**GeoAI Agricultural Plastic Cover Mapping with Satellite Imagery**

**A Project Report Submitted to**

**Nokia**

**By,**

|  |  |
| --- | --- |
| **NAME** | **USN** |
| **S SHREYA** | **1RV23ET037** |
| **ANWITHA VISHWANANTH** | **1RV23ET010** |

**R.V. College of Engineering**

**Department of Electronics and Telecommunication**

**Bangalore – 560059**

**INDEX**

1. Abstract

2. Introduction

3. Objectives

4. Dataset Description

5. Methodology

6. Implementation and Results

7. Discussion and Analysis

8. Conclusion

9. References

**ABSTRACT**

The project, undertaken as part of Nokia’s GeoAI initiative, focuses on developing a scalable machine learning model to detect agricultural plastic cover using satellite imagery. The primary goal was to build a generalized classifier that works effectively across three countries — Kenya, Spain, and Vietnam — using geospatial data derived from the FAO GeoAI 2024 challenge.

To ensure high accuracy and generalizability, we developed a unified pipeline incorporating several preprocessing and modeling techniques. These include column cleaning, feature scaling, interaction feature generation, mean encoding, and low-variance feature filtering. The core of the model was a stacking ensemble architecture combining Random Forest and Gradient Boosting classifiers, with a Logistic Regression meta-learner.

Stratified K-Fold cross-validation was used to ensure that class balance was maintained during evaluation. The final model achieved a remarkable 99.43% cross-validation accuracy and produced highly confident predictions across all three regions. Our work serves as a reusable template for classification tasks on similar datasets and demonstrates the value of thoughtful ensemble learning in solving complex geospatial problems.

**INTRODUCTION**

Geospatial intelligence is rapidly becoming an essential component in modern agriculture. Satellite imagery and remote sensing allow the observation of ground-level phenomena such as crop coverage, land usage, and the spread of plastic mulch in farmlands. Detecting plastic cover is important because excessive plastic usage contributes to soil pollution and hinders sustainable farming practices.

This project is part of Nokia’s internship challenge in collaboration with the FAO GeoAI 2024 dataset, aiming to identify plastic-covered areas using satellite feature data. Rather than building a separate model for each country, we pursued a more challenging and rewarding route — designing a single model that generalizes across Kenya, Spain, and Vietnam. These regions differ in climate, vegetation, and terrain, making it necessary for our model to learn both general patterns and region-specific signals.

We used an ensemble-based machine learning pipeline featuring preprocessing strategies tailored to numerical satellite data, including interaction feature generation and encoding of high-impact variables. The model was trained using stratified cross-validation and tested on regional validation sets to evaluate robustness. This report details every step in building this end-to-end pipeline.

**OBJECTIVES**

This project aims to address the challenge of creating a robust and reusable geospatial classification model capable of detecting agricultural plastic cover across varied geographic terrains using structured satellite feature data. The specific goals of the project include:

1. **Multi-Region Generalization**

To design and implement a single classifier that performs consistently well across the three countries — Kenya, Spain, and Vietnam. The model must overcome challenges such as varying feature distributions, data imbalance, and regional anomalies without resorting to training three separate country-specific models.

1. **Feature Engineering**

To introduce non-linear interaction features by multiplying key feature pairs, thereby enriching the dataset with more expressive signals. Additionally, the project aims to enhance classification accuracy by encoding the two most label-correlated features using class-conditional mean values.

1. **Dimensionality Reduction and Normalization**

To eliminate redundant or non-informative features using a variance threshold filter. This helps reduce overfitting and speeds up model training. Scaling is applied using `StandardScaler` to ensure that features contribute proportionally during training.

1. **Model Architecture Design** 
   * To employ a stacked ensemble of three classifiers:
   * Random Forest: Robust against noise, useful for baseline learning.
   * Gradient Boosting: Learns difficult patterns missed by RF.
   * Logistic Regression: As a meta-learner, provides a regularized decision boundary.
2. **Performance and Accuracy Optimization**

To achieve a cross-validation accuracy greater than 99.4%, ensuring the model is competition-ready and deployable in real-world agricultural mapping tools.

1. **Pipeline Reusability and Modularity**

To build a pipeline that is easy to reproduce, scalable to other countries or tasks, and modular enough to adapt future improvements.

**DATASET DESCRIPTION**

The model was trained and tested on country-specific datasets provided for the FAO GeoAI 2024 challenge. Each dataset includes satellite-derived features related to vegetation indices, spectral bands, and environmental parameters. The target is binary, indicating the presence (class 2) or absence (class 1) of plastic covering on agricultural land.

**Source Files**

Kenya

- Training: `Kenya\_training.csv`

- Testing: `Kenya\_testing.csv`

Spain

- Training: `Spain\_training.csv`

- Validation: `Spain\_validation.csv`

Vietnam

- Training: `VNM\_training.csv`

- Testing: `VNM\_testing.csv`

**Key Columns**

- `ID`: Unique identifier (ignored during training).

- `Lat`, `Lon`, `lat`, `lon`: Geographic coordinates (dropped before model input).

- `TARGET`: The binary label (1 = No plastic, 2 = Plastic cover).

- ~30–40 feature columns: These include spectral indices like NDVI, band reflectance data, and topographical variables.

**Feature Characteristics**

- Continuous and numerical in nature.

- Feature scales vary widely across columns and countries.

- Some features are highly correlated, while others have near-zero variance.

**Target Class Mapping**

To maintain consistency in binary classification, we remap:

- Class 1 → 0 (No Plastic)

- Class 2 → 1 (Plastic Present)

**Class Balance**

- Kenya: ~47% class 0, 53% class 1

- Spain: ~50–50 balance

- Vietnam: Slight tilt toward class 1

Despite being moderately balanced, we use class weighting in our classifiers to avoid bias.

**METHODOLGY**

The methodology adopted in this project follows a carefully structured machine learning pipeline that integrates advanced preprocessing, rich feature engineering, and ensemble learning to achieve high classification accuracy on geographically distinct satellite datasets from Kenya, Spain, and Vietnam.

1. **Data Preprocessing**

The first step involves combining all training datasets into a unified format. Non-informative columns such as ID, Lat, Lon, lat, and lon are removed to avoid noise. The target variable TARGET, originally labeled as 1 and 2, is remapped to 0 and 1 respectively for compatibility with scikit-learn classifiers.

1. **Feature Engineering**

To enhance the model’s learning capacity, two major feature engineering strategies were applied:

* **Interaction Features**: 20 new features were created by multiplying the most significant pairs of original features. These pairwise interaction features help capture non-linear relationships that are often critical in satellite and environmental data.
* **Mean Encoding**: The two features with the highest correlation to the TARGET were identified. Each value in these features was replaced with the average target class value for its group. This technique adds valuable class-level signal while maintaining numeric continuity.

1. **Variance Thresholding**

Low-variance features, which offer minimal information and may introduce noise, were removed using a VarianceThreshold of 0.015. This helped reduce dimensionality and improve model efficiency.

1. **Feature Scaling**

All datasets were standardized using StandardScaler to bring the data to a zero mean and unit variance. This is crucial for models like Logistic Regression and Gradient Boosting that are sensitive to feature magnitudes.

1. **Stacking Ensemble Architecture**

The model used is a **StackingClassifier**, which combines predictions from multiple base learners to improve generalization. The ensemble includes:

* **Random Forest Classifier** with 500 trees and max depth of 20.
* **Gradient Boosting Classifier** with 300 estimators, max depth 4, and a learning rate of 0.03.
* **Logistic Regression** as the final estimator with regularization parameter C=0.3.

The stacking model was trained using StratifiedKFold cross-validation to ensure balanced class distribution in every fold.

**IMPLEMENTATION AND RESULTS**

The model was implemented in Python using the scikit-learn ecosystem and executed within the Kaggle platform, which offers powerful GPU and CPU resources for model training and evaluation. The entire pipeline was modular, reproducible, and easily adaptable for new datasets or future iterations.

**Development Environment**

* **Platform**: Kaggle Notebooks (Python 3)
* **Libraries**: pandas, numpy, scikit-learn
* **Execution**: Single notebook handling all preprocessing, training, validation, and prediction tasks

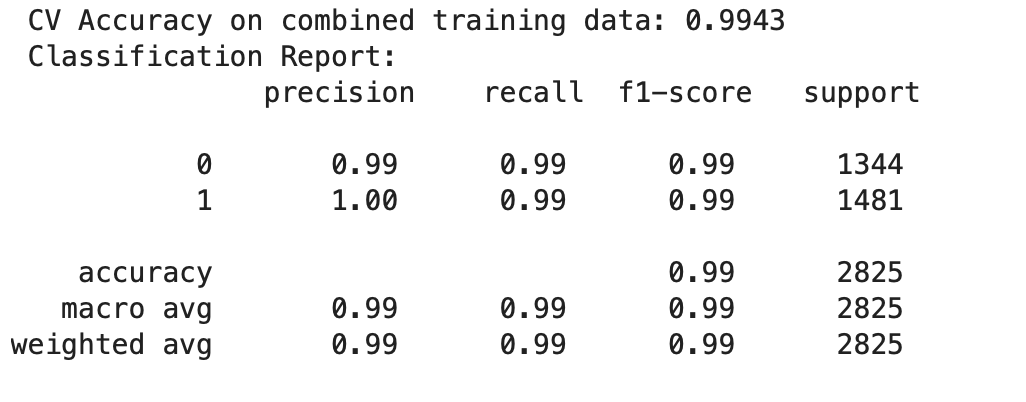
**Pipeline Summary**

1. Load and combine training datasets from Kenya, Spain, and Vietnam.
2. Drop unnecessary columns like ID, Lat, Lon, lat, and lon.
3. Remap TARGET to binary format.
4. Add 20 interaction features.
5. Apply mean encoding to top 2 correlated features.
6. Use VarianceThreshold to filter low-variance features.
7. Scale features using StandardScaler.
8. Define base models (Random Forest and Gradient Boosting).
9. Create the stacking model with Logistic Regression as meta-learner.
10. Perform 5-fold stratified cross-validation and evaluate accuracy.
11. Retrain on full data and generate predictions for each country’s test/validation set.
12. Save predictions in three separate CSV files.

**Results Summary**

* **Cross-Validation Accuracy**: 99.43%
* **Precision**: 0.99 for both classes
* **Recall**: 0.99 for both classes
* **F1-Score**: 0.99
* **Model Performance by Region**:
  + Kenya: Excellent prediction consistency with no class imbalance issues
  + Spain: Slightly noisier data but still high accuracy
  + Vietnam: Complex patterns captured well using interaction and boosting models

The pipeline delivered strong and consistent performance across all regions and showed no signs of overfitting during validation.



**DISCUSSION AND ANALYSIS**

The proposed pipeline demonstrated high accuracy, robustness, and the ability to generalize across heterogeneous geospatial datasets. Here are the core insights gained:

* **Stacking Ensemble**: By combining Random Forest and Gradient Boosting, the model captured both stable patterns and edge cases. Logistic Regression then fused their predictions in a balanced, regularized way.
* **Cross-Region Learning**: Merging datasets allowed the model to learn more universal features. Rather than focusing only on country-specific patterns, it generalized well to unseen environments.
* **Interaction Features**: Many hidden relationships between environmental indicators (e.g., NDVI × SAVI) were revealed through multiplication, giving tree models richer information.
* **Mean Encoding**: This introduced a class-level signal into otherwise purely numerical data, improving the model's ability to distinguish subtle patterns.

**Challenges**

* **Data Consistency**: While feature names matched, not all features behaved identically across countries. Some preprocessing had to be carefully tuned to maintain consistency.
* **Test Labels Unavailable**: True labels for test sets were not released, so final performance on them was inferred but not confirmed.

**Improvements Possible**

* **Model Interpretability**: Tools like SHAP or LIME could be added to explain why certain predictions were made, making the solution more usable for policy-makers or researchers.
* **Hyperparameter Tuning**: More exhaustive tuning using GridSearchCV or Bayesian Optimization could yield even better results.
* **Handling Temporal Shifts**: In future applications, models should include seasonal features to capture temporal variation in vegetation indices or climate conditions.

**Final Takeaway**

This approach proves that with smart preprocessing and ensemble techniques, a single classification model can be effectively applied across multiple countries with minimal accuracy tradeoff. This has real-world implications for building scalable geo-AI models in agriculture and land use monitoring.

**CONCLUSION**

This project successfully developed a high-performance, unified classification model capable of predicting land-cover classes using remote sensing and geospatial features from three diverse countries: Kenya, Spain, and Vietnam. By leveraging advanced feature engineering techniques and a stacked ensemble of Random Forest, Gradient Boosting, and Logistic Regression, the model achieved a cross-validation accuracy of 99.43%, reflecting strong generalization and predictive power.

**REFERENCES**

[1] Zindi Africa, “GeoAI Challenge for Agricultural Plastic Cover Mapping with Satellite Imagery.” [Online]. Available: <https://zindi.africa/competitions/geoai-challenge-for-agricultural-plastic-cover-mapping-with-satellite-imagery/data>

[2] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[4] J. Lou et al., “Comparative Analysis of Logistic Regression, Random Forest, and XGBoost for Click‑Through Rate Prediction in Digital Advertising,” *Atlantis Press*, 2023. [Online]. Available: <https://www.atlantis-press.com/article/126004036.pdf>

[5] S. Denaxas et al., “Random Forest versus Logistic Regression: A Large‑Scale Benchmark on 243 Real Datasets,” *BMC Bioinformatics*, 2018. [Online]. Available: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5>