Python solution

Project Report

Classification of Crop Types using Multi-temporal NDVI and Random Forest Classifier

Name: Rezaul Hasan Bhuiyan
Student Id: S3003337
Specialization: Natural Resources Management
Email: r.h.bhuiyan@student.utwente.nl

Table of Contents

1. Introduction

Crop type mapping plays a pivotal role in food security by providing insights into agricultural landscapes. It enables informed decision-making and optimizing resource allocation for sustainable production. Several supervised and unsupervised machine learning techniques have been extensively used in land use mapping studies (Tatsumi et al., 2015). Random Forest, an ensemble classifier, has demonstrated robustness across diverse landscapes, providing high accuracy and efficient processing of large datasets. Spectral bands, more specifically vegetation indices such as NDVI, are sensitive to crop phenology. This makes multi-temporal Normalized Difference Vegetation Index (NDVI) images suitable for crop mapping variables. Python is a powerful scripting language with rich resources to conduct such geospatial analysis efficiently and effectively.

Objectives: The aim of this project is to achieve following objectives through python scripting:

- Classify different crop types from multi-temporal NDVI (sentinel-derived) using the Random Forest classifier.
- Assess the sensitivity of hyperparameter combinations with respect to accuracy and out of bag error (OOB)
- Assess the performance of optimized Random Forest classifier.
- Identify important features for crop mapping.

2. Dataset and Method:

2.1 Dataset

1. NDVI raster:

The NDVI raster comprises 12 layers representing the NDVI values for each month of 2019 in Noord Beveland, Netherlands. Each layer has a resolution of 10 meters. The 10 m resolution allowed the classification of small agricultural fields. Additionally, this resolution makes the study applicable to tropical and subtropical regions where smaller Agro-land parcel system is dominant (Zeng et al., 2015). The layers are not stacked in monthly serial. The chronology as follows:

Band 1: November, Band 2: May, Band 3: February, Band 4: April Band 5: January' Band 6: September, Band 7: July, Band 8: August, Band 9: June, Band 10: March, Band 11: December, Band 12: October

2. Sampling Points:

The crop data for labelling were sourced from the Base Registration Crop Parcels agency in the Netherlands. 929 sampling points were collected. The numbers of samples available for each crop type are as follows: 353 for cereals, 193 for onions, 144 for potatoes, 110 for sugar beet, 76 for lucerne, 31

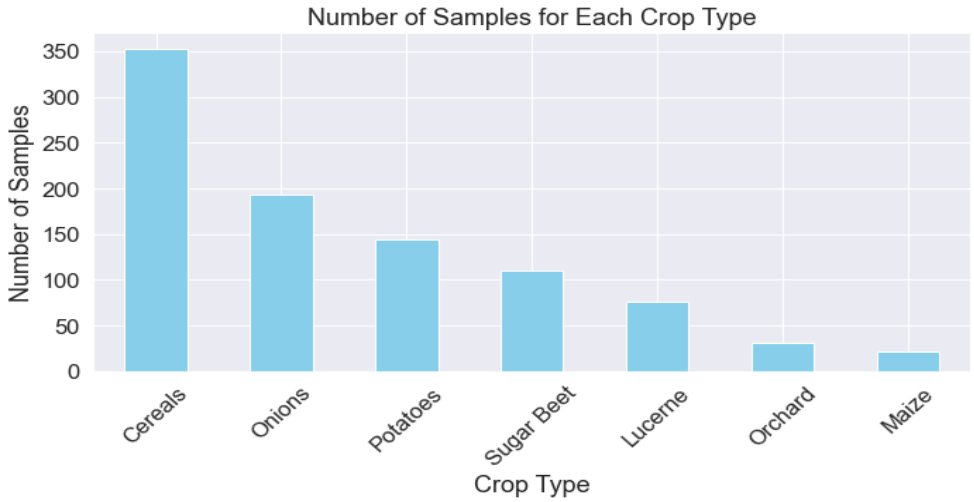

Figure 1 Number of samples for each crop type.

for orchard crops, and 22 for maize (figure1).

2.2 Project Workflow:

The workflow outlines how the project goals were achieved throughout the study.

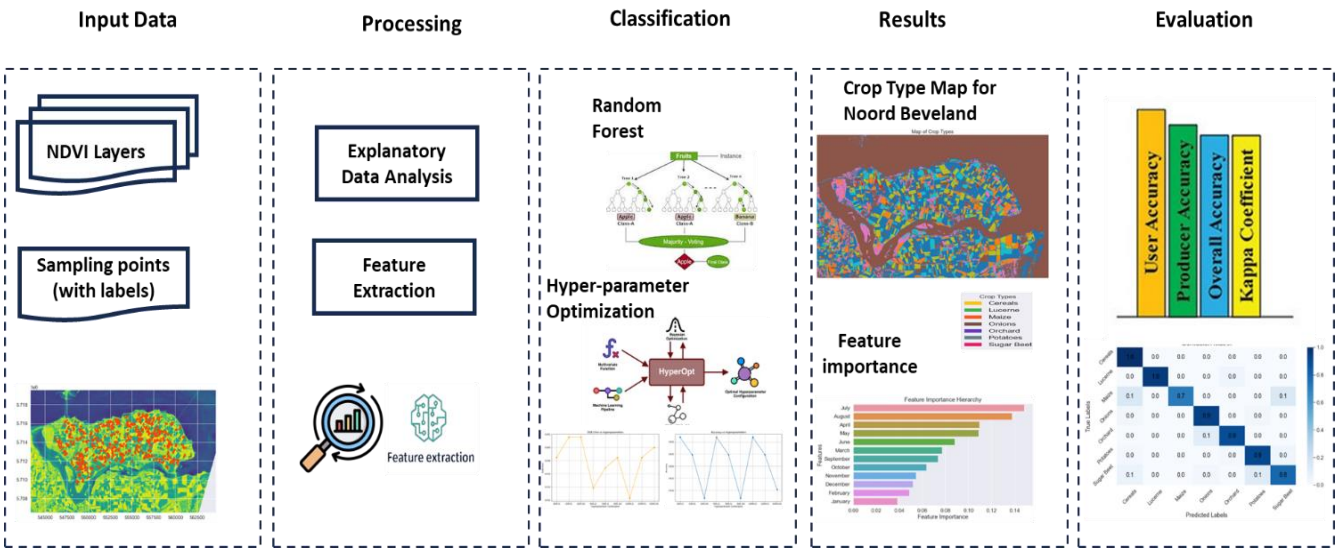

Figure 2 Workflow of the project.

3. Result:

3.1 Crop type classification: The agricultural fields of Noord Beveland were classified into cereals, lucerne, maize, onions, orchard, potatoes, and sugar beet classes (see figure 3). It is evident from the figure that a significant portion of the map is covered by onion fields. However, water bodies were misclassified as onion fields. This occurred because there were no training points made available for water bodies. Additionally, since testing dataset did not have water class, it didn't affect the model accuracy report. The classified map has overall accuracy of 93.18%.

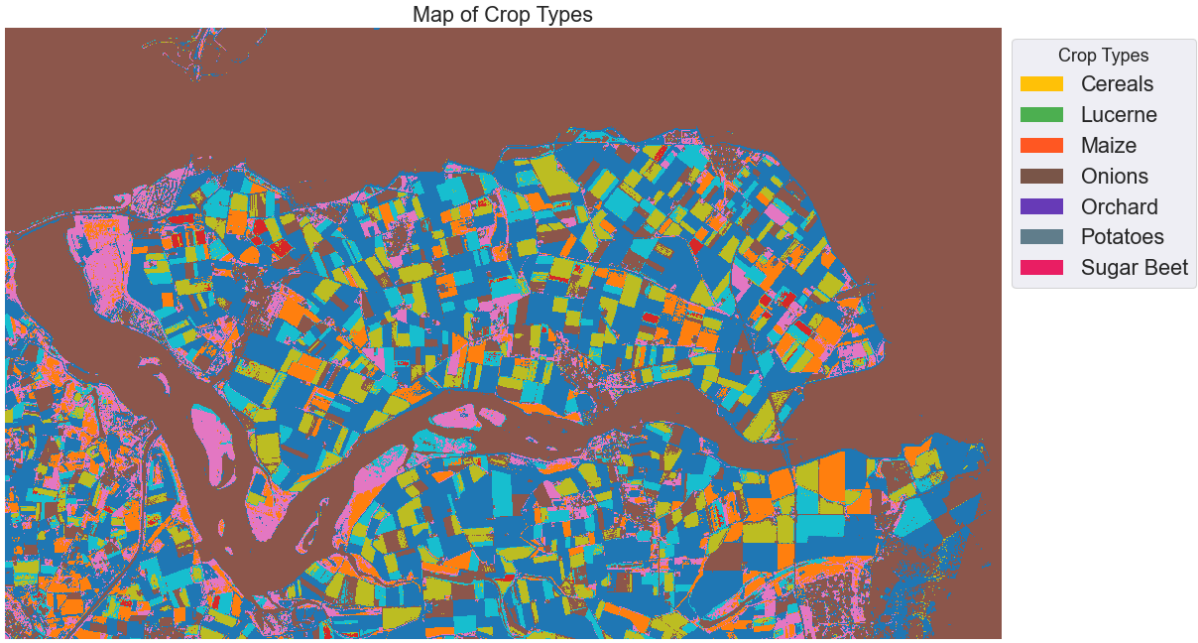

Figure 3 Crop type map of Noord Beveland

3.2 Hyperparameter optimization:

The Random Forest model consisted of two hyperparameters: the number of trees (n_tree) and the number of variables used for splitting (mtry). The number of trees (n_tree) in a Random Forest model

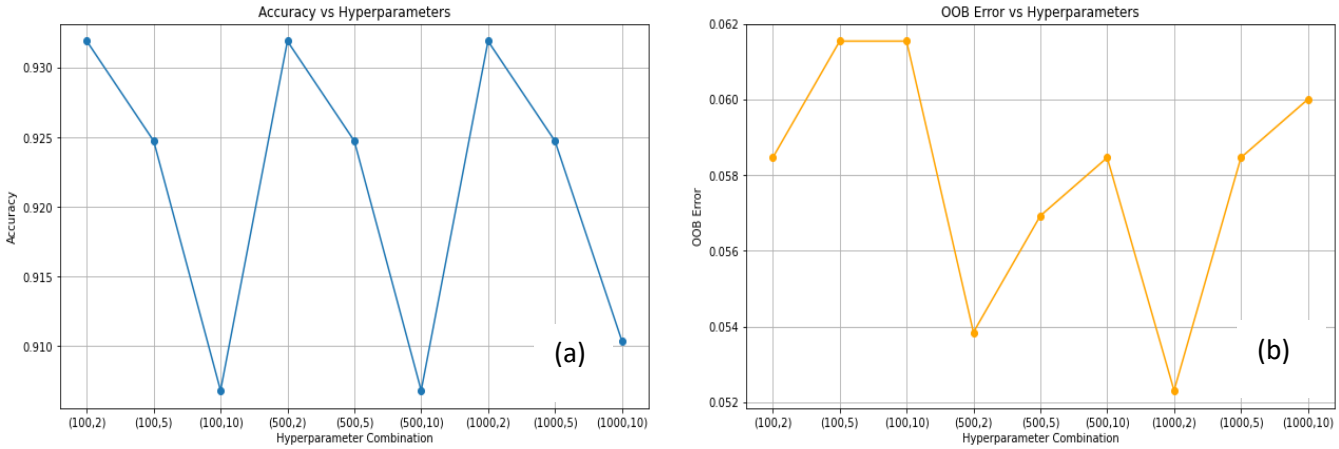

Figure 4 (a) Accuracy of the model for different hyperparameter combinations, (b) OOB of the model for different hyperparameter combinations.

defines the quantity of decision trees that will be grown during the training process. Meanwhile, the number of variables used for splitting (mtry) determines how many predictors are randomly selected at each split in each tree. This study explored the Cartesian combination of n_tree and mtry sets, predicting the map accuracy and out-of-bag error (OOB). Out-of-bag error (OOB) is an estimate of the model's performance on unseen data, calculated by evaluating the model on instances not included in the bootstrap sample used to train each tree. After analysis, it was determined that [500, 2] was the optimum combination for this dataset. While other combinations may have offered high accuracy, they often exhibited high OOB error, and vice versa (see Figure 4). Therefore, we selected the balanced combination for further analysis.

3.3 Accuracy assessment:

With the optimized hyperparameter setting, the classified map of this study achieved an overall accuracy of 93.18%. Notably, the model demonstrated an exceptional accuracy of 100% for cereals and Lucerne crop type. However, maize exhibited the lowest accuracy among the classes. This is due to the small number of samplings for Maize class (Figure 1). Additionally, the model achieved accuracies ranging from 80% to 90% for other classes such as orchard, potatoes, and sugar beet. To provide further insights into the model's performance, a confusion matrix (Figure 5) was created to depict the classification results.

Further analysis revealed the following classification metrics:

Overall Producer Accuracy: 0.88

Overall User Accuracy: 0.93
Kappa Coefficient: 0.88

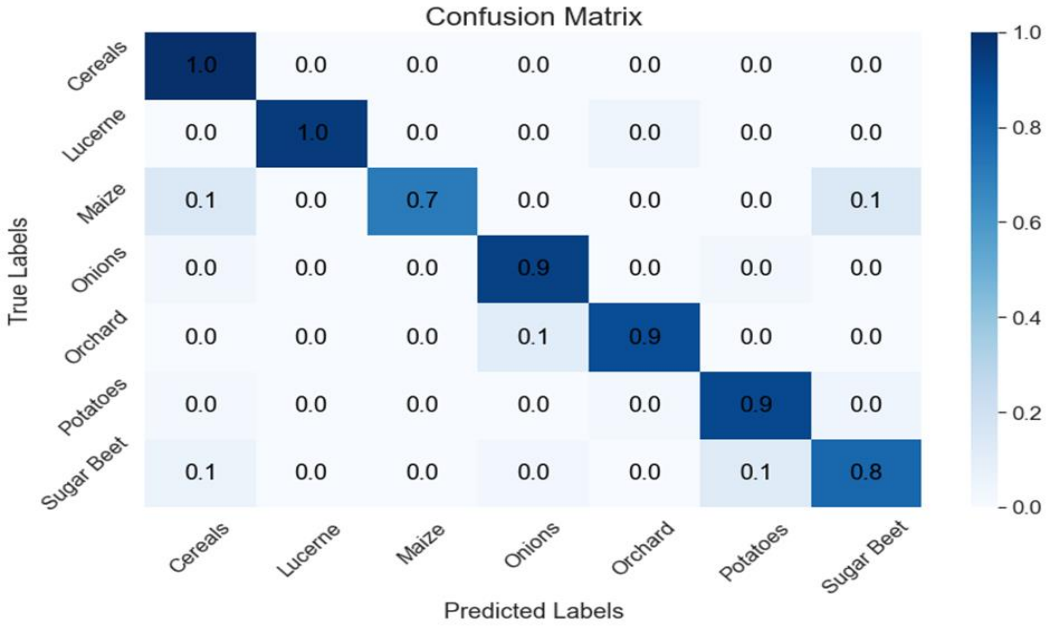

Figure 5 Confusion matrix for the optimized model.

3.4 Feature Importance Analysis:

To identify the important features of the model, feature importance analysis was conducted in this study. The results (Figure 6) indicate that the Normalized Difference Vegetation Index (NDVI) of July has the highest influence on the prediction, followed by August, April, and May. Conversely, January has the lowest impact on the model's predictions. Additionally, the NDVI of other months of the year has feature importance ranging between 0.05 and 0.08.

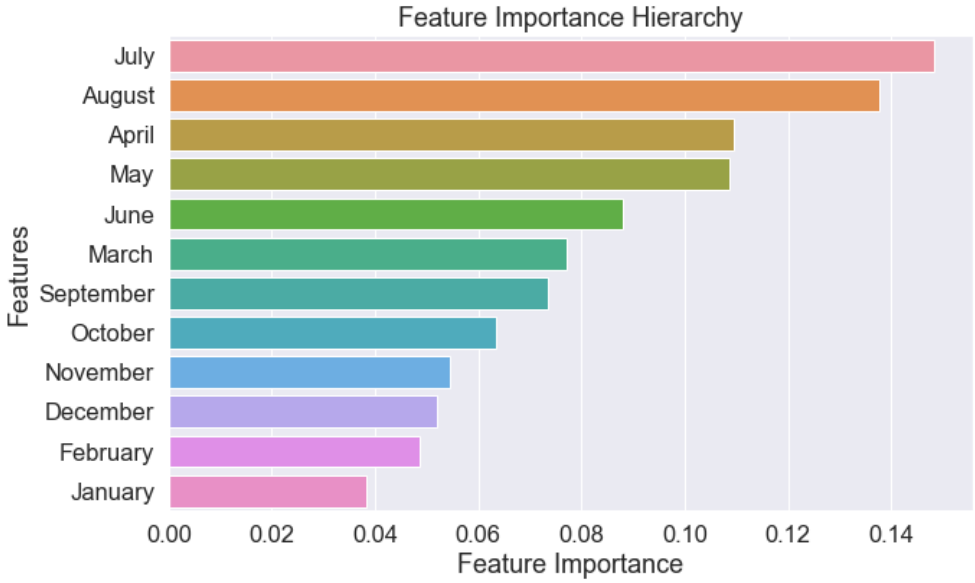

Figure 6 Feature importance hierarchy of model input.

4 References

Zheng, B., Myint, S. W., Thenkabail, P. S., & Aggarwal, R. M. (2015). A support vector machine to identify irrigated crop types using time-series Landsat NDVI data. International Journal of Applied Earth Observation and Geoinformation, 34, 103-112.

Tatsumi, K., Yamashiki, Y., Torres, M. A. C., & Taipe, C. L. R. (2015). Crop classification of upland fields using Random Forest of time-series Landsat 7 ETM+ data. Computers and Electronics in Agriculture, 115, 171-179.