EY Data Challenge

Rice Crop Classification

Ling Xiang, Zou
Ling Feng, Zhou
Fan, Ye

1 Introduction

Rice is a staple food for over half the world's population and plays a vital role in global food security. Accurate and timely rice crop classification is essential for agricultural planning, resource allocation, and sustainable development. Remote sensing technology, particularly satellite imagery, has become a powerful tool for monitoring agricultural resources. High-resolution data from Sentinel-1 and Sentinel-2 missions allow for large-scale crop classification with improved accuracy.

This report aims to create a machine learning model for rice crop classification using multi-temporal satellite imagery. We will explore feature engineering techniques, such as deriving vegetation indices, and evaluate different machine learning algorithms to select the most suitable one.

2 Exploratory Data Analysis

Figure 1 displays a scatter plot of Sentinel-1 VV and VH data, revealing the presence of outliers that can negatively impact model performance. Anomaly detection was conducted using the IQR method to identify and replace these outliers with the mean value of non-outliers.

Figure 2 shows time series data for various vegetation indices. Blue lines represent rice-growing areas, while red lines represent non-rice areas. Patterns in the data help differentiate between rice and non-rice areas.

For certain vegetation indices like VV and VH, we can see that larger values tend to correspond to non-rice areas, while values close to 0 are more indicative of rice-growing areas. In the case of RVI, rice-growing areas seem to be represented by the extreme values (both the highest and lowest values), while non-rice areas are located in the middle range. For SAVI, NDVI, and LAI, the patterns appear to be very similar, with only differences in the magnitude of the values. This observation suggests that these variables may be strongly correlated with each other, and they could potentially be conveying redundant information in the model. This insight could help us further refine our feature selection process and focus on the most relevant and independent predictors for our analysis.

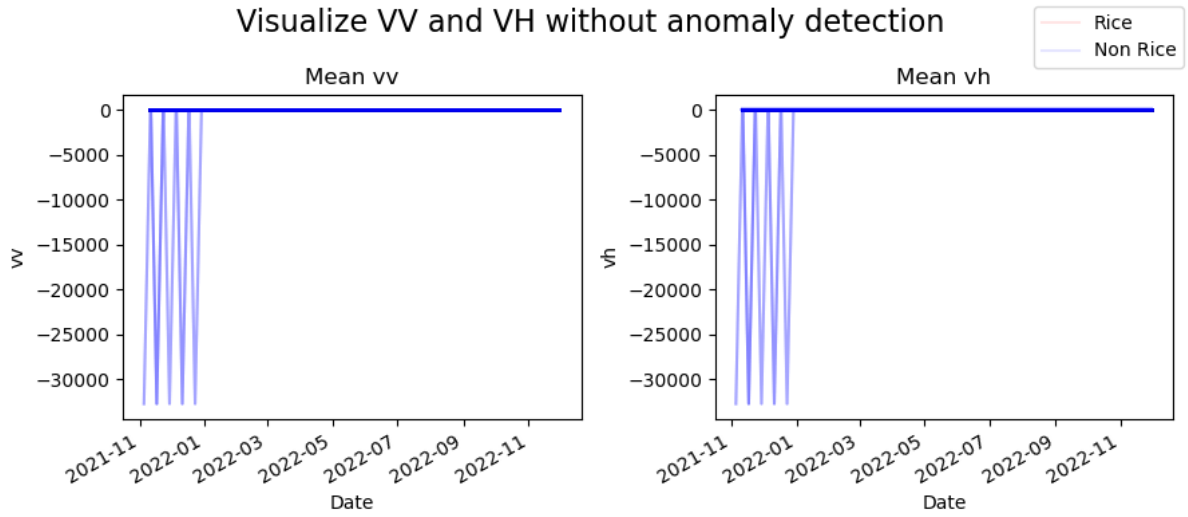

Figure 1: Scatter plot of VV and VH data from Sentinel-1, with extreme values.

Figure 3 visualizes the top 20 features with the strongest correlation with the target variable. As expected, we found that SAVI, NDVI, and LAI were strongly correlated, which is consistent with earlier analyses.

3 Methodology

3.1 Assumption

Our approach assumes accurate data from Sentinel-1 and Sentinel-2 satellites and a known rice crop growing cycle. To ensure consistency, the data was acquired under similar conditions, such as at similar times of day with similar weather conditions. Additionally, the training data is representative of the population of the test data, and labels are obtained from reliable sources for accuracy. We expect that the classification model will generalize well to new, unseen data.

3.2 Data Collection and Integration

Sentinel-1 and Sentinel-2 data were collected from 2021-11-01 to 2022-12-01, using coordinates with 3x3, 5x5, and 7x7 bounding boxes as well as coordinates without bounding boxes, ensuring the most accurate information available for the analysis.

3.3 Data Preparation

We performed data cleaning and exploratory analysis to ensure data quality. For missing values, we calculated statistical measures such as mean, median, variance, max, and min ignoring the missing values. Outliers were detected using the interquartile range (IQR) and replaced with the mean of non-outlier elements to minimize their impact on the model.

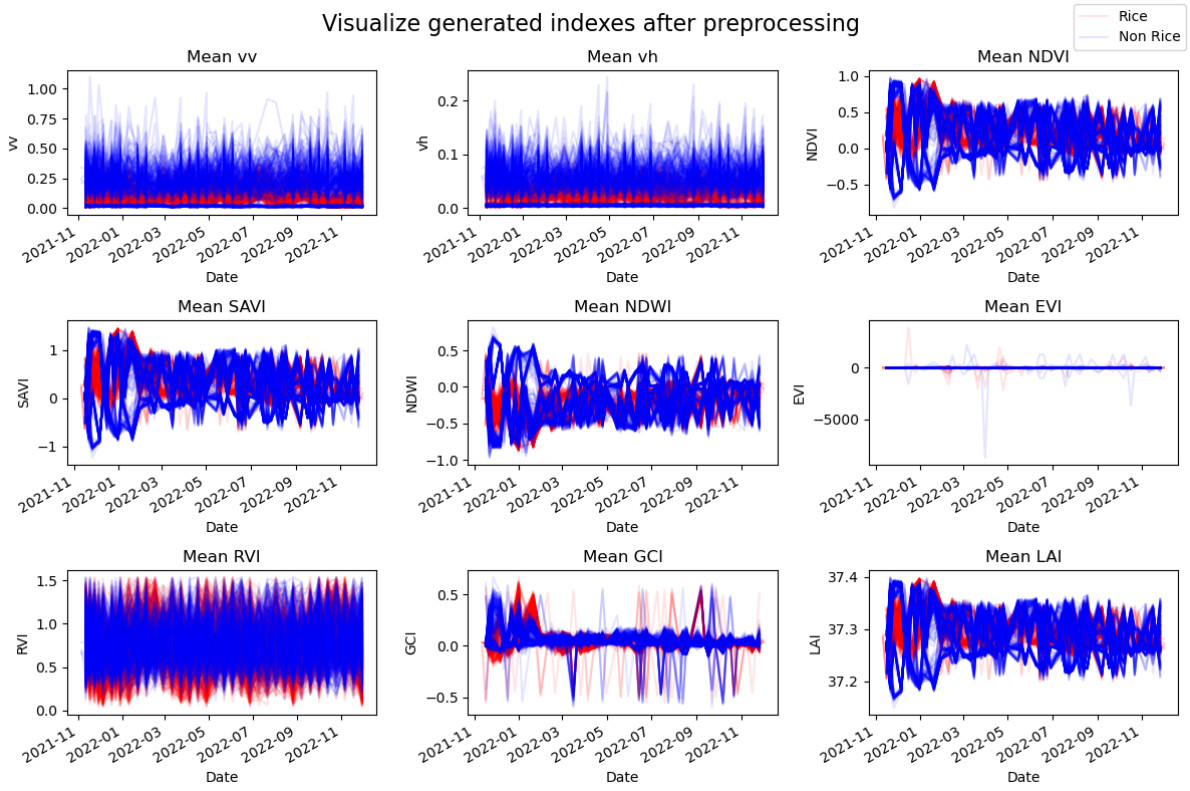

Figure 2: Feature visualization

3.4 Feature Engineering

Feature engineering: To improve the accuracy of our rice crop classification model, we conducted extensive research on representative indices in the field and selected seven features:

- Normalized Difference Vegetation Index (NDVI)
- Soil-Adjusted Vegetation Index (SAVI)
- Normalized Difference Water Index (NDWI)
- Enhanced Vegetation Index (EVI)
- Green Chlorophyll Index (GCI)
- Leaf Area Index (LAI)
- Ratio Vegetation Index (RVI)

After calculating these indices, we applied various statistical aggregation methods, such as mean, median, variance, standard deviation, maximum, and minimum, to each index for every coordinate throughout the year. The resulting data provided valuable insights for the analysis and helped to better understand the underlying patterns in the dataset.

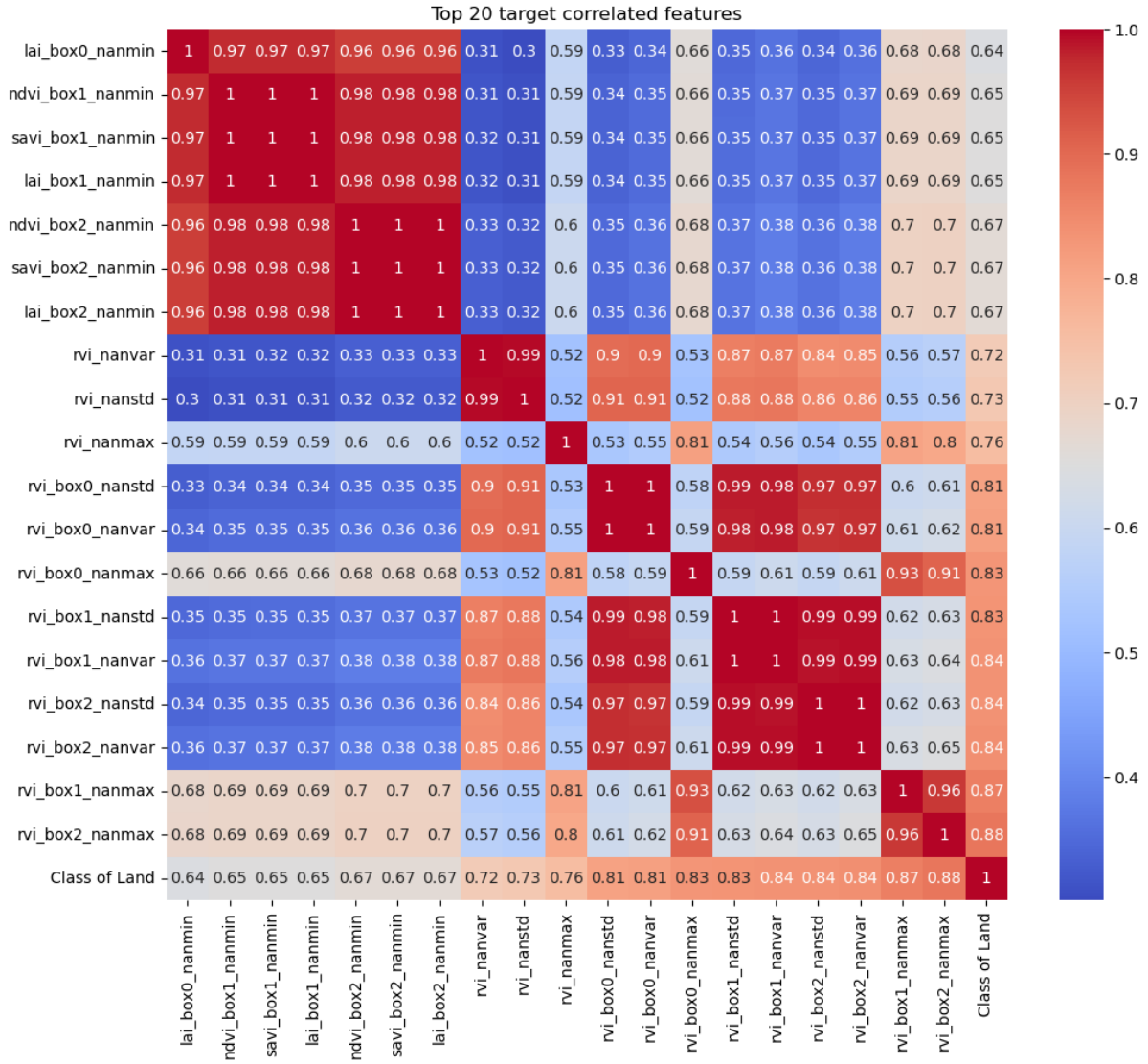

Figure 3: Correlations visualization

3.5 Model Development

Our approach to selecting the best model evolved through several stages, with key decisions made at each point to ensure the best possible performance:

Data Preprocessing approach: To handle the missing data, we utilized a mean imputer, which replaces missing values with the mean value of the corresponding feature. Additionally, we employed anomaly detection to identify and eliminate any outliers in the dataset. Finally, we used an oversampling technique to increase the diversity of the data and improve the performance of our models.

Model validation approach: We employed a test/train split approach, dividing the data into separate 80% training and 20% testing sets. This strategy helped us evaluate the model's performance on unseen data and avoid overfitting.

Model selection approach: We explored various scalers and machine learning algorithms, developing pipelines for each scaler-estimator combination using default hyperparameters. Pipelines

were evaluated using 5-fold cross-validation with F1-score as the performance metric. This led to the selection of *RandomForestClassifier* with a *MinMaxScaler*.

Feature selection approach: We used ANOVA F-value to identify the most relevant features, reducing model complexity and potentially increasing accuracy.

Hyperparameter tuning approach: To further enhance the performance of our chosen model, we conducted a grid search for hyperparameter tuning. We experimented with various parameter combinations, such as the *n_estimators*, *max_depth*, *min_samples_split*, and *max_features*, to identify the optimal configuration for our random forest model.

Model Results: The best model was evaluated using various metrics, such as accuracy, recall, precision, and F1-score, that is trained on the complete dataset, and employed to generate predictions for the provided coordinates. This systematic approach led to exceptional accuracy, further enhanced by hyperparameter tuning and robust feature selection.

4 Innovative Aspects and Breakthroughs

Advanced Data Cleaning and EDA: Rigorous data cleaning and EDA processes, including outlier removal in the VV and VH bands, ensured data quality and reliability, leading to improved model performance. Visualization and correlation analyses helped identify strongly correlated variables and guided model selection to address multicollinearity.

Customized Feature Engineering: Generating a comprehensive set of features, such as multiple vegetation and water indices and their statistics, enabled valuable information extraction from the dataset, resulting in accurate predictions and better model performance.

Anomaly Detection and Oversampling Techniques: Integrating anomaly detection and oversampling methods enhanced model performance by helping the model learn more complex patterns and relationships, leading to better generalization and performance on unseen data.

Comprehensive Model Comparison and Continuous Refinement: A pipeline was developed for each scaler-estimator combination, and each pipeline underwent k-fold cross-validation. Promising combinations were subjected to hyperparameter tuning, enabling the identification of the optimal model. Focusing on continuous improvement and adaptation resulted in a more robust and effective solution, reflected in the improved score.

5 Results and Discussion

5.1 Model Performance

Our optimized random forest model excelled in handling multicollinearity, leading to exceptional performance. By partitioning data and using subsets of features at each split, the model minimized the impact of correlated features on predictions. This approach maximized the use of available information, resulting in a 1.0 accuracy score on unseen data.

The model's parameters were tuned to maximize its performance and generalization capability. Below is a list of the key optimized parameters for our model:

feature_k: 100

feature_score_func: *f_classif* (ANOVA F-value)

model_bootstrap: True

model_max_depth: None

model_max_features: 'sqrt'

model_min_samples_leaf: 1

model_min_samples_split: 2

model_n_estimators: 20

model_n_jobs: -1

model_random_state: 7

5.2 Feature Importance

The top 4 features included RVI indices with 3x3 and 7x7 bounding boxes, and VV and VH. RVI indices measure reflectance differences between red and near-infrared bands, providing insights into vegetation patterns. These features effectively distinguished rice cultivation from non-rice areas by capturing rice crop properties, making them top-performers in our model.

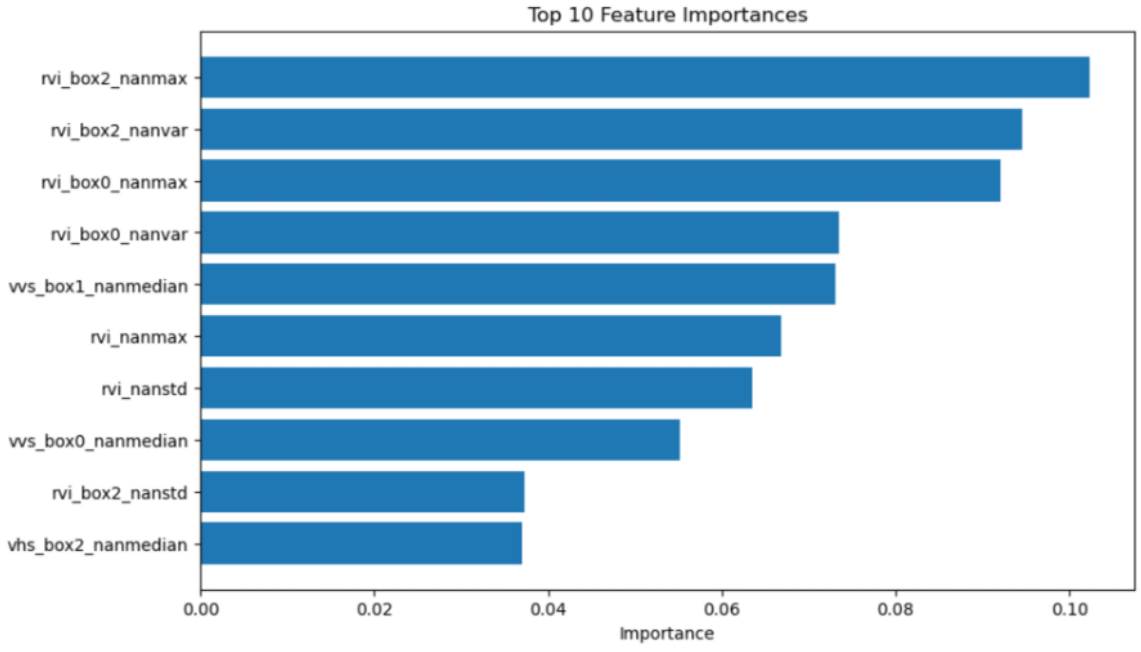

Figure 4: Feature importances.

6 Challenges and Efficiency

6.1 Challenges

Dealing with complex and noisy satellite data posed challenges in identifying outliers and extracting valuable information. The time-consuming process of collecting satellite data also posed a challenge.

6.2 Efficiency Improvements

Parallelization through running multiple Jupyter notebooks simultaneously on the planetary computer can significantly speed up the time-consuming data acquisition process, while the inference runtime for generating results is instant with only 200 data points.

7 Real World Impact

The accurate classification of rice and non-rice areas using remote sensing data can have significant real-world impact, including:

- **Precision agriculture:** Farmers can better manage their crops by identifying areas of high and low productivity, leading to increased crop yield and reduced costs.
- **Scaling the solution:** The solution can be scaled to other parts of the world, particularly in developing countries, to provide accurate data on rice cultivation. This can lead to improved food security, environmental conservation, and economic development.

In summary, the accurate classification of rice and non-rice areas using satellite data has the potential to contribute to more sustainable and efficient agricultural practices, improve food security, and enhance environmental conservation efforts.

8 Conclusion

In summary, this paper presents a novel approach to rice crop classification using data from Sentinel-1 and Sentinel-2 satellites. Our well-designed methodology, which combines advanced data cleaning techniques, customized feature engineering, and a comprehensive comparison of model performances, has enabled us to produce a highly accurate and reliable model for rice crop classification. By addressing the challenges of dealing with complex and noisy satellite data, we have demonstrated the potential of our solution to contribute to more sustainable and efficient agricultural practices, improve food security, and enhance environmental conservation efforts, with the ability to be scaled to other parts of the world.