EY Open Science Data Challenge 2023

Level 1: Rice Crop Discovery Report

— Skyler Aliya

Project Objectives

Introduction:

This report outlines a data science challenge aimed at addressing the UN Sustainable Development Goal 2: Zero Hunger. With over 800 million people going to bed hungry, and climate change being one of the main causes of hunger, understanding crop yields and how to maximize efficiency of crops is an urgent need. This challenge focuses on rice production in Vietnam, which is among the world's leading rice producers and where rice makes up more than half of the total calories consumed. Vietnam is also one of the most vulnerable countries to climate change, losing an estimated US$10b in 2020 alone. Using satellite data from Microsoft's Planetary Computer to predict the presence of rice crops at 250 geo locations (latitude and longitude) in the An Giang province of Vietnam. With rice being a staple food for over four billion people and a livelihood for a fifth of the world's population, solutions focused on this staple are scalable and could transform global rice production.

Goal:

The goal of Level 1 is to develop a rice crop classification model using satellite data. The model will predict the presence of rice crops at a given location and distinguish between rice and non-rice fields. The higher accuracy of the prediction is, the better score will be evaluated on the challenge.

Techniques and Tools

General structure :

To develop an accurate model for predicting the presence of rice crops, I started by utilizing features from the Sentinel-1 Radiometrically Terrain Corrected (RTC) dataset as predictor variables, specifically the VV (Vertical polarization – Vertical polarization) and VH (Vertical polarization – Horizontal polarization) bands. In order to train the model, I extracted the VV and VH band data from the Sentinel-1 dataset for an entire year and summarized using time-series metrics for a given location. The same date as the given latitude and longitude dataset 2022/12/01 was used to ensure consistency.

To explore different approaches for building the model, I experimented with building bounding boxes - 3x3, 5x5 or 7x7 window around the given latitude and longitude positions. I then extracted the aggregated average and median band values to get normalized band values to build the model.

After splitting the dataset into training and testing sets, I applied feature scaling techniques such as Min Max Scaler, Max Absolute Scaling, and Robust Scaling to improve the accuracy of the model. For the model training, I experimented with different algorithms including LogisticRegression, RandomForestClassifier, SVC, MLPClassifier. Through multiple iterations of testing and optimization, I found that Min-Max Scaling and MLPClassifier together provided the highest accuracy for the testing dataset, with an accuracy of 0.86.

To evaluate the performance of the model, I used in-sample evaluation and out-sample evaluation to generate a confusion matrix and gauge the robustness of the model in terms of F-1 score. I then used "plot_confusion_matrix" to visualize the results.

Finally, I used my trained model to make predictions about the presence of rice crops for a set of test coordinates and uploaded the file onto the challenge platform, achieving a score of 0.72. This project demonstrates the effectiveness of using Sentinel-1 data and machine learning techniques to predict the presence of rice crops, which can have significant implications for crop management and food security.

Methods:

Dataset Preparation:

The initial step in this project was to obtain the data from EY, which includes crop location data with geolocation (latitude and longitude) and crop classification as either rice or non-rice fields. To perform the analysis, I chose Python as the primary tool. The predictor variables used in this study are derived from Sentinel-1 radar data, which can penetrate through clouds and provide reliable values with minimal atmospheric attenuation. The VV (gamma naught values of signal transmitted with vertical polarization and received with vertical polarization with radiometric terrain correction applied) and VH (gamma naught values of signal transmitted with vertical polarization and received with horizontal polarization with radiometric terrain correction applied) bands were selected as predictor variables, as they help differentiate between rice and non-rice crops. To obtain the VV and VH band values for a particular location, I utilized the function get_sentinel_data, which fetches the band values over a specified time window. Additionally, I applied various bounding boxes, such as 3x3, 5x5, and 7x7 windows, and extended the time frame to the entire year of 2020 and 2022 to obtain more data for the feature model. Moreover, I employed mathematical combinations of VV and VH data, such as

Normalized Difference Vegetation Index (NDVI) and Radar Vegetation Index (RVI), which can be used as additional features in the model. After obtaining the predictor variables, I combined them with the response variable using the *combine_two_datasets* function, where the concat function from pandas was used to facilitate this process.

Model building:

Next, regression models were trained on the dataset, and the accuracy was calculated to compare the models. I began by applying different feature scalings, including Min Max Scaler, Max Absolute Scaling, Robust Scaling, and Standard Scaler, to produce good results. I then utilized various models, such as Logistic Regression, Random Forest Classifier, SVC, and MLP Classifier, to test and train the dataset. Additionally, I employed Hyperparameter Tuning to determine the optimal combination of hyperparameters for the given dataset, such as the learning rate, number of hidden layers, and number of neurons in each hidden layer, for the MLPClassifier. Finally, Ensemble models were used to combine multiple models, such as a random forest and an SVM, to reduce overfitting and improve the accuracy of the predictions.

Evaluation and Visualization:

After building the model, it is important to evaluate and visualize its performance. In-sample evaluation and out-sample evaluation are commonly used techniques to evaluate the robustness of the model. In-sample evaluation involves testing the model on the same data that was used to train it. This evaluation method may lead to overfitting, which can result in a high accuracy rate but poor performance on new data. Out-sample evaluation, on the other hand, involves testing the model on data that was not used for training. This method can provide a more accurate representation of the model's performance on new data.

To evaluate the performance of our model, I generated a confusion matrix, which is a table that compares the actual and predicted classes. The confusion matrix allows us to calculate performance metrics such as precision, recall, and F-1 score. The F-1 score is the harmonic mean of precision and recall and is often used as an evaluation metric for binary classification problems. To visualize the confusion matrix, I used the "plot_confusion_matrix" function in Python. This function generates a graphical representation of the confusion matrix that makes it easier to interpret and analyze the performance of the model.

Submission:

Once I have built and evaluated the model, I are ready to make predictions on new data. To submit our predictions to the challenge, I used the data from "challenge_1_submission_template.csv" and applied our model to make predictions for the presence of rice crops at specific latitudes and longitudes. I used the "get_sentinel_data" function to extract the VV and VH data for the specified location and time frame, and then used the trained model to predict the class of land as either rice or non-rice. Finally, I exported the predicted classes to a .csv file and submitted it to the challenge platform for evaluation.

Observations and Conclusions

Observations:

During the experimentation process, various techniques such as testing, training, model selection, and optimization were employed to achieve the best possible results. After evaluating the performance of several models, the combination of Min-Max Scaling and MLPClassifier provided the highest accuracy rate, reaching 0.86 on the testing dataset. Min-Max Scaling is a normalization method that scales the values of features between 0 and 1, ensuring that all the features contribute equally to the model's predictions. For example, using the same model with the features NDVI and RVI produces a lower F-1 score and accuracy score, so I dropped those two features in the final model. MLPClassifier is a type of neural network that is often used for classification tasks due to its ability to learn non-linear relationships between features. By using these techniques, the model was able to accurately distinguish between rice and non-rice fields in the An Giang province of Vietnam.

To validate the effectiveness of the model, the predictions were tested on a set of test coordinates, and the results were uploaded to the challenge platform. The model achieved a score of 0.72 on the platform, demonstrating its capability to identify rice crops accurately. Although there was a slight decrease in accuracy compared to the testing dataset, this could be due to the variations in the environmental factors, which could affect the growth of rice crops. Nevertheless, the model's performance was promising, and it could be further improved by incorporating additional features such as weather data, soil conditions, and crop management practices, which could enhance the model's prediction accuracy. Overall, the model presented a viable solution for predicting rice crop classification, which could have practical applications in optimizing rice production and ensuring food security in regions vulnerable to climate change.

Applications:

Looking to the future, the rice crop classification model developed in this challenge could have a significant impact on global rice production. With rice being a staple food for over four billion people and a livelihood for a fifth of the world's population, maximizing the efficiency of rice crops is crucial for ensuring food security and reducing hunger. By accurately predicting the presence of rice crops at specific locations and distinguishing between rice and non-rice fields, this model could help farmers and policymakers make more informed decisions about crop management and resource allocation. In addition, the use of satellite data and machine learning techniques in this challenge could pave the way for more efficient and sustainable agricultural practices, particularly in vulnerable areas affected by climate change. Ultimately, the success of this challenge and the resulting model could have far-reaching implications for global food security and sustainable agriculture.

Conclusions:

In conclusion, this project aimed to develop a rice crop classification model using satellite data to address the UN Sustainable Development Goal 2: Zero Hunger. The model was developed and optimized through a series of testing and training steps, ultimately resulting in a model with a testing dataset accuracy of 0.86. This model was used to make predictions about the presence of rice crops for a set of test coordinates, achieving a score of 0.72 on the challenge platform. While there is still room for improvement, this project represents a step towards addressing the urgent need for understanding crop yields and maximizing efficiency of crops, particularly in vulnerable areas such as Vietnam. The potential impact of solutions focused on rice production is significant, as rice is a staple food for over four billion people and a livelihood for a fifth of the world's population. With continued research and development in this area, we may be able to transform global rice production and make progress towards the goal of zero hunger.

The code of this project is available on Github.