Avlonitis Ektor

Gaitanis Pavlos

Kalogeris Spyros

Vavakas Alex

Satellite-based
Rice Yield Forecasting

EY – Level 2 Challenge

Table of Contents

1. Abstract

In this report, we present the results of our work in building a machine learning model to estimate rice crop yield using satellite data in the An Giang province of Vietnam. The primary goal of this study was to leverage the European Copernicus (Sentinel-1) program and NASA's Landsat program to track the location and growth of rice crops, and subsequently predict their productivity. Our machine learning model demonstrated promising results in estimating rice crop yield, with precision, recall, accuracy, and R2 scores serving as measures of success. By integrating satellite data, weather data and advanced analytics, we achieved a deeper understanding of rice crop phenology, providing essential information for agricultural decision-making, resource allocation, and policy formulation.

2. Introduction

a. Problem statement

i. What is a high-level description of the problem you are addressing?

The high-level description of the problem to create a machine learning model that can calculate the rice crop production for a specific site. The objective is to follow the position and growth (phenology) of rice crops using satellite data from the European Copernicus (Sentinel-1) program and NASA's Landsat program. Understanding the growth and productivity (yield) of rice fields in the An Giang province of the Mekong Delta, Vietnam, requires combining these satellite datasets, which include optical data from Landsat and radar data from Sentinel-1.

The features include the district where the crops are located, latitude and longitude coordinates of the fields, the season of cultivation (Summer-Autumn or Winter-Spring), the intensity of rice cropping (Double or Triple cropping), the date of harvest, the field size in hectares, and the rice yield in kilograms per hectare. The goal is to make predictions about their productivity or yield based on these features.

ii. How many data you have in total?

The dataset provided for this problem consists of 557 rows in a CSV file. Each row represents a data point containing information about rice crop yield for a given location. The dataset includes data for the period of late-2021 to mid-2022, focusing on the Winter-Spring 2021-2022 season (November to April) and the Summer-Autumn 2022 season (April to August). The data is specific to the Chau Phu, Chau Thanh, and Thoai Son districts in the An Giang province of Vietnam, which is known for its dense rice crop region with a mixture of double and triple cropping cycles. The dataset provides valuable information for participants to analyze and build machine learning models that estimate rice crop yield based on the available data points.

iii. What is your high-level approach ML?

The high-level approach used for the model is supervised machine learning. The use of an Extra Trees Classifier indicates that the model is trained with labeled data, where the input features are known, and the corresponding rice crop yield is provided as the target variable.

iv. What is the measure of success/ progress?

The measures of success/progress used are precision, recall, accuracy, and R2 (coefficient of determination). These measures of success/progress help evaluate and assess the performance and effectiveness of the model in estimating rice crop yield based on the provided data and chosen Extra Trees Classifier.

b. Motivation for the work or Context of problem or Importance of the problem

The difficulty of the problem focuses on calculating rice crop production using satellite data, which is crucial for the agriculture industry and food security worldwide. For a sizable section of the world's population, rice is a staple crop and an essential source of nutrition.

The allocation of resources, supply chain management, and agricultural policy-making all depend heavily on accurate crop yield estimation. It is now possible to acquire insights into the development and productivity of rice crops on a bigger scale, enabling better decision-making processes, by utilizing satellite data and machine learning techniques.

Using satellite data analysis to better understand rice crop yields can help reduce food shortages, improve farming methods, and solve difficulties with agricultural sustainability. The work completed for this challenge has the potential to significantly improve food security, reduce global hunger, and advance our understanding of the use of remote sensing technology for agricultural monitoring and management.

4. Data Exploration

The dataset contains information related to rice crop characteristics in three different districts. Each row represents a specific field in a district and provides the following information:

- District: The name of the district where the field is located (Chau_Phu, Chau_Thanh, Thoai_Son) (Figure 1).
- Latitude: The latitude coordinate of the field.
- Longitude: The longitude coordinate of the field.

- Season: Indicates the season of the rice crop. "SA" represents Summer Autumn, and "WS" represents Winter Spring.
- Rice Crop Intensity: Indicates the crop intensity of the rice field. "D" stands for Double crop intensity, and "T" stands for Triple crop intensity.
- Date of Harvest: The date when the rice crop was harvested, represented in the format dd-mm-yyyy.
- Field size (ha): The size of the field in hectares.
- Rice Yield (kg/ha): The yield of rice per hectare in kilograms (target variable).

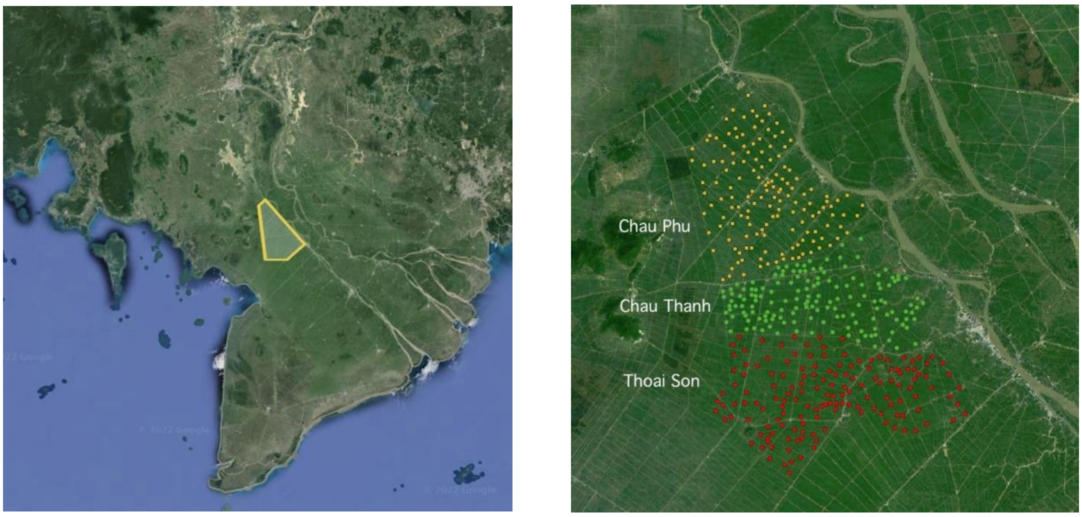


Figure 1. Three Vietnam districts (Chau_Phu, Chau_Thanh, Thoai_Son) that rice crop data is available.

For this challenge, all of the training and test data assume triple cropping (3 cycles per year) but focus on 2 cycles: the Winter-Spring 2021-2022 season (November to April) and the Summer-Autumn 2022 season (April to August). (Figure 2)

Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Cycle #1 - Winter-Spring													
					Cycle #2 - Summer-Autumn								
								Cycle #3 - Autumn-Winter					

Figure 2. Many rice crops in the An Giang province of Vietnam have 3 growth cycles per year. The data provided for this challenge is focused on the 1st and 2nd cropping cycle.

The typical growing stages of rice are shown in the figure below. It is possible to use optical and radar data to track these growing stages over time. Optical data will tell us about the "greenness" of the plant and radar data will tell us about the "structure" of the plant. For example, peak "greenness" occurs before full plant maturity as the rice grain is formed prior to harvest. In the case of "structure", which can be measured with radar data, we are able to see differences in scattering at the various growth stages. Early stages might see less scattering due to reflections from background soil or flooded fields. The peak flowering stage may see maximized scattering due to dense foliage, whereas the ripening stage prior to harvest may see a drop in scattering due to rice tassel formation and "layover" of the plant. In the end, there will be differences between optical and radar phenology, but it is still possible to relate this data to the growth stages and build a good yield model.

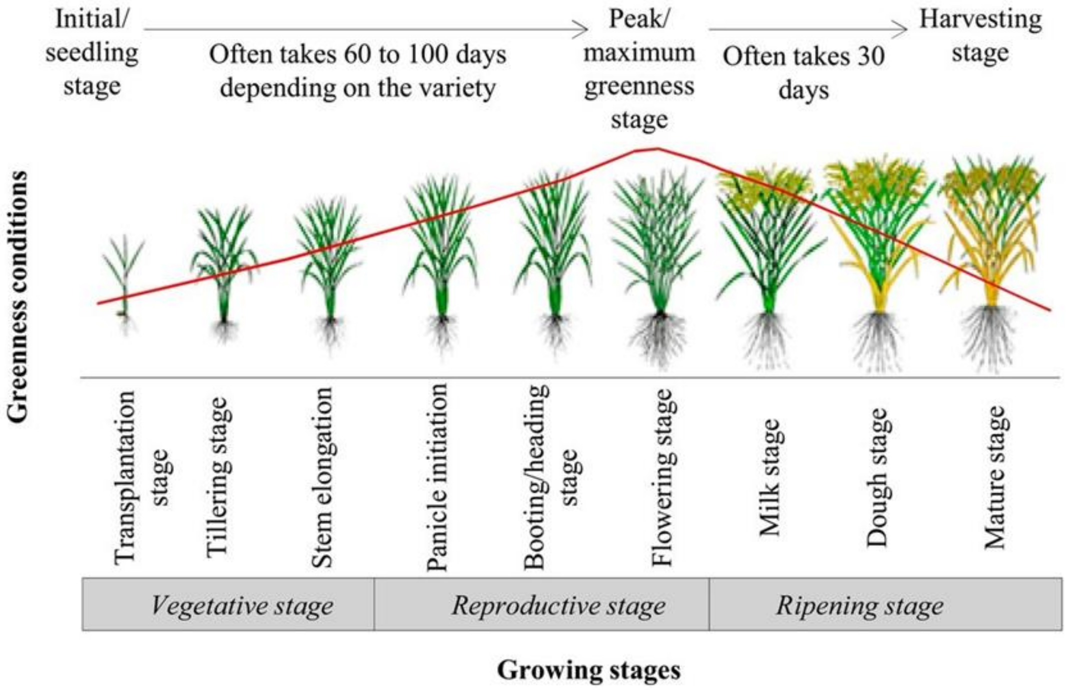


Figure 3. Plant Growing Stages

Sentinel – 1 Radar Data

The Sentinel-1 missions run at a frequency in the C-band, which corresponds to a wavelength of 5.6 cm. Either horizontal (H) or vertical (V) polarizations can be used to broadcast and receive the radar signal. The two primary polarization "bands" in the Sentinel-1 data are VV and VH, where the first letter denotes the transmitted polarization and the second, the received polarization. At any location, the VV and VH bands provide us with surface backscatter. This backscatter tells us about the "structure" of the crop, such as its growth progress from small plants to large plants, and then to bare soil after harvesting.

The Sentinel-1 satellite provides VV (Vertical Vertical) and VH (Vertical Horizontal) data for a range of dates rather than offering data for specific individual dates. It is important to note that not all dates within the desired range may have available data. The availability of data depends on various factors such as satellite imaging schedules, weather conditions, and the specific region of interest. Therefore, while requesting data from Sentinel-1, it is necessary to specify a date range and understand that data may only be available for a subset of those dates. That's why we request ranges of data based on the season we require.

We can also visualize the data. For Seninel-1 RTC, this produces a false-color composite from a combination of the VV and VH bands, as it can be seen below.

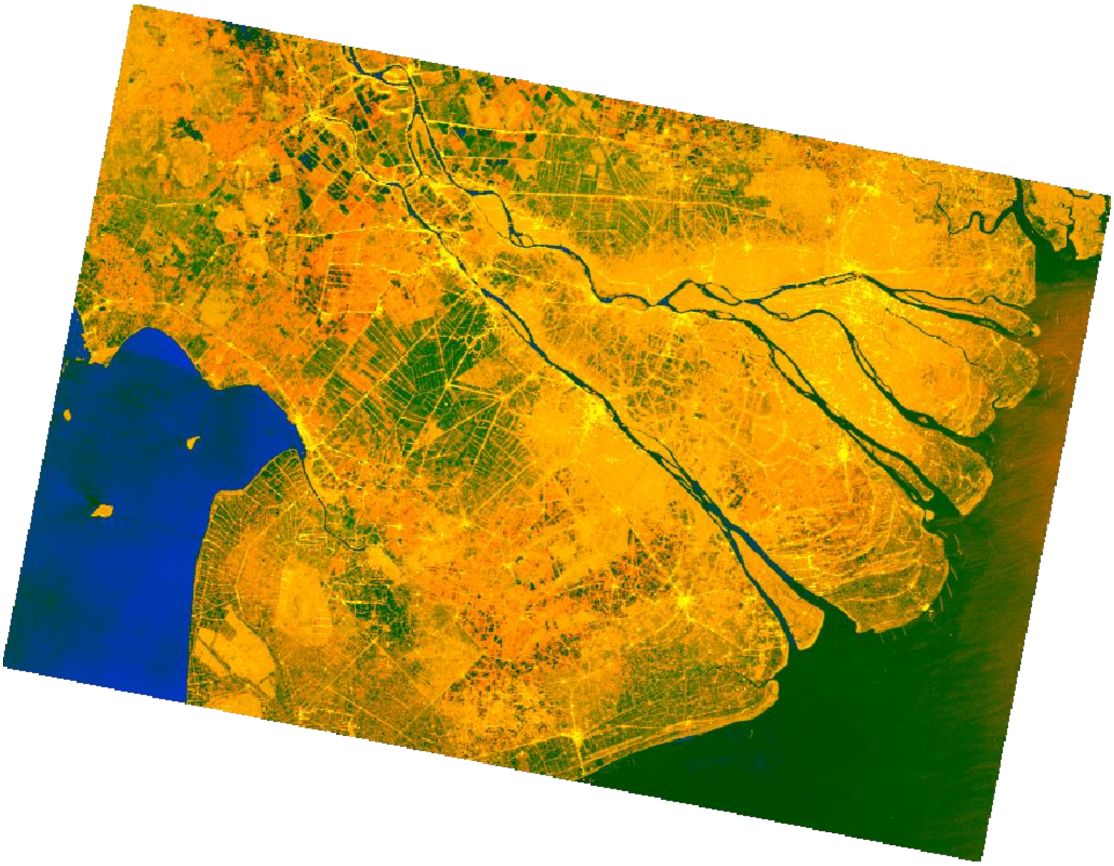


Figure 4. Example of Sentinel-1 RTC photo data.

The VV band backscatter tends to peak at the start of January and May, while the VH band backscatter tends to peak at the start of December and April, as observed in the figures below. The VV band backscatter tends to peak at the start of January and May due to the specific characteristics of radar waves interacting with the vegetation and soil during those periods. This can be influenced by factors such as changes in vegetation growth, moisture content, and the overall agricultural cycle. Similarly, the VH band backscatter tends to peak at the start of December and April for similar reasons.

It is important to note that the peaks in backscatter intensity do not correspond to the date of harvest. We don't observe a pattern between the values VV and VH and the growing stage of the plant. The timing of the harvest can vary depending on various factors such as crop type, geographical location, climate conditions, and farming practices. While mid-April and July are often associated with harvest in certain regions, the specific dates can differ.

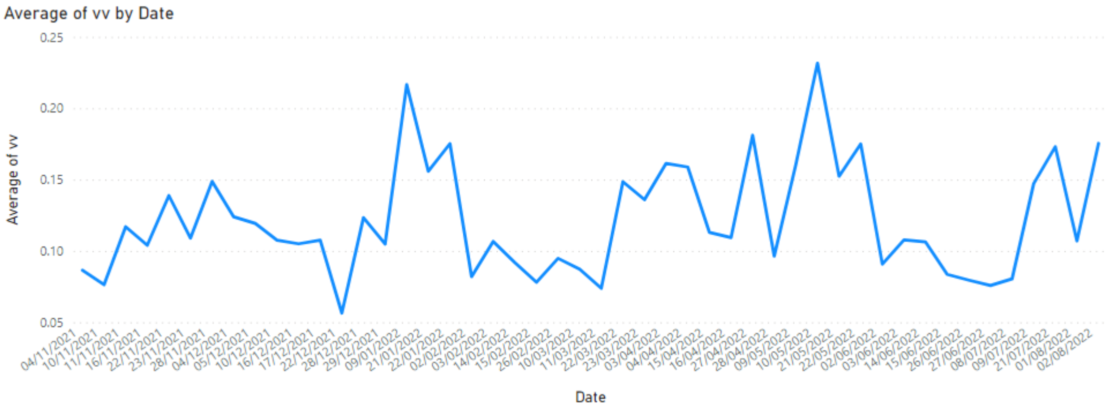


Figure 5. Average VV value by date.

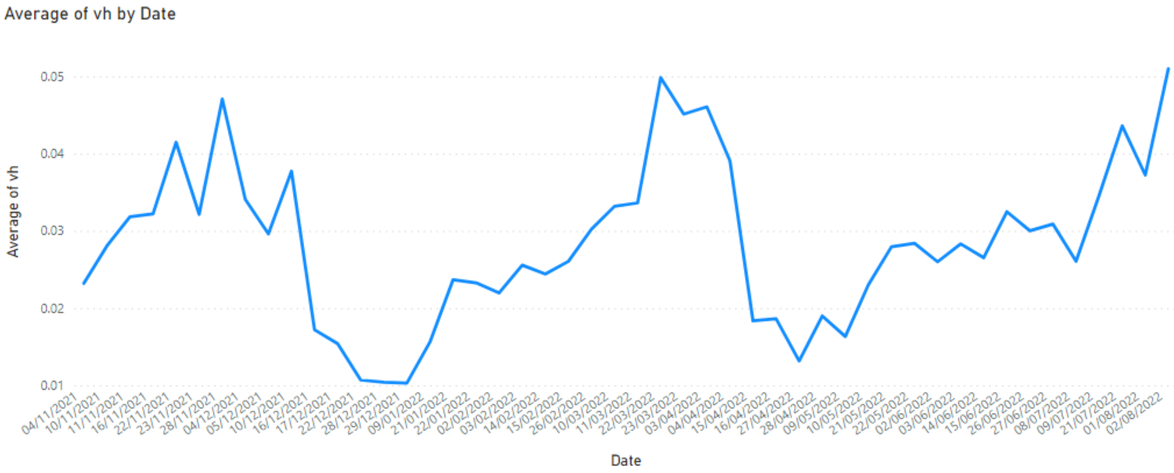


Figure 6. Average VH value by date.

For each row in the given CSV file, representing a specific longitude and latitude, a list of VV and VH data is obtained for the corresponding dates within the range of the season specified in the row. It is important to note that the availability of VV and VH data is limited to the dates within the specific season. To improve the accuracy of the model, certain assumptions and adjustments were made based on the available data within the season range.

Firstly, we keep in mind that the Summer-Autumn season follows the Winter-Spring season. Also, for most of the specific locations of the CSV, there is data for both of these seasons. The following adjustments concern these data. If the row belongs to the WS season (Winter-Spring), the collected data ranges from the 1st of November till the date of harvest for that cycle. However, if the row belongs to the SA season (Summer-Autumn), the collected data ranges from either the 1st of April or the next day of the previous cycle's date of harvest (if that day is after the 1st of April) to the date of harvest for that cycle (SA). The reason for doing that is the noticeable difference between the values of VV and VH before and after the day of harvest. If for a location, there is data for only one season, we collect the VV and VH lists without making these assumptions.

The generated lists contain values corresponding to each date for which data is available from Sentinel-1. These lists capture the VV and VH backscatter data obtained for the specific dates, and thus several statistical combinations were created based on the available data points.

- Maximum (max) and minimum (min) values of VV and VH: These provide information about the highest and lowest backscatter values observed at a specific location over the available dates of the season (SA or WS).
- Range of VV and VH: We calculate the difference between the maximum and minimum backscatter values observed at a specific location over time. It provides information about the variability of backscatter values at that location.
- Mean and standard deviation (std) of VV and VH: These provide information about the average and spread of backscatter values at a specific location over time.
- Ratio of VV to VH: This provides information about the relative strength of the backscatter signals in the two polarizations. This can be useful in distinguishing between different types of crops or vegetation.

RVI

The Radar Vegetation Index (RVI) is a vegetation index derived from Sentinel-1 radar satellite data. It is calculated using the backscatter values obtained from the radar signals. The RVI is used to assess vegetation density and health by measuring the radar signal response from vegetation cover. It is one of the most common indices which tends to mimic the properties of optical NDVI, which will be discussed on the next satellite data. A typical equation for RVI is shown below, but there are other variations of this index used by researchers. In this one the index is defined by a square root scaling for a better dynamic range of the RVI index. This root is multiplied with the vegetation depolarization power fraction as seen below.

RVI (Radar Vegetation Index) = sqrt (1- VV / (VV+VH)) * 4 * (VH / (VV + VH))

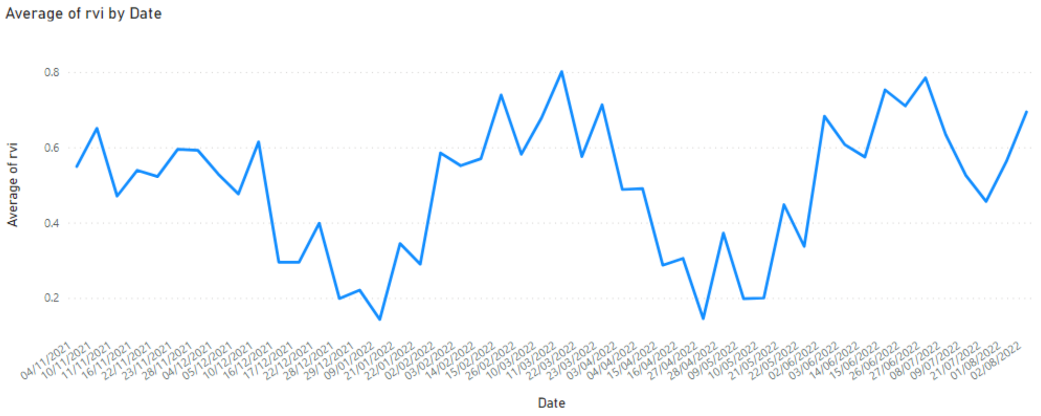


Figure 7. Average RVI value by date.

In the figure above, we can observe that the RVI (Radar Vegetation Index) reaches its peak at mid-March, which is before the harvest season for winter spring crops. This peak in the RVI indicates a dense foliage and active vegetation growth during the peak flowering stage. Similarly, for the summer autumn season, we can observe that the peak in RVI occurs at the beginning of July, which is before the harvesting season that typically takes place later in July. This peak in RVI suggests dense foliage and active growth during the peak flowering stage for the summer autumn crops.

Landsat – Optical Data

Landsat data played a pivotal role in the 2023 EY Open Science Data Challenge, where it served as a valuable resource for agricultural monitoring and crop yield estimation. With its unparalleled capabilities in capturing high-resolution, multispectral images of the Earth's surface, Landsat provided critical insights into the dynamics of agricultural landscapes.

In the challenge, we harnessed Landsat data to assess crop health, identify different crop types, and mainly monitor changes in vegetation over time. Leveraging this satellite's multispectral bands, including Red, Green, Blue, Near-Infrared (NIR), and Short-Wave Infrared (SWIR), we aimed to build comprehensive machine learning models. We dig into our approach, data pre-processing steps, feature selection, and transformation techniques. Additionally, we discuss the methodologies used to develop accurate prediction models, including the application of popular indices such as the Normalized Difference Vegetation Index (NDVI).

Leveraging the powerful PySTAC library and the Planetary Computer API, we gained access to a vast array of Earth observation data, enabling us to delve into the agricultural landscape of Vietnam.

Landsat provides a wealth of valuable information through various spectral bands, capturing the reflected light from the Earth's surface. These satellites observe the study area at regular intervals, allowing for frequent monitoring of rice crops' growth stages.

To derive more actionable insights, we integrated Landsat imagery with other critical bands such as "nir08," "red," "green," "blue," "qa_pixel," and "lwir11." However, the challenge lay in matching multiple Landsat images to create a unified and consistent dataset for analysis.

NDVI

The different spectral bands provided by these satellites are sensitive to various properties of vegetation, such as chlorophyll content, water content, and biomass. Researchers often use statistical combinations of these bands, known as indices, to build models for rice crop growth stages and yield estimation. The Normalized Difference Vegetation Index (NDVI) is one of the most commonly used indices for agriculture, but other indices such as Enhanced Vegetation Index (EVI) and Soil Adjusted Vegetation Index (SAVI) also play essential roles in tracking crop health and growth patterns. The most used index for tracking rice crop growth stages is the Normalized Difference Vegetation Index (NDVI). NDVI is calculated using the near-infrared (NIR) and red bands and provides a measure of vegetation greenness and vigor. During the rice crop lifecycle, NDVI values vary, reflecting changes in plant health and growth.

In the early growth stages, when rice plants are small and sparse, NDVI values are relatively low. As the crop develops and foliage becomes denser, NDVI values increase, indicating healthier and more vigorous vegetation. The peak NDVI value typically corresponds to the crop's maximum greenness, signifying the critical growth stage for optimal yield estimation. Subsequently, NDVI values decline as the crop matures and approaches the harvest stage.

NDVI (Normalized Difference Vegetation Index) = (NIR-Red) / (NIR+Red)

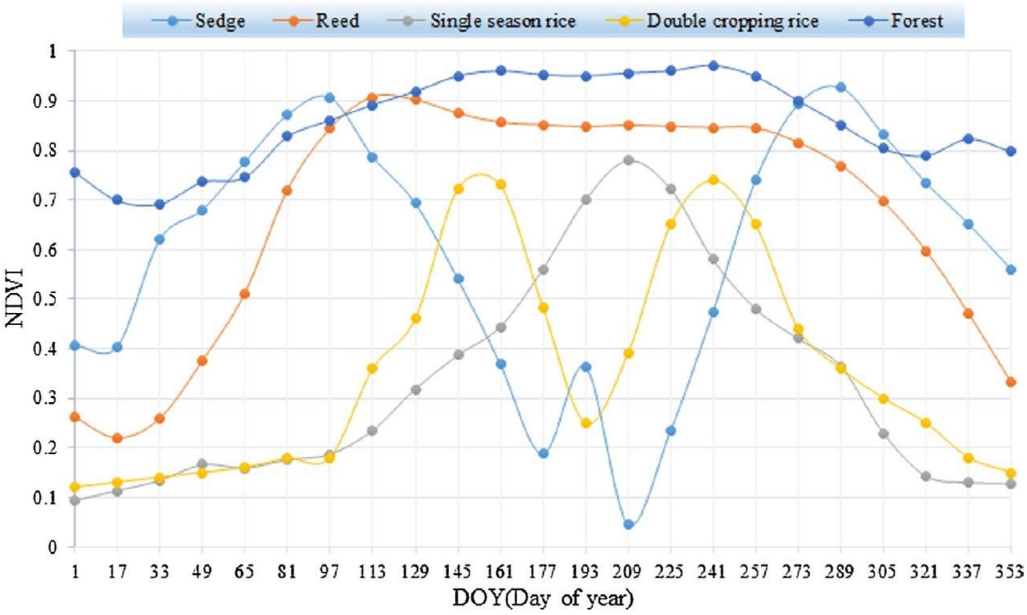


***Figure 3**. Sentinel-2 NDVI varies considerably for different crop types. Forests are stable and have high NDVI values. In the case of rice, it is easy to see its variability for double cropping cycles (2 per year, yellow line) and single cropping cycle (1 per year, grey line).*

Near-infrared (NIR) and red bands are specific spectral bands of electromagnetic radiation that are commonly used in remote sensing and satellite imaging to study vegetation and various Earth surface features.

Near-Infrared (NIR) Band

The Near-Infrared band is a region of the electromagnetic spectrum that lies adjacent to the visible red band. It ranges from approximately 700 nanometers (nm) to 1,300 nm in wavelength. NIR radiation is just beyond the red end of the visible spectrum, making it invisible to the human eye. Despite being invisible, NIR radiation interacts differently with different types of surfaces, including vegetation.

In the context of remote sensing, the NIR band is particularly valuable for studying vegetation health and density. Healthy vegetation strongly reflects NIR radiation due to the high chlorophyll content, while stressed or sparse vegetation reflects less NIR radiation. This property allows researchers and data analysts to calculate vegetation indices, such as the

Normalized Difference Vegetation Index (NDVI), which compares the reflectance of red and NIR bands to quantify the "greenness" and health of vegetation.

Red Band

The red band is a part of the visible spectrum, which humans can perceive. It spans approximately 620 nm to 750 nm in wavelength. When it comes to satellite imaging and remote sensing, the red band is often used in conjunction with the NIR band for vegetation analysis.

Similar to the NIR band, healthy vegetation also reflects red light, although to a lesser extent. Chlorophyll absorbs red light for photosynthesis, causing plants to appear green to our eyes. As vegetation density and health change, the amount of red light reflected from the surface also changes, making the red band a valuable component for vegetation-related studies.

Challenges & Cloud Filtering

Optical data, such as Landsat, provides essential spectral bands that researchers use to build models. However, one significant challenge with optical data is its vulnerability to cloud cover, especially in regions like Vietnam, where clouds persist over one location about two-thirds of the time, making only one-third of the data available in a given year. Hence, filtering the data to remove clouds is crucial for accurate analysis.

The heart of our analysis was ensuring the reliability and quality of the satellite data. The presence of cloud cover presented a significant challenge, affecting the spectral information of the imagery. Our solution entailed applying stringent data filtering techniques based on the "eo:cloud_cover" property, where scenes with cloud cover exceeding 10% were eliminated. By doing so, we maintained data integrity and minimized the influence of atmospheric conditions.

The challenge we encountered related to cloud cover led to missing values in the Normalized Difference Vegetation Index (NDVI) data. To address this issue and ensure the continuity of our analysis, we implemented a strategy of using the median value of NDVI in locations where cloud cover obscured the data. By choosing the median value from nearby areas, we leveraged the assumption that locations in close proximity tend to exhibit similar phenological patterns, especially when dealing with agriculture crops like rice within a relatively confined region. Using the median value from neighboring locations for imputation helps minimize the potential impact of cloud contamination on the analysis, as clouds often affect adjacent areas similarly. This approach enables us to maintain the spatial coherence of NDVI values while still addressing the issue of missing data, ultimately enhancing the reliability of the phenology analysis. Additionally, we preserved the spatial structure of the dataset and avoided introducing artificial fluctuations in the NDVI time series. Employing this approach, we effectively filled in the gaps caused by cloud-contaminated observations, allowing us to maintain a more complete and consistent time-series dataset for our analysis. The use of the median value in place of missing NDVI values allowed us to derive meaningful insights and accurately track rice crop phenology, contributing to the development of robust yield forecasting models, as it now benefits from a completer and more consistent dataset.

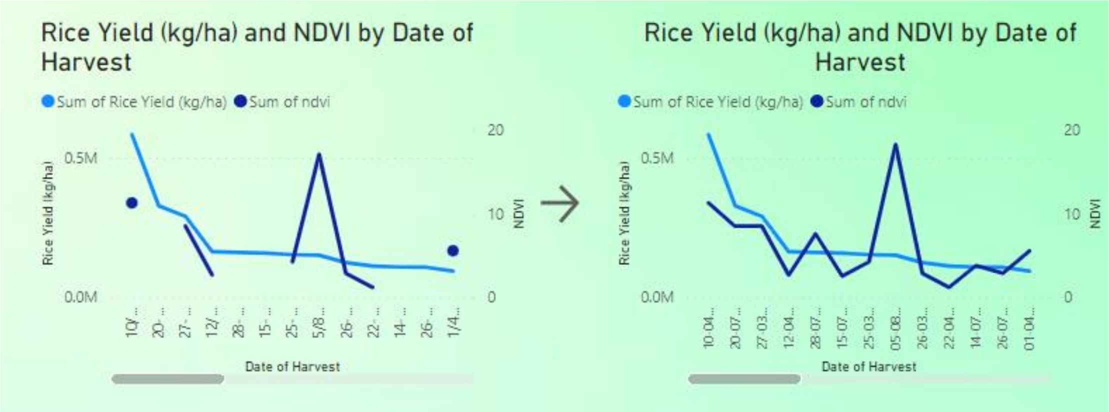


Figure 8. Rice Yield and NDVI by Date of Harvest

Above you will find our two line graphs that compare the Rice Yield and the NDVI values at certain data points. Both graphs are created for comparison on how the NDVI values follow the line of the Rice Yield.

On the left graph as mentioned previously the NDVI line has many points that are missing. That is the result of extracting the data from NIR and Red bands from the satellite at the data points where the clouds are covering more than 10% of the area. Having as a result the missing NaN values of NIR and Red data and as outcome the NaN value of the NDVI. The proof is shown in the left graph while changing the data points between seasons. We can see that in the Summer- Autumn (SA) season the NDVI values have much less data-values than the season Winter- Spring (WS). The reason this happens is that at Vietnam the SA season has generally a lot of rain and therefore many clouds.

The satellite cannot retrieve the data when there are clouds covering the area at specific data points.

In the right line graph, you may find that the NDVI values are all presented in a line that follows the Rice Yield line. On the right graph the difference is, as we have mentioned above, that we have added on all the NDVI missing values, the median NDVI value of the nearest (neighbor) area with clouds that cover less than 10% of this area. As a result, the right graph has a connected line on all the data points and throughout the length of the data, in both seasons. The comparison here shows us that the Rice Yield line follows very well the line of the NDVI, proving that our calculation of the NDVI, adding even the median NDVI on the NaN values, can provide a good understanding from the historic data and can help in the prediction of futures events.

Weather data

Since the growth of a crop is dependent on its environment, it is important to consider its weather variables. Features describing elements like the temperature that surrounds the crop, or the amount of water supply and solar energy which are important to the plant's development would increase the accuracy of a prediction if they were taken into consideration. The same applies to those describing cases of hazardous events or other elements that could harm a plant and in result decrease its rice yield.

To gather weather data for the crop's location, we utilized the website https://www.visualcrossing.com/, which provides an API system capable of extracting various climate variables for a specified location and date range. By referring to the documentation and obtaining an API key through account creation, we were able to make requests for the required data. The API was designed to accommodate data input in the form of latitude and longitude for the crop's location and the date range from the seeding date to the harvesting date. We utilized the 'requests' library to facilitate this process.

After passing a string containing the location for a crop and a specific range of time, we use the function 'request' from the library and as a return we get an object which contains the wanted data in a JSON form. However, there is a limit to the number of requests that can be made, in which case the program stops requesting new data.

Data Exploration

The obtained weather data is structured as an array, with each element representing a day within the provided date range. Each day is described using a dictionary that contains several key-value pairs representing various weather conditions concerning the specified crop.

The key features in each dictionary include:

- **datetime and datetimeEpoch:** Indicating the date of the weather data in the local time zone and the number of seconds since 1st January 1970 in UTC time, respectively.
- **temp, tempmax, and tempmin:** Average, maximum, and minimum temperature of the location throughout the day.
- **feelslike, feelslikemax, and feelslikemin:** Representing what the temperature feels like, accounting for heat index or wind chill. Daily values are provided as average (mean), maximum, and minimum values.
- **dew and humidity:** Describing the natural phenomenon of dew formation and the amount of moisture or water vapor present in the air, respectively. Humidity is represented in percentage.
- **precip, precipprob, preciptype, and precipcover:** Referring to the amount of liquid precipitation that fell or is predicted to fall in each period, including the liquid-equivalent amount of any frozen precipitation such as snow or ice. Precipitation probability is expressed as a percentage, and preciptype is an array indicating the type(s) of precipitation expected or that occurred (e.g., rain, snow, freezing rain, and ice).

- **snow and snowdepth:** Indicating the amount of snowfall and the depth of accumulated snow in the area.
- **windgust, windspeed, and winddir:** Representing instantaneous wind speed, sustained wind speed (average over the preceding one to two minutes), and wind direction. The first two daily values are the maximum hourly values for the day.
- **pressure:** Denoting the sea level atmospheric or barometric pressure in millibars (or hectopascals).
- **cloudcover and visibility:** Providing information on the amount of sky covered in clouds, ranging from 0 to 100%, and the distance at which distant objects are visible.
- **solarradiation and solarenergy:** Indicating the solar radiation power at the instantaneous moment of observation (or forecast prediction) in watts per square meter (W/m2) and the total energy from the sun that builds up over an hour or day in megajoules per square meter (MJ/m2).
- **uvindex:** Representing a value between 0 and 10 indicating the level of ultraviolet (UV) exposure for that hour or day. A value of 10 represents a high level of exposure, while 0 indicates no exposure. The UV index is calculated based on short-wave solar radiation, considering factors such as cloudiness, type of cloud, time of day, time of year, and location altitude. Daily values represent the maximum value of the hourly values.
- **severerisk (forecast only):** A value between 0 and 100 representing the risk of convective storms (e.g., thunderstorms, hail, and tornadoes). This measure combines various fields such as convective available potential energy (CAPE), convective inhibition (CIN), predicted rain, and wind. A value below 30 indicates a low risk, between 30 and 70 represents a moderate risk, and above 70 denotes a high risk.
- **sunrise and sunset:** Providing the formatted time of the sunrise and the sunset (e.g., '05:56:16').
- **sunriseEpoch and sunsetEpoch:** Indicating the number of seconds since 1st January 1970 in UTC time until the specific times of sunrise and sunset.
- **moonphase:** Representing the fractional portion through the current moon lunation cycle, ranging from 0 (new moon) to 0.5 (full moon) and back to 1 (the next new moon).
- **conditions and description:** Offering textual representations of the weather conditions and longer text descriptions suitable for display in weather displays. These descriptions combine the main features of the weather for the day, such as precipitation or the amount of cloud cover.
- **Icon:** A fixed, machine-readable summary that can be used to display an icon representing the weather conditions.
- **stations and source:** Identifying the weather stations used when collecting historical observation records and the type of weather data used for this weather object. Values include historical observation (“obs”), forecast (“fcst”), historical forecast (“histfcst”), or statistical forecast (“stats”). If multiple types are used in the same day, “comb” is used.

These features are used as insights of the conditions of the environment and their impact of the development of the plant. After processing, they would be very helpful in increasing the accuracy of our models. The website offers a complete documentation of the API and the ways to use it which is annotated bellow:

Timeline Weather API – Visual Crossing Weather

Feature Selection

To make use of this data, it is important to distinguish them by their value and contribution to the predictions. After looking at each feature, it was discovered that some of them maintain the same values throughout both cycles, or that are unusable due to their form or are better expressed by other values. At the same time some can be converted into a different feature to better suit the passing of information to the model.

For us to decide, we need to examine how each element is connected to the development of the plant. Plants need water for various vital processes and functions. Water is essential for their growth, survival, and overall health, so elements like the amount of water in the air or the probabilities of hazards like rain are of key interest. Additionally, the amount of solar energy or ultraviolet (UV) and the duration of which the crop could harvest it can be useful in this case, since plants use photosynthesis as one mean of collecting energy. With these someone might have been able to calculate the temperature of the environment, however the data regarding it from the website was covering some forms of aggregations for the day for our disposal. At the same time, wind can be a small factor in the health of the agricultural, for it can have both positive and negative effects, like seed dispersal or pollination or even drying of fields which reduces the number of available resources. Finally, pressure can also significantly affect agriculture, particularly in terms of weather patterns, plant growth, and overall crop health by manipulating weather patterns or affect plant growth. The rest of the features remained unused and are not included in the final form of the data.

Pre-processing

In the end, the variables temp, tempmax, tempmin were chosen for their indication of temperature regarding the area, while solarradiation, solarenergy, uvindex, sunrise and sunset to approximate the availability of solar energy for the plant. In regards for water supply we chose precip, precipcover as well as dew and humidity which depict both the amount of water and of time that it was obtainable throughout the day. As a final feature, we selected cloud cover to have as an extra measure calculating sun blockage and chance of rain.

In total these are over ten features, each a form of aggregation to describe a whole day. All were found directly connected and ready to be used in matters of their form of values except for sunrise and sunset which were in the form of strings. As such, they were converted from strings depicting the time, to integers depicting the amounts of seconds that have passed since the beginning of the day. Their difference was used to calculate the amounts of seconds that the fields are exposed to the sun throughout the day and the feature describing the value would be named Sunlight duration.

However, climate variables usually differ in small amounts from day to day, so feeding all the features for every day to the model without performing some aggregation would be unwise. The goal was to furtherly aggregate the variables and fuse them from each describing daily values to describing weekly values. Since the date ranges vary in number of days to 100 or even 150, it was decided to split the date ranges to 14 sections each so that each section would approximate a week. That way, performing an overall average which would lose some information is avoided, and at the same time a form of convolution is achieved making the difference of the climate variables more easily visible and accessible for the entirety of the cycle. That way, we reduce the number of variables that pass into the model, making it less computationally expensive and more agile.

The final form of the data in regards of weather, would be 557 rows with the season, latitude and longitude as the first 3 values and followed by the rest of 210 values, which resulted from the 15 final features of the weather data, all for each of 14 sections for the entire cycle.

5. Methodology / Implementation-Experimental design

The task of forecasting crop yields has garnered significant attention, given its direct implications on food security, economic planning, and agricultural resource management. By identifying key influential factors and predicting the outcomes accurately, farmers can streamline their strategies, optimize resource usage, and enhance productivity. This report explores the application of machine learning techniques, specifically K-Nearest Neighbors (KNN), Random Forest, and Extra Trees regression models, for crop yield prediction. The paper also delves into aspects of data preprocessing, model evaluation, and feature importance identification.

a. Data pre–processing

The dataset employed in this study is a compilation of numerous Comma Separated Values (CSV) files that include features pertinent to the crop, climate conditions, soil composition, and additional factors that influence agricultural output. The initial stage involves data cleaning and preprocessing to eliminate irrelevant variables, handle missing values, and prepare the data for modeling.

The categorical variables in the dataset are transformed into a machine-readable format using one-hot encoding. The 'Date of Harvest' feature is decomposed into multiple time-based features such as 'Year', 'Month', 'Quarter', and 'Day of the Year', offering more insights into temporal influences on crop yield.

Missing values pose a significant challenge in data analysis. For our dataset, the K-Nearest Neighbors Imputation technique is utilized to fill missing values. This method preserves the structure of the data by using the mean or median values of the k-nearest neighbor samples in the training set.

b. Modeling

Following the data preprocessing, the dataset is divided into a training set, which is used to train our models, and a test set, which serves to evaluate their performance (the dataset was split into 70% for training, 15% for validation, and 15% for testing.). The features are standardized using the StandardScaler to ensure that all features contribute proportionally to the final distance computation in the KNN model, and to the element of randomness in the ensemble methods. The GridSearchCV function is employed for hyperparameter optimization, ensuring the best performance of the models.

The first model implemented is the K-Nearest Neighbors (KNN) Regressor. A grid search for optimal parameters such as the number of neighbors and distance metric is performed. Once the model is trained, its predictive performance is evaluated using Root Mean Squared Error (RMSE) as the primary metric.

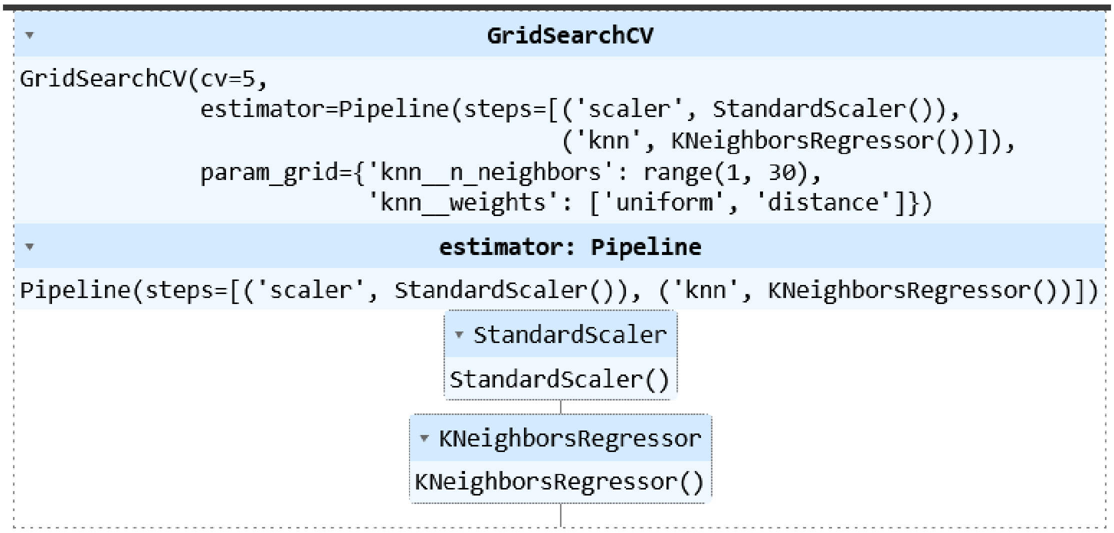
The K-Nearest Neighbors (KNN) imputation is a technique that estimates missing values in a dataset based upon the values of other non-missing elements. It operates under the assumption that those who are close to each other are more similar than those who are further apart.

Here is how it works:

1. **Distance Calculation**: The KNN algorithm calculates the distance between the input pattern for which the missing value needs to be imputed and all the complete case patterns in the dataset. There are various ways to calculate this distance, and the method usually used is Euclidean, though other methods like Manhattan or Minkowski can also be used.

2. **Determining the Neighbors**: Based on these distances, it selects the 'k' closest complete case patterns to the input pattern. The 'k' in KNN is a user-defined constant, and this value determines the number of neighbors we look at when we compute the imputation.

3. **Imputation**: Finally, it estimates the missing values of the input pattern based on the values of these 'k' neighbors. This estimation can be as simple as the average (for mean imputation) or median (for median imputation) of these 'k' neighbors. More complex imputations might take a weighted average of the 'k' neighbors, where the weights are the reciprocal of the distance from the neighbor to the input pattern.



```
GridSearchCV(cv=5,
             estimator=Pipeline(steps=[('scaler', StandardScaler()),
                                       ('knn', KNeighborsRegressor())]),
             param_grid={'knn__n_neighbors': range(1, 30),
                         'knn__weights': ['uniform', 'distance']})
```

```
Pipeline(steps=[('scaler', StandardScaler()), ('knn', KNeighborsRegressor())])
```

```
StandardScaler()
```

```
KNeighborsRegressor()
```

Random Forest Regressor

Model Training and Tuning

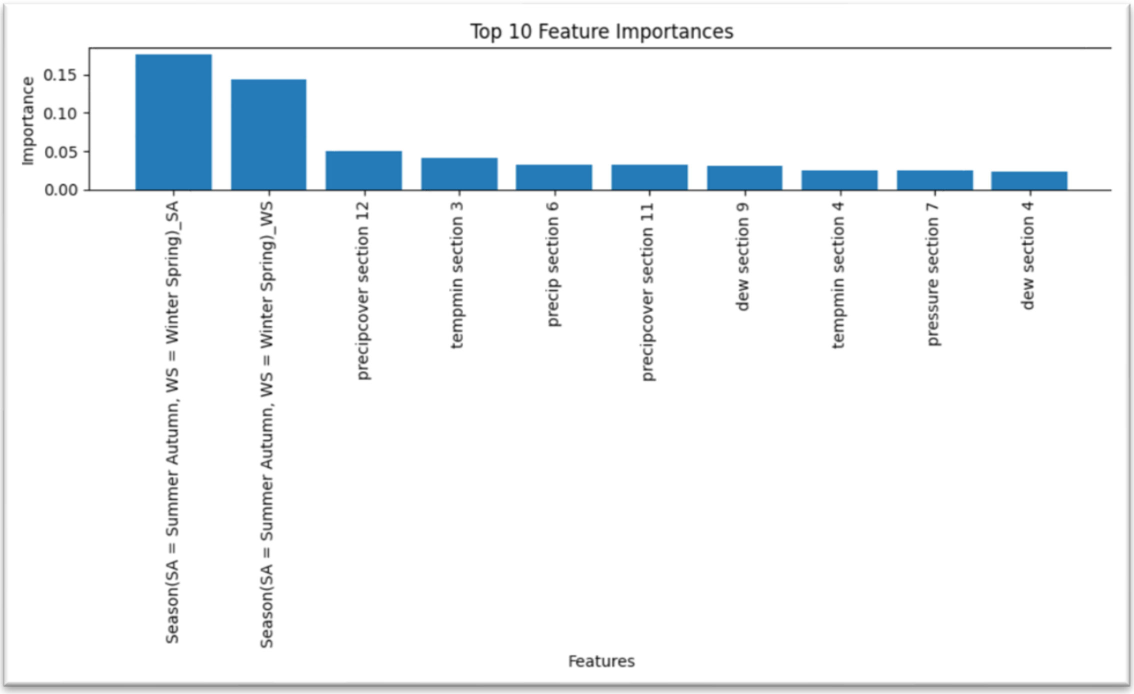
The Random Forest model was the first machine learning model to be implemented in this study. Random Forest is an ensemble learning method, which builds a multitude of decision trees during training and outputs either the mode of classes in the case of classification, or the mean prediction of individual trees in the case of regression.

The hyperparameters of the model were tuned using GridSearchCV. which is a handy tool that performs a search over the specified parameter values for an estimator. The following hyperparameters were adjusted in this case:

- **max_depth**: The maximum depth of each tree. This parameter is crucial in preventing overfitting by limiting how deep the tree can go. The optimal max_depth for this model was found to be 5.
- **min_samples_leaf**: The minimum number of samples required for a leaf node. This was set to 1, implying that every leaf must have at least one sample.
- **min_samples_split**: The minimum number of samples required to split an internal node. This was set to 5 in the model.
- **n_estimators**: The number of trees in the forest. The more trees, the more robust the model becomes. However, too many trees can slow down the model. The optimal number of trees for this model was found to be 200.

Feature Importance

Understanding the model's inner workings is crucial for gaining trust and making the model more useful. Therefore, feature importance was derived from the Random Forest Regressor. Feature importance indicates the contribution of each feature towards the prediction made by the model. In this case, the top 10 features were selected and their relative importances were visualized, offering an in-depth view into the variables that significantly affect crop yield predictions.



Model Evaluation

The performance of the Random Forest regressor was assessed using two key metrics:

- **Root Mean Squared Error (RMSE):** This metric measures the average magnitude of the error. It offers a straightforward way to measure the accuracy of the model. The RMSE on the validation set was found to be 381.247, and on the test set, it was 438.038. A lower RMSE signifies a better fit to the data.
- **Coefficient of Determination (R^2):** R^2 indicates the proportion of the variance for the dependent variable (in this case, the crop yield) that's explained by the independent variables in the model. The R^2 on the validation set was 0.731, and on the test set, it was 0.713. Higher R^2 values imply that the model can explain a larger portion of the variance.

Extra Trees Regressor

Model Training and Tuning

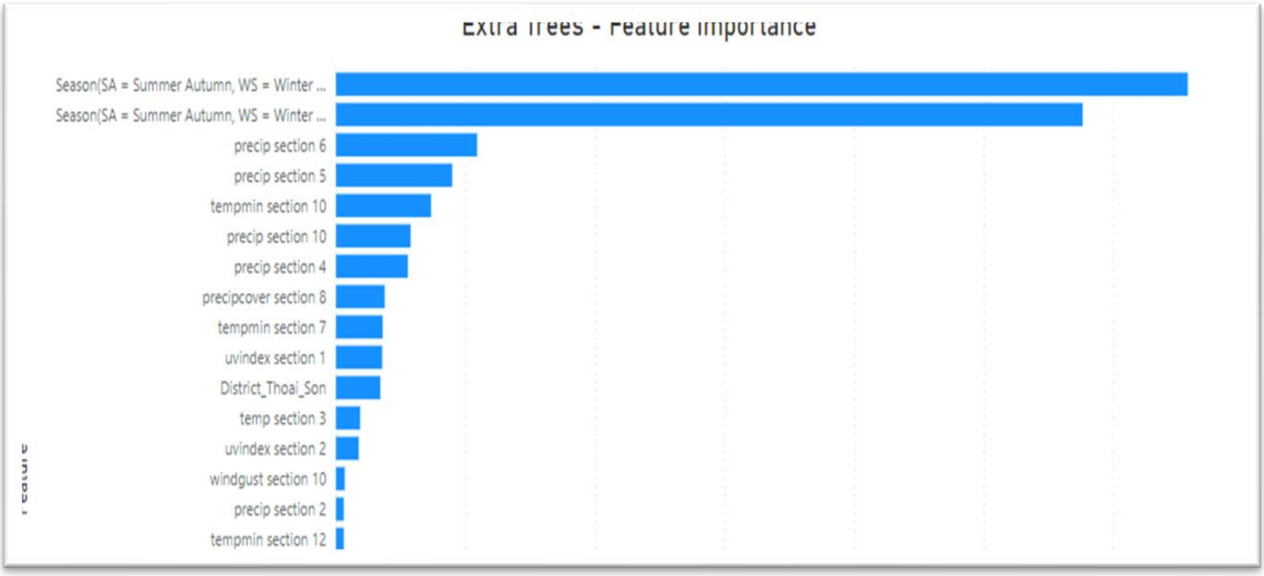
The Extra Trees Regressor was the second machine learning model implemented in this study. Similar to the Random Forest, the Extra Trees algorithm is an ensemble learning method, but it introduces more randomness in the creation of the trees. This added randomness can make Extra Trees more robust to overfitting, while also potentially introducing a slightly higher bias.

The hyperparameters of the Extra Trees Regressor were meticulously tuned using GridSearchCV. This tool offers the advantage of performing an exhaustive search over the specific parameter values for an estimator. For this particular model, the following hyperparameters were modified and optimised:

- **max_depth**: The optimal max_depth for this model was determined to be 5.
- **min_samples_leaf**: The optimal value for this parameter was found to be 4.
- **min_samples_split**: The optimal value for this parameter was determined to be 10.
- **n_estimators**: The optimal number of trees in the forest for this model was found to be 100.

Feature Importance

The top 10 features selected and visualized.



Model Evaluation

The performance of the Extra Trees regressor was evaluated using two key metrics:

- **Root Mean Squared Error (RMSE):** This metric offers a clear way to measure the average magnitude of the prediction error. The RMSE on the validation set was

386.402 and on the test set was 451.188. Lower RMSE values indicate a better fit of the model to the data.

- **Coefficient of Determination (R^2):** The R^2 score represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. The R^2 on the validation set was 0.724, and on the test set, it was 0.696. Higher R^2 values indicate that the model can explain a greater proportion of the variance.

c. Results

Based on the comprehensive analysis and evaluation conducted, both the Random Forest and Extra Trees Regressors demonstrated significant potential as models for crop yield prediction. They each offered a unique balance of performance, interpretability, and complexity.

The Random Forest Regressor, with an RMSE of 438.038 and R^2 of 0.713 on the test set, displayed slightly superior performance in terms of its fit to the data compared to the Extra Trees Regressor, which achieved an RMSE of 451.188 and R^2 of 0.696 on the test set. The Random Forest's performance indicates a slightly better predictive accuracy, suggesting it might be the more suitable choice for this particular prediction task.

However, the Extra Trees Regressor also performed admirably, displaying robustness against overfitting due to its inherent randomness. This characteristic can be particularly valuable in scenarios where there is high variance in the data or when predictions need to be made for data that significantly deviates from the training data.

Both models' feature importance analyses were insightful, highlighting key variables influencing crop yield predictions. This not only boosts our understanding of the models but also offers valuable information that can be used to guide agricultural strategies and decisions.

In summary, while the Random Forest Regressor showed marginally better performance, both models exhibited strong capabilities in crop yield prediction. The final choice between the two models may therefore depend on the specific needs and constraints of the implementation scenario, such as the tolerance for bias versus variance and the interpretability requirements. Regardless, both models can provide valuable insights and predictions to assist in optimizing crop yields and guiding future agricultural research and practice.

6. Discussion, conclusions, future work.

Sentinel – 1

As mentioned above, while the RVI can provide valuable insights into vegetation dynamics and growth stages, there may be differences between optical and radar phenology. These differences can arise due to factors like variations in reflectance properties and the specific interactions of radar waves with vegetation and soil.

Furthermore, the observed patterns in the RVI are what we expected. During the ripening stage, which precedes harvest, we anticipate a decrease in scattering due to the formation of rice tassels and the "layover" effect caused by the plant's bending over. These factors contribute to reduced scattering and can explain the drop in RVI values leading up to the harvest season. By considering these phenomena, we can better interpret the RVI data and its correlation with the growth stages of the crops. That’s why even though on Figure 4. Average VV value by date. and Figure 5. Average VH value by date. we don’t observe a pattern between the values VV and VH and the growing stage of the plant, with RVI we can see a clear pattern.

Landsat

Landsat, as a vital satellite dataset, plays a crucial role in the 2023 EY Open Science Data Challenge. We can leverage optical data from Landsat and other datasets to track rice crop phenology and forecast yield. By addressing the challenges associated with cloud cover and leveraging the advantages of radar data, we have the opportunity to contribute to food security and agricultural advancements on a global scale.

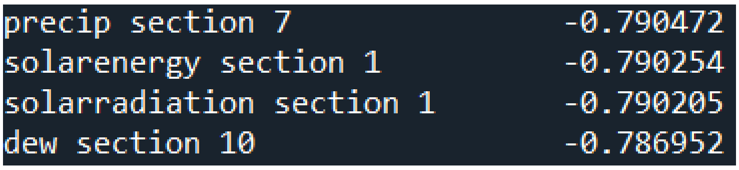
In conclusion, optical data from satellites like Landsat offers valuable insights into rice crop phenology by leveraging various spectral bands and indices. By tracking the NDVI and other relevant indices over time, researchers and agricultural stakeholders can better comprehend the growth stages of rice crops, predict yield potential, and make informed decisions for sustainable agricultural management. The combination of high temporal resolution and multispectral capabilities provided by these satellites has revolutionized phenology analysis and enhanced our understanding of crop dynamics for improved food security and agricultural planning in Vietnam and beyond.

By combining optical and radar data, participants can build more comprehensive yield models. Optical data measures the "greenness" of the plant, while radar data provides information about the "structure" of the crop. Participants can relate this data to the growth stages and build accurate yield models.

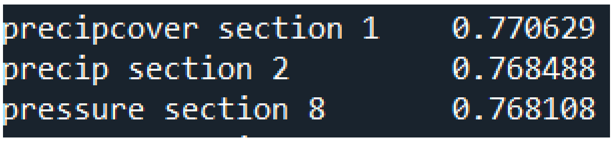
Additionally, considering the effects of climate change on rice crop growth is crucial. Rising temperatures, altered precipitation patterns, and extreme weather events can influence the phenological cycle of rice crops, leading to changes in yield.

Weather Data

While studying and analyzing the data, it was discovered that some features regarding specific sections can be more connected to the target variable than other with significant difference. This was observed by calculating the correlation and realizing that there are results of values regarding the same feature from different sections that have highly negative correlation and some of them having highly positive correlation. This is happening due to the cycle of the life of each plant, where it has different needs from each stage of life. Additionally, an observation of the means of the features for the entire cycle would result in highly negative correlation with the yield crop variable which seems unreasonable considering these are supplies needed for the development of the crop. In conclusion, it is important to have a thorough understanding of ones features to effectively pre-process them so that they may obtain the best form useful for the training and the predictions of the model.



```
precip section 7                -0.790472
solarenergy section 1           -0.790254
solarradiation section 1        -0.790205
dew section 10                  -0.786952
```



```
precipcover section 1    0.770629
precip section 2         0.768488
pressure section 8       0.768108
```

7. References

Wikifarmer. (n.d.). Rice Harvesting, Yield per Hectare and Storage. Retrieved May 20, 2023, from https://wikifarmer.com/rice-harvesting-yield-per-hectare-and-storage/

Wang, J., Sun, X., Xu, Y., Wang, Q., Tang, H., & Zhou, W. (2021). The effect of harvest date on yield loss of long and short-grain rice cultivars (Oryza sativa L.) in Northeast China. European Journal of Agronomy, 131, 126382. https://doi.org/10.1016/j.eja.2021.126382

Sentinel-Hub by Sinergise (no date) *Radar Vegetation index for sentinel-1 SAR data - RVI4S1 script*, *Sentinel Hub custom scripts*. Available at: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-1/radar_vegetation_index/ (Accessed: 29 June 2023).