Enhancing Vegetation and Land Cover Classification Accuracy: A Supervised and Unsupervised Learning Framework with OpenStreetMap and Satellite Imagery

Sajib (Ryhan) Suny, Sharmi Das, Shalini Ivaturi and Siddhartha Choudhary

E-mail: ryhansunny@gmail.com (Sajib (Ryhan) Suny), dassharmi6@gmail.com (Sharmi Das), vshalini227@gmail.com (Shalini Ivaturi), siddhurc@gmail.com (Siddhartha Choudhary)

This study explores the integration of crowdsourced mapping data from OpenStreetMap with Landsat time-series satellite imagery to automate and enhance the classification of satellite images and its accuracy, into distinct land cover classes. Motivated by the potential for crowdsourced data to enhance the accuracy of land cover classification, the project aims to address the inherent challenges of data noise from cloud cover and labeling inaccuracies. It also compares supervised and unsupervised learning approaches and explores the efficacy of feature reduction techniques to refine classification accuracy as additional approaches.

Key words: Vegetation and Land Cover Classification, Supervised Learning, Unsupervised Learning, Crowdsourced Mapping, OpenStreetMap, Landsat Time-Series Imagery, K-Means Clustering, Random Forest, XGBoost.

INTRODUCTION

The aim of this study is to leverage crowdsourced data from OpenStreetMap and Landsat time-series satellite imagery to automate and enhance the classification of satellite images into different land cover classes. The dataset used in this project consists of geospatial data collected during the years 2014 and 2015, incorporating both Landsat imagery and crowdsourced polygons with land cover labels. The primary challenge lies in addressing noise in both the imagery and crowdsourced data due to factors such as cloud cover and inaccuracies in labeling. This investigation not only anticipates enhancing the precision of land cover classification but also aspires to contribute significantly to the broader field of environmental monitoring and resource management.

MOTIVATION

The motivation for this study is driven by the pressing need to enhance land cover classification, a process crucial for environmental management, sustainable development, and resource allocation. In a world increasingly shaped by climate change and rapid urbanization, precise mapping of land cover is indispensable. While satellite imagery has been instrumental in this domain, it is often hampered by issues such as atmospheric interference. Crowdsourced data emerges as a valuable asset that can potentially fill the gaps left by remote sensing, offering detailed local knowledge that may otherwise be overlooked Crowdsource mapping provides a crucial supplement to satellite imagery, offering insights on ground level that also enhances the robustness of classification models. This research is therefore propelled by the goal of synthesizing these two data streams—each powerful on its own—to create a more accurate and reliable approach to land cover classification. By doing so, it aims to address common challenges of data noise and labeling errors, thereby advancing our capacity to monitor and respond to environmental changes effectively.

DATA DESCRIPTION

The dataset for this study comprises 10,546 instances and 29 features, sourced from Landsat time-series satellite imagery (2014-2015) and OpenStreetMap's crowdsourced mapping. It is a multivariate collection designed for the classification of six land cover classes: impervious, farm, forest, grass, orchard, and water. Each instance includes a class label and a series of NDVI (normalized difference vegetation index) measurements reflective of vegetation health and land cover status across the two-year span.

The primary challenge in using this dataset for machine learning is the noise present in both the satellite images (due to cloud cover) and the crowdsourced labels (due to potential inaccuracies in polygon labeling). However, the dataset is complete with no missing values, ensuring consistency for algorithm training and testing.

The features include:

- **class:** The categorical target variable indicating the land cover class.
- **max_ndvi:** The highest NDVI value from the time-series, signaling peak vegetation vigor.
- **20150720_N - 20140101_N:** Sequential NDVI values from specific dates, providing temporal insights into land cover dynamics.

ANALYSIS AND METHODOLOGY

In this section, we describe the methodology and analysis steps employed in our project. We start with **Data Collection and Visualization**, where we obtain the dataset from a reliable source and provide an overview of its structure and data types. Additionally, we visualize key aspects of the dataset, such as class distribution and summary statistics. To gain insights into the relationships between variables, we present a **Correlation Heatmap Visualization** using the normalized difference vegetation index (NDVI) values extracted from satellite images. Further, we explore additional steps in our analysis, such as feature engineering and classification models.

a. Data Collection and Visualization

The dataset utilized in this study is a compilation of geospatial information from two primary sources: Landsat time-series satellite imagery and crowdsourced mapping data obtained from OpenStreetMap. The total number of **instances** are **10,546** and the total number of **features** are **29**.

Landsat Time-Series Satellite Imagery

Time Frame: The Landsat imagery covers the period be- tween 2014 and 2015.

Temporal Information: The time-series nature of the im- agery provides temporal insights into land cover changes over the specified period.

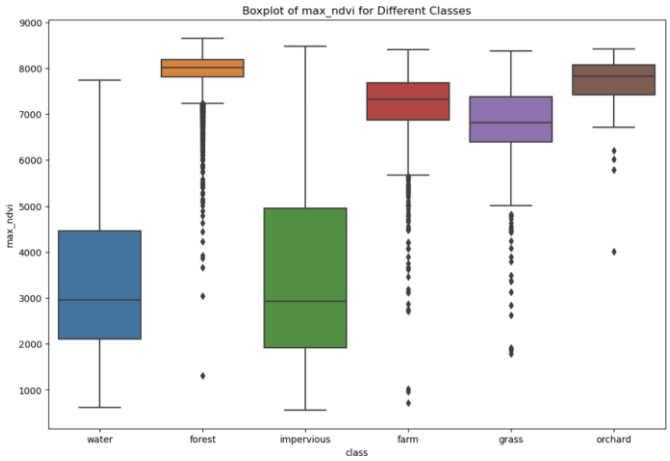

Figure 1. Box Plot of Various Classes Under max_ndvi Feature.

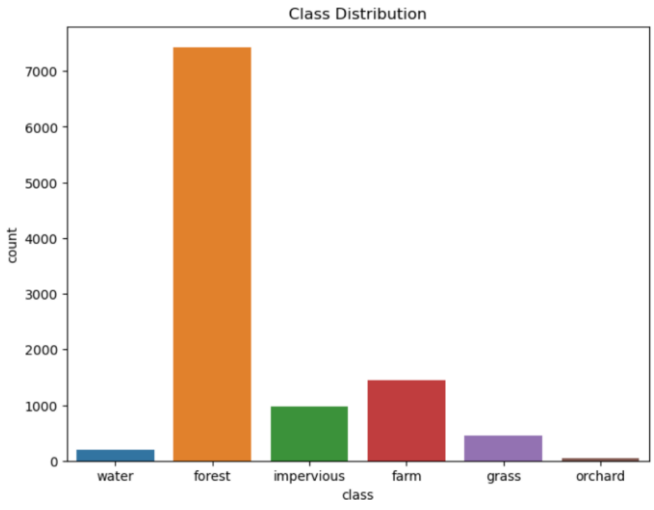

Figure 2. Class Distribution

b. Correlation Heatmap Visualization : The heatmap displays a color-coded matrix visualization which interprets the correla- tion between two features. In the heatmap, the positive correla- tion is depicted in warmer colors, while the negative correlation is represented by cooler colors. This type of visualization aims in identifying feature relationships and multicollinearity. In the

image the diagonal line represents in warmer color as each feature is highly correlated with itself. Also 20130202_n feature is highly correlated with feature max_ndvi.

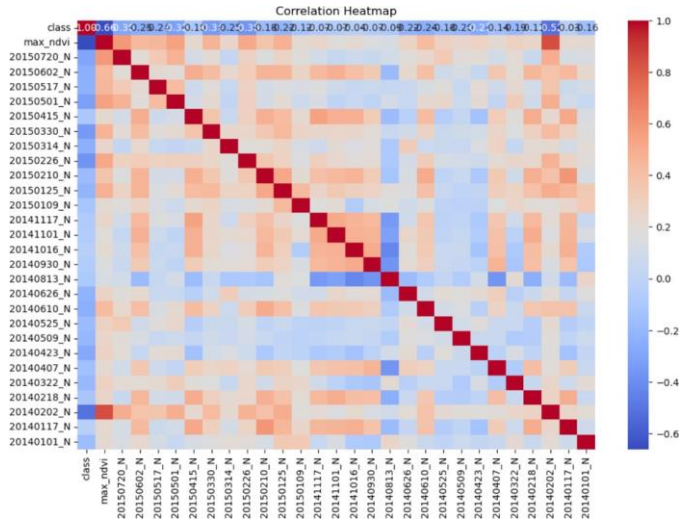

Figure 3. Correlation Heatmap

FEATURE ENGINEERING

Feature engineering is an important step in data pre-processing pipeline which aims to transform the raw data into format suitable for ML and AI models. It involves transforming exist- ing features and ensuring that the data effectively captures the patterns and the relationship between various features. In this project we have applied categorical variable encoding to the class variable which involves encoding categorical features and transform them to numerical data. This facilitates their utilization into machine learning algorithms.

```
In [42]:
from sklearn.preprocessing import LabelEncoder

# Create a copy of the DataFrame
df_encoded = df.copy()

# Instantiate LabelEncoder
label_encoder = LabelEncoder()

# Encode the 'class' column
df_encoded['class'] = label_encoder.fit_transform(df['class'])
```

Figure 4. Label Encoding

CLASSIFICATION MODELS

Unsupervised Learning (K-Means)

K-Means clustering is employed in this project. The relevant columns for clustering, namely 'class' (land cover class) and 'max_ndvi' (maximum NDVI value), are selected. The categorical variable 'class' is transformed into numerical values using label encoding to enable its use in the K-Means algorithm.

The resulting clusters are visualized in a scatter plot, where each point represents a data instance with color-coded clusters based on the K-Means grouping.

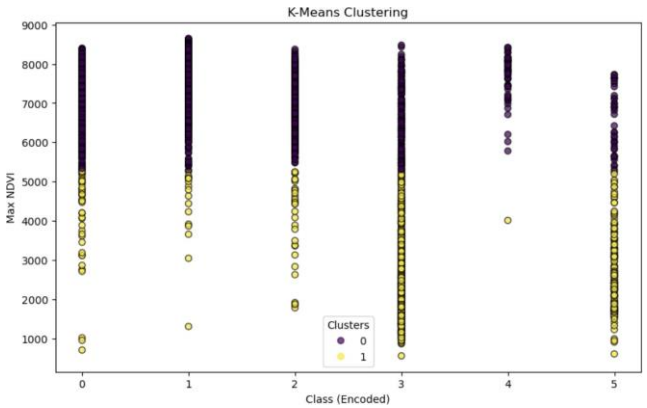

Figure 5. K-Means Clustering

The silhouette score, a metric measuring the separation and compactness of the clusters, is calculated to assess the quality of the clustering. In this instance, the silhouette score is 0.8535, indicating a high degree of separation and cohesion among the identified clusters.

Supervised Learning (Random Forest, XGBoost)

In the context of the project, supervised learning techniques have been instrumental in the classification of land cover classes based on a dataset derived from satellite imagery and crowd-sourced mapping data. Two powerful ensemble learning algorithms, Bagging and Random Forest, were employed.

RESULTS AND MODEL EVALUATION

The Bagging Classifier utilized a Decision Tree as its base classifier, achieving an accuracy of approximately 93.79%. Subsequently, the Random Forest Classifier, comprising 100 decision trees, was applied with additional measures taken to address class imbalance through standardization and oversampling. This model exhibited improved accuracy, reaching approximately 94.17%. Evaluation metrics, including precision, recall, and the F1-score, provided detailed insights into the models' performance across various land cover classes (impervious, farm, forest, grass, orchard, water).

The classification report and confusion matrix presented a comprehensive overview of the models' effectiveness in accurately categorizing land cover types.

```
Accuracy: 0.9416785206258891
Classification Report:
              precision    recall  f1-score   support

        farm       0.87      0.85      0.86       268
      forest       0.96      0.99      0.97      1506
       grass       0.90      0.74      0.81        85
  impervious       0.90      0.84      0.87       202
     orchard       1.00      0.18      0.31        11
       water       1.00      0.81      0.90        37

    accuracy                           0.94      2109
   macro avg       0.94      0.74      0.79      2109
weighted avg       0.94      0.94      0.94      2109

Confusion Matrix:
[[ 228   33    0    7    0    0]
 [   7 1493    2    4    0    0]
 [   4   13   63    5    0    0]
 [  18   10    4  170    0    0]
 [   2    7    0    0    2    0]
 [   2    2    1    2    0   30]]
```

Figure 6. Model Evaluation Score Table

COMPARING RESULTS

The unsupervised learning algorithm, K-Means clustering, achieved a high Silhouette Score of 0.8535, indicating robust cluster separation within the dataset.

In the realm of supervised learning, the DecisionTreeClassifier and BaggingClassifier achieved an accuracy of 0.9379. Subsequently, the RandomForestClassifier, coupled with RandomOverSampler to address class imbalance, achieved an accuracy of 0.9417. The overall accuracy of the models stands at 0.9417.

Additionally, the XGBoost model achieved an impressive accuracy of 0.9635 on the test dataset, showcasing its efficacy in land cover classification.

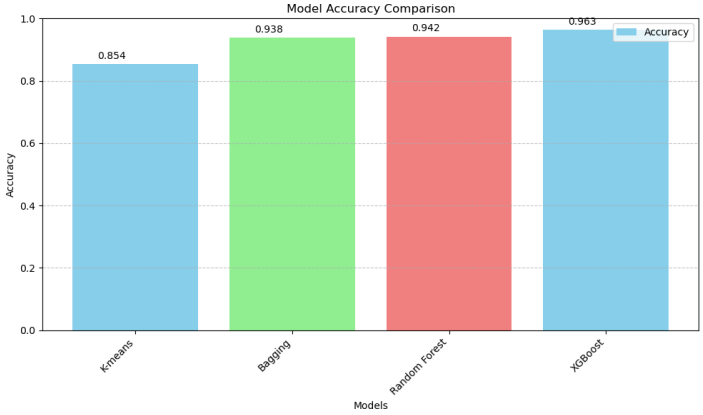

Figure 7. Model Accuracy Comparison

ADDITIONAL APPROACHES WITH

FEATURE REDUCTION

As part of our analysis, we explore an additional approach—Feature Reduction Using Algorithm. We discuss how this technique helps us reduce the of the while preserving important features, ultimately improving model efficiency and interpretability.

First, we determined the random forest feature importance:

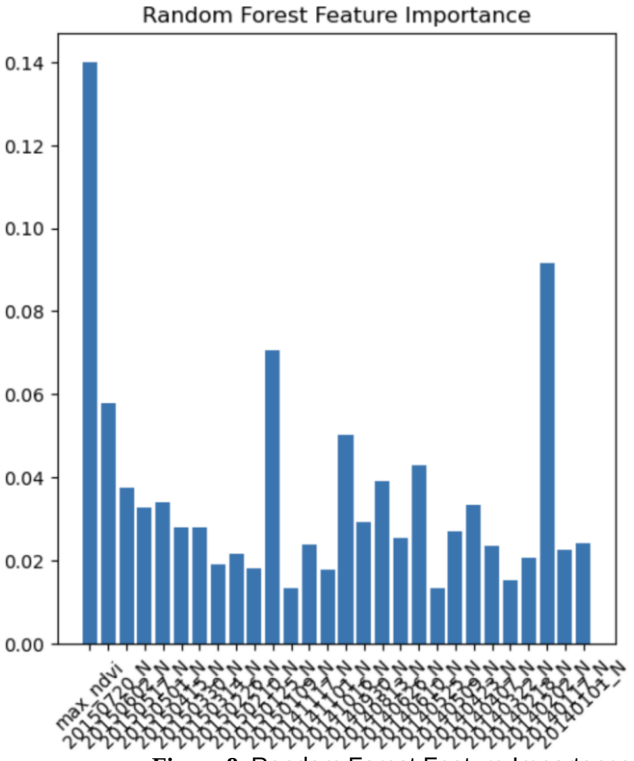

Figure 8. Random Forest Feature Importance

Then, we determined the XGBoost feature importance:

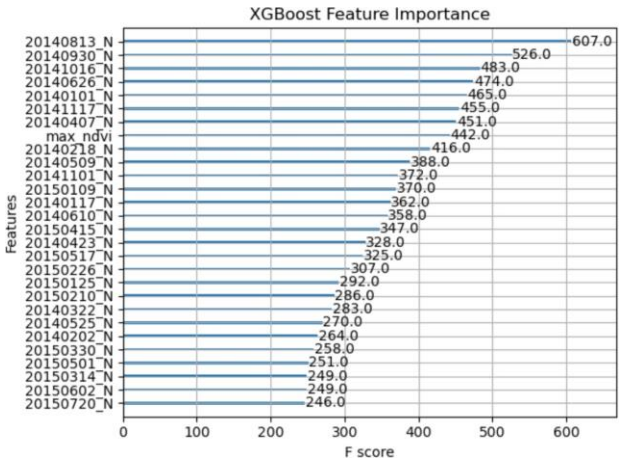

Figure 9. XGBoost feature importance.

Feature Reduction Using XGBoost Algorithm

The analysis of feature importance revealed that XGBoost model's least important feature was '20150720_N' with an importance value of 246.0. To assess the impact of feature importance's, the least important feature was removed from the dataset and the model is trained. Various metrics were calculated like cross-validation and accuracy sures for both the original model and the modified model are calculated. The results is, the original model achieved accuracy of about 91.72 % while the modified model has accuracy of about 91.67%. This demonstrates that the feature importance's are all distribute somewhat equally among various features and that removing one feature removes a bit of information from the data and hence the reason for low accuracy.

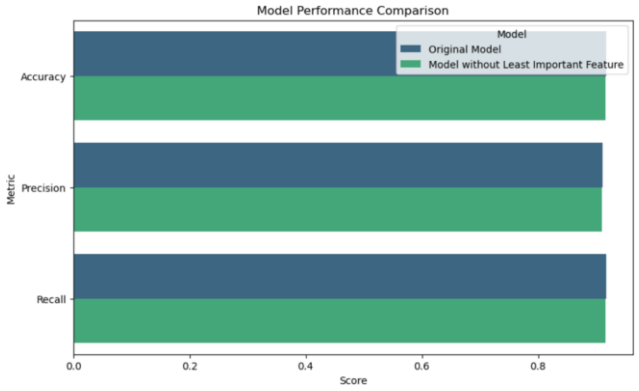

Figure 10. Model Performance Comparison After Feature Reduction

CONCLUSION

In conclusion, the comprehensive analysis of the feature importance, supervised learning, and unsupervised learning components provides valuable insights into the strengths and nuances of the predictive modeling for the given dataset. The examination of feature importance, particularly through the XGBoost model, identified '20150720_N' as the least important feature. However, the subsequent evaluation of model performance following the removal of this feature revealed only marginal changes in accuracy, with the original model achieving 91.72% accuracy and the modified model without the least important feature achieving 91.67% accuracy.

In the context of supervised learning, while utilizing the method of BaggingClassifier and RandomForestClassifier the resultant accuracy is 93.79% and 94.17% respectively. While the XGBoost demonstrated the accuracy of about 96.35%.

Unsupervised learning, as exemplified by K-Means clustering, contributed to the understanding of inherent patterns within the data. The Silhouette Score of 0.8535 indicated a well-defined separation between clusters, reaffirming the dataset's inherent structure.

REFERENCES

Johnson,Brian. (2016). Crowdsourced Mapping. UCI Machine Learning Repository. https://doi.org/10.24432/C56315.

Institute for Global Environmental Strategies. (2012). Integrating Volunteered Geographic Information into Land Use and Land Cover Change Models [PowerPoint slides]. IGES. Retrieved from https://www.iges.or.jp/en/publication_documents/pub/presentation/en/5054/ceres_presentation_12_02.pdf

Nguyen KA, Chen W. DEM- and GIS-Based Analysis of Soil Erosion Depth Using Machine Learning. *ISPRS International Journal of Geo-Information*. 2021; 10(7):452. https://doi.org/10.3390/ijgi10070452

Fritz, S., See, L., Perger, C. *et al.* A global dataset of crowdsourced land cover and land use reference data. *Sci Data* **4**, 170075 (2017). https://doi.org/10.1038/sdata.2017.75

Zhang, G., Roslan, S.N.A.b., Wang, C. *et al.* Research on land cover classification of multi-source remote sensing data based on improved U-net network. *Sci Rep* **13**, 16275 (2023). https://doi.org/10.1038/s41598-023-43317-1