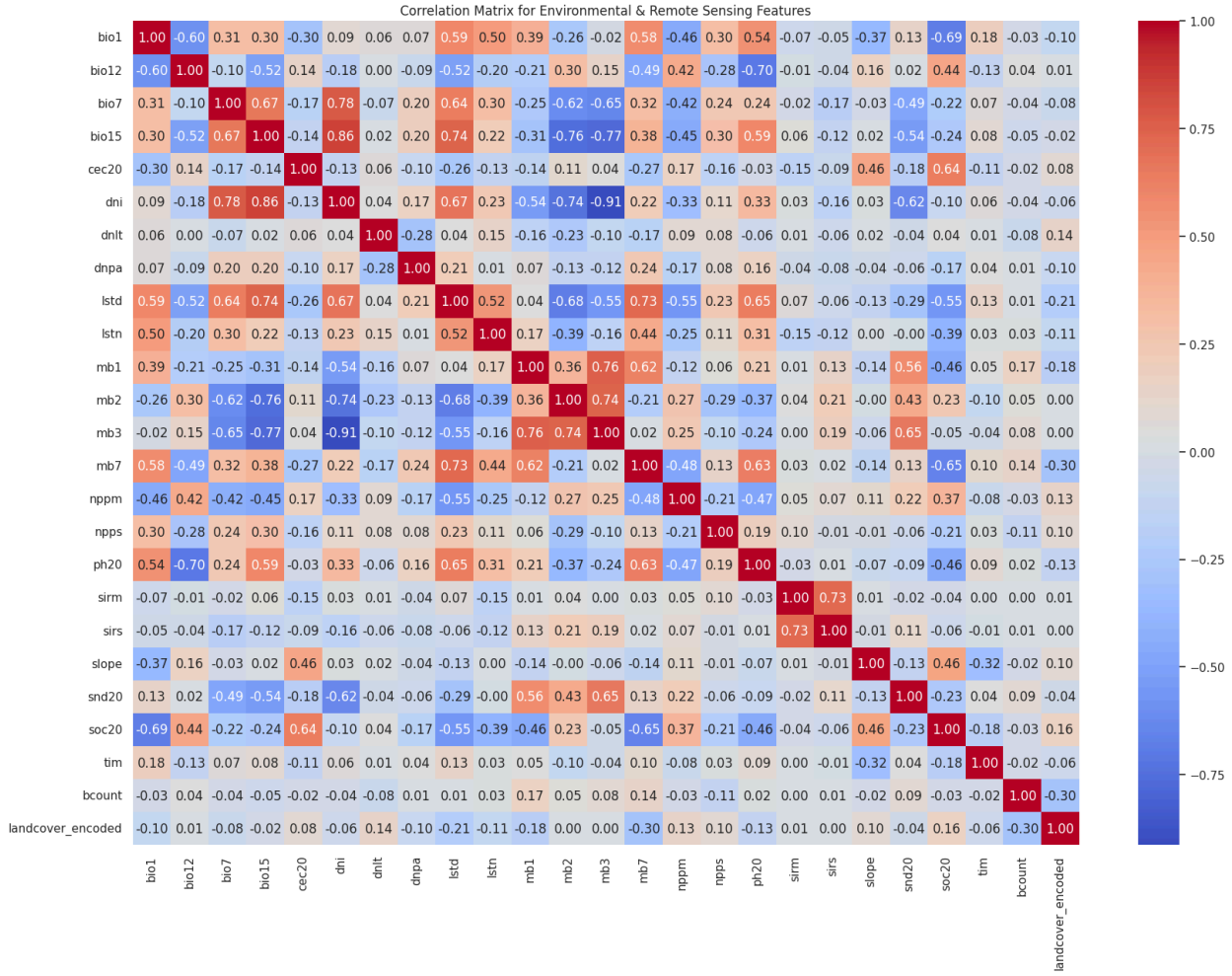
Spatial Distribution:

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:

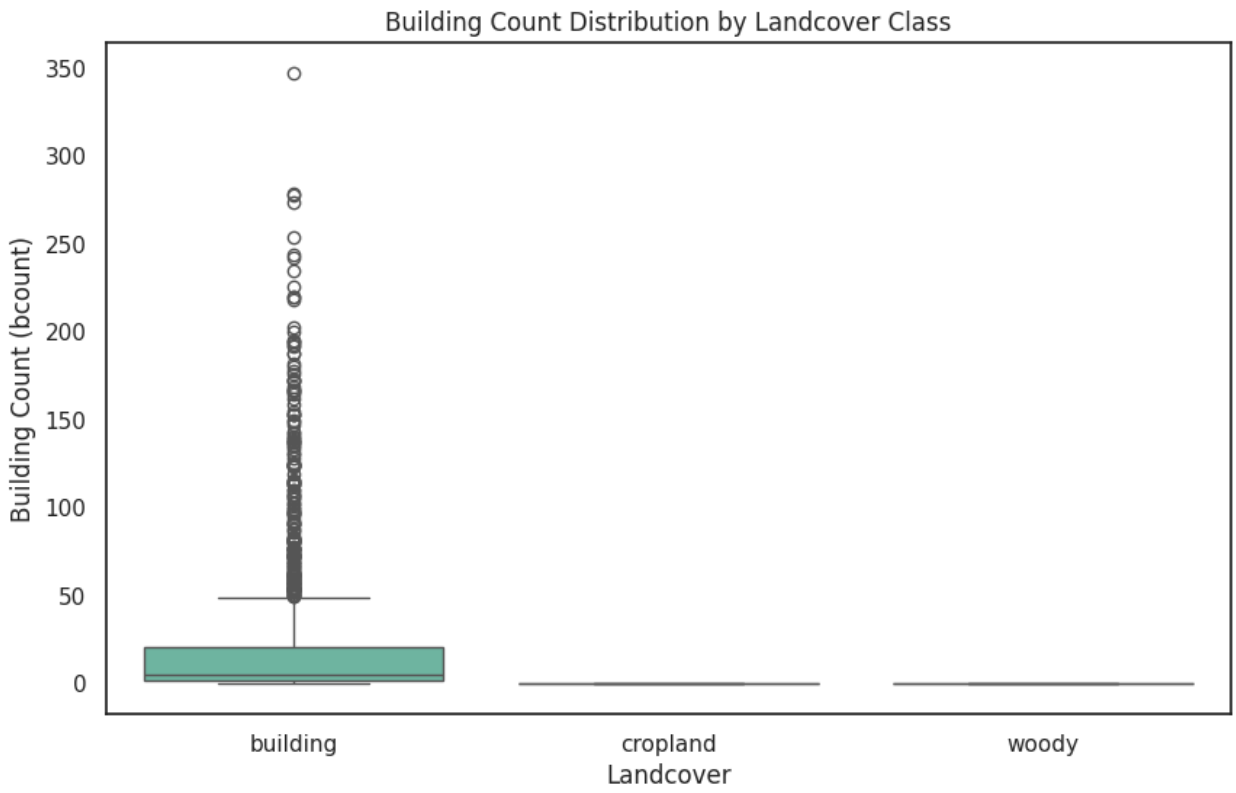
Which environmental and remote sensing features correlate with the land cover classes?
An extended correlation matrix showed that features such as bcount (Which we assimed to be building count) have a strong (negative) correlation, while others (e.g., dnlt, nppm) show moderate positive correlations,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



Impact of Urbanization:

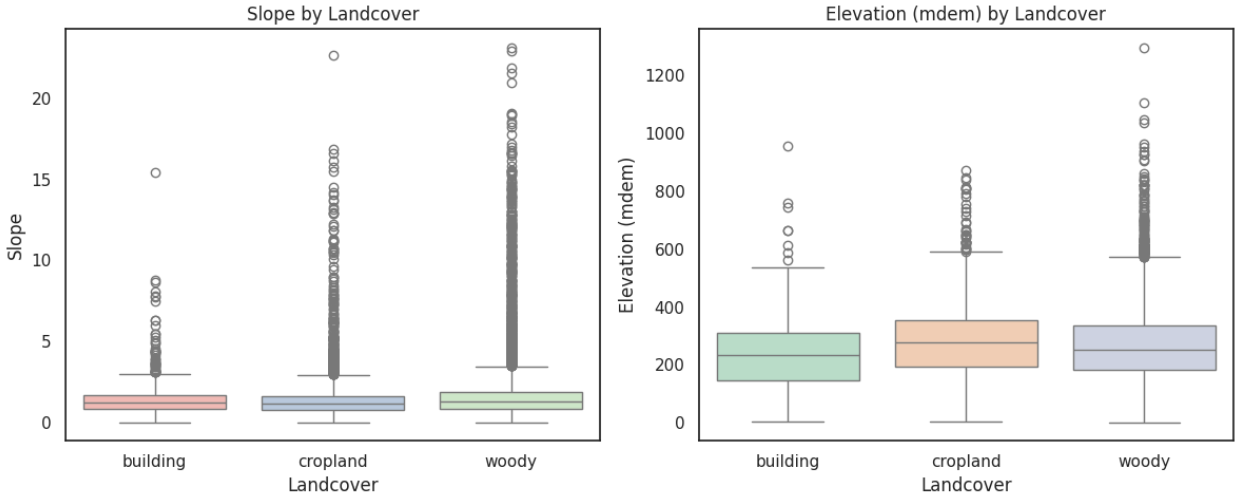
How does building count (bcount) influence the classification?

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

Feature Engineering and Encoding

To simplify the prediction task, we consolidated the three target columns into a single unified target variable ("landcover") using a hierarchical rule:

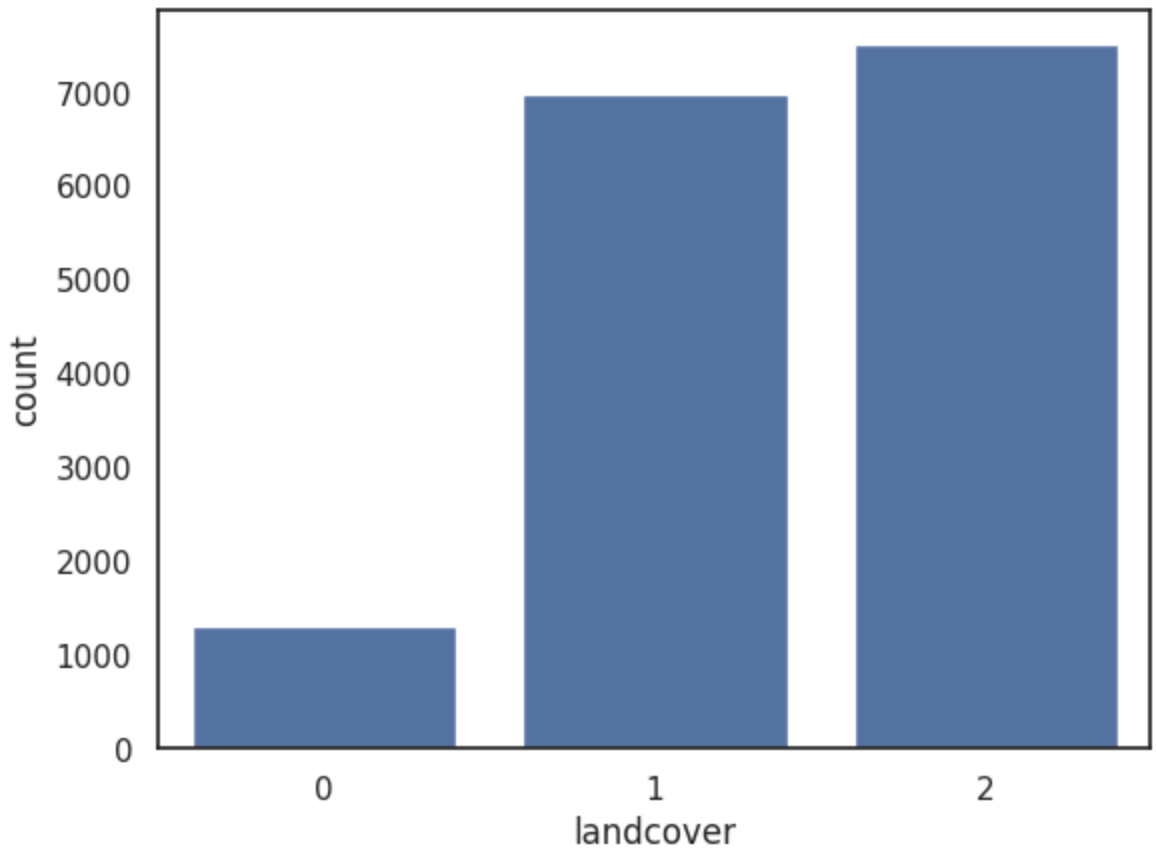
If building is "Yes", the observation is labeled as building.
Else, if cropland is "Yes", label it as cropland.
Else, if wcover indicates ">60%", label it as woody.

Handling Class Imbalance

The initial distribution of the target was imbalanced:

Building: 1,308 instances
Cropland: 5,232 instances (calculated)
Woody: 7,511 instances

The graph below shows the distribution:



To counteract this, we applied the SMOTE (Synthetic Minority Over-sampling Technique) to the training data. This technique generated synthetic samples for the minority class (buildings) so that all classes were balanced before training, thus mitigating bias towards majority classes.

Model Training and Evaluation

Model Selection and Tuning:

Multiple classifiers (Random Forest, Gradient Boosting, K-Nearest Neighbors, Decision Tree, and XGBoost) were evaluated.
Random Forest outperformed other models with an accuracy of ~74.2%, a macro F1 score of ~81.1%, and a ROC-AUC of ~87.8%.

	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 using GridSearchCV further refined the Random Forest settings (e.g., n_estimators=300, max_depth=None, min_samples_split=2, min_samples_leaf=1).

Feature Importance

- Analysis showed that bcount is the dominant feature (≈32.6% importance).
- Other features, including several remote sensing variables and environmental indices, contributed meaningful but lower signals.

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 Signatures:

Features like dnlt and nppm correlate moderately with the encoded target, hinting at distinct signatures between woody vegetation and other classes

- Effective Handling of Imbalance:

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

- Value of Undocumented Features:

Even without formal descriptions, several variables (e.g., dnlt, nppm) contribute useful predictive signals.