**DATA SCIENCE FOR AGRICULTURE**

**FINAL PROJECT**

Raj Saha

**Semester:** Fall 2023

**Course:** DSA 5900 Professional Practicum

**Credit Hours:** 4

A red and white logo

Description automatically generated

**Faculty Supervisor:** Dr. Danala Gopichandh

**Company & Sponsor:** Data Institute for Societal Challenges (DISC) – Dr. Ebert

**Credits:** United States Department of Agriculture. Agricultural Research Service (USDA-ARS), El-Reno, Oklahoma, USA

**ACKNOWLEDGEMENT**

I want to express my sincere gratitude to my professor Dr. Matthew Beattie for conducting this beautiful course and to my sponsor Dr. Danala Gopichandh for providing me with constant guidance and support throughout the project. I am grateful for the invaluable knowledge and experience I have gained under their direction. Their contributions and encouragement were instrumental in the successful completion of this project.

Once again, thank you to my professors, and sponsors for their unwavering support and assistance.

Raj Saha

**Table of Contents:**

Contents

[1. Introduction: 6](#_Toc152502117)

[2. Objectives: 6](#_Toc152502118)

[a) Technical Objectives: 7](#_Toc152502119)

[b) Individual Learning Objectives: 7](#_Toc152502120)

[3. Data: 7](#_Toc152502121)

[3.1 Data Ingestion: 7](#_Toc152502122)

[Predictive Modeling of EVI and LSWI: 8](#_Toc152502123)

[Predictive Modeling of ER: 9](#_Toc152502124)

[3.2 Data Preparation: 9](#_Toc152502125)

[Predictive Modeling of EVI and LSWI: 9](#_Toc152502126)

[Predictive Modeling of ER: 14](#_Toc152502127)

[3.3 Data Exploration: 17](#_Toc152502128)

[Predictive Modeling of EVI and LSWI: 17](#_Toc152502129)

[Predictive Modeling of ER: 20](#_Toc152502130)

[4. Methodology: 21](#_Toc152502131)

[4.1 Techniques: 21](#_Toc152502132)

[4.2 Procedures: 22](#_Toc152502133)

[5. Results and Analysis: 25](#_Toc152502134)

[Predictive Modeling of EVI and LSWI: 25](#_Toc152502135)

[Predictive Modeling of ER: 30](#_Toc152502136)

[6. Deliverables: 35](#_Toc152502137)

[7. References: 36](#_Toc152502138)

[8. Self-Assessment: 37](#_Toc152502139)

**List of Figures and tables:**

[Figure 1: Flowchart from Data Preparation to Dimensionality Reduction 8](#_Toc152346411)

[Figure 2: No. of Exponents vs Explained Variance for EVI 8](#_Toc152346412)

[Figure 3: No of Components vs Explained Variance for LSWI 10](#_Toc152346413)

[Figure 4: No of Exponents vs Explained Variance for the overall dataset 11](#_Toc152346414)

[Figure 5: No of Exponents vs Explained Variance for GPP, NEE, ET at a time 13](#_Toc152346415)

[Figure 6: Correlation Analysis Plot for objective 1 16](file:///C:\Users\tapar\Downloads\Mid%20term%20report_Saha_Raj_DSA%205900%20(2).docx#_Toc152346416)

[Figure 7: Histogram of independent variables for objective 1 17](#_Toc152346417)

[Figure 8: Scatter plot of SAVG for EVI and LSWI 17](file:///C:\Users\tapar\Downloads\Mid%20term%20report_Saha_Raj_DSA%205900%20(2).docx#_Toc152346418)

[Figure 9: Scatter Plot of TAVG for both EVI and LSWI 18](#_Toc152346419)

[Figure 10: Correlation Plot for the objective 2 18](#_Toc152346420)

[Figure 11: Histogram Analysis for Objective 2 19](#_Toc152346421)

[Figure 12: Scatter Plot of ER with both rH and VPD 19](#_Toc152346422)

[Figure 13: Flowchart of Overall Process 21](#_Toc152346423)

[Figure 14: Regression plot of XGBoost for target variable EVI 27](#_Toc152346424)

[Figure 15: Regression plot of Random Forest for target variable LSWI 28](#_Toc152346425)

[Figure 16: Top 5 features extracted from XGBoost for target variable EVI 29](#_Toc152346426)

[Figure 17: Top 5 features extracted from XGBoost for target variable LSWI 30](#_Toc152346427)

[Figure 18: Top 5 features extracted from Random Forest Regressor for target variable ER using NEE 34](#_Toc152346428)

[Table 1: Weather and Soil Variables 8](#_Toc152507830)

[Table 2: Satellite Measurements Data 8](#_Toc152507831)

[Table 3: Environment Variables 9](#_Toc152507832)

[Table 4: Features contributing to the first two principal components for target variable EVI. 11](#_Toc152507833)

[Table 5: Features contributing to the first two principal components for target variable LSWI. 12](#_Toc152507834)

[Table 6: Features contributing PC1 and PC2 for overall Dataset. 14](#_Toc152507835)

[Table 7: Features contributing to the principal components while considering GPP. 16](#_Toc152507836)

[Table 8: Features contributing to the principal components while considering NEE. 16](#_Toc152507837)

[Table 9: Features contributing to the principal components while considering ET. 17](#_Toc152507838)

[Table 10: Hyperparameter Tuning for EVI 27](#_Toc152507839)

[Table 11: Results of the target variable EVI without hyperparameter tuning 27](#_Toc152507840)

[Table 12: Results of the target variable EVI with hyperparameter tuning. 27](#_Toc152507841)

[Table 13: Results of the target variable LSWI without hyperparameter tuning. 27](#_Toc152507842)

[Table 14: Results of the target variable LSWI with hyperparameter tuning. 28](#_Toc152507843)

[Table 15: Hyperparameter Tuning for GPP Modeling. 32](#_Toc152507844)

[Table 16: Results of ER without using hyperparameter tuning considering GPP. 32](#_Toc152507845)

[Table 17: Results of ER using hyperparameter tuning considering GPP. 32](#_Toc152507846)

[Table 18: Hyperparameter Tuning for NEE Modeling. 33](#_Toc152507847)

[Table 19: Results of ER without using hyperparameter tuning considering NEE. 33](#_Toc152507848)

[Table 20: Results of ER using hyperparameter tuning considering NEE. 33](#_Toc152507849)

[Table 21: Hyperparameter Tuning for ET Modeling. 34](#_Toc152507850)

[Table 22: Results of ER without using hyperparameter tuning considering ET. 34](#_Toc152507851)

[Table 23: Results of ER using hyperparameter tuning considering ET. 34](#_Toc152507852)

**Glossary:**

**Enhanced Vegetation Index(EVI):** Enhanced Vegetation Index (EVI) is a vegetation health indicator that combines multiple spectral bands to mitigate atmospheric influences and provide a more accurate measure of vegetation density and vigor. It offers a nuanced assessment of plant health, incorporating adjustments for factors such as soil background and atmospheric conditions, thereby enhancing its sensitivity to variations in vegetation cover.

**Land Surface Water Index(LSWI)**:The Land Surface Water Index (LSWI) is a remote sensing-derived metric that quantifies the presence of surface water on the Earth's land surface. It combines information from different spectral bands to distinguish between water and non-water features, providing valuable insights into the spatial distribution of land surface water.

**Evapotranspiration:** Evapotranspiration is the combined process of water vapor evaporation from the Earth's surface and transpiration from plants, contributing to the overall water cycle. It represents the total water loss to the atmosphere through these dual mechanisms.

**Gross Primary Production:** Gross Primary Production (GPP) refers to the total amount of energy plants captured during photosynthesis within a specific period. It represents the gross assimilation of carbon dioxide into organic compounds by autotrophs, laying the foundation for the overall energy flow within ecosystems.

**Net Ecosystem Exchange:** It is a measure in ecology that quantifies the overall exchange of carbon dioxide between an ecosystem and the atmosphere, encompassing photosynthetic carbon uptake and respiratory carbon release. A positive NEE value indicates net carbon uptake by the ecosystem, while a negative value suggests a net release of carbon into the atmosphere.

**Ecosystem Respiration:** Ecosystem respiration(ER) refers to the collective exhalation of carbon dioxide by all living organisms within an ecological system. It represents a fundamental process in the carbon cycle, indicating the total respiratory activity of plants, animals, and microorganisms contributing to carbon flux in an ecosystem.

# Introduction:

The environment and global food security depend on sustainable agriculture, which involves using farming techniques that are environmentally friendly to produce crops while preserving the fertility and health of the soil. To accomplish this, farmers need accurate information on soil parameters such as nutrients, soil moisture, and pH levels. Important decisions, such as maintaining soil health, minerals, and nutrients, preventing erosion, deciding which crops to grow, determining care needs like the necessary amount of water and fertilizer for plants, and identifying the best planting and growing conditions, all depend on this information.

In addition, the United States Department of Agriculture (USDA) must comprehend various field aspects to provide sound advice to the government on managing the land, utilizing ecological water, and combating rising CO2 levels. Eddy covariance (EC) or an infrared gas analyzer is used for measurements (Wagle, 2019), and these systems have remote monitoring capabilities. They must also be scalable for improved monitoring, which might incur significant expenses when established in the field. The USDA is primarily interested in values such as the Enhanced Vegetation Index (EVI) and Land Surface Water Index (LSWI). Net Ecosystem Exchange (NEE), Gross Primary Product (GPP), and environmental respiration (ER) are all critical factors in managing CO2 levels. The USDA has been exploring ways to use weather data, including average temperature (TAVG), average relative humidity (HAVG), average daily vapor deficit (VDEF), heating degree days (Hdeg), cooling degree-days (65 standards) (CDEG), average wind speed (Wspd), solar radiation, and rainfall, to predict field variables, rather than solely relying on sensor data. The USDA measures the Enhanced Vegetation Index (EVI) and Land Surface Water Index (LSWI) using moderate-resolution spectroradiometer (MODIS) satellite images.

Earlier research has demonstrated that weather and Enhanced Vegetation Index (EVI) data can predict several field and soil characteristics effectively. The data also indicates variations in the values of field parameters across different pastures. In this context, our approach involves predicting the field characteristics of multiple pastures using information from a single pasture instead of relying on sensor data. This methodology is expected to assist the USDA in optimizing resource allocation and avoiding unnecessary expenditures.

# Objectives:

## Technical Objectives:

From a technical perspective, the project's main objective is to provide the USDA with a better awareness of the conditions in all fields so that decisions can be made without installing expensive monitors.

* + Predictive Modeling of EVI and LSWI: The first objective of our project is to use one pasture data (Weather and soil variables of P13) to predict satellite readings such as Enhanced Vegetation Index (EVI) as well as Land Surface Water Index (LSWI) and then validate the same for two different pastures. The data provided here is a daily value, so we must interpolate the 8-day satellite readings to create daily estimates using linear interpolation.
  + Predictive Modeling of ER: The project's second objective is to use weather, and soil variables, and combine them with one of the 3-sensors measurements of Gross Primary Production (GPP), Net Ecosystem Exchange (NEE), Evapotranspiration (ET) at a time to model for environmental respiration (ER).

## Individual Learning Objectives:

* + My learning objectives are to work on a data science project, from analyzing the data to creating a working model and evaluating the results of the model.
  + This project will be the entire data science process, going from data Ingestion, data analysis, creating a model, and analyzing the results. This project will also develop Python skills to create a proper model and explain the results in a paper.

# Data:

The USDA-ARS provided two data sets for this project.

* + The first data set is the Pasture 13 vs. weather data including soil variables. The dataset size is 8403 rows, and the timeframe for the data collection is from 2000 to 2022. This data set comprises 8-day composite satellite-derived vegetation greenness data for a native tallgrass prairie pasture from a MODIS Satellite, including the daily weather containing daily averages of air temperature, relative humidity, cooling degree days, heating degree days, wind speed, solar radiation, and rain.
  + The second data set is the Environmental respiration (ER) data, the size of which is 65808 rows, collected in 30-minute intervals for pasture 20 in the timeframe of 2019 to 2022, The USDA took both the ER and environment variables.

# 3.1 Data Ingestion:

# Predictive Modeling of EVI and LSWI:

For this first objective, in the first step, we combined weather variables, including maximum air temperature, minimum air temperature, average air temperature, average relative humidity, average daily vapor deficit, heating degree-days, cooling degree-days, minimum wind chill index temperature, average wind speed, solar radiation, and daily rainfall, with soil variables, which encompassed average soil temperatures at 10 cm under sod and bare soil, as well as soil moisture calibrated measurements at 5 cm, 25 cm, and 60 cm depths. These variables were merged into a single dataset, creating a comprehensive repository for further analysis. To bridge gaps in the 8-day MODIS satellite dataset, we employed linear interpolation techniques. This process involved estimating values for the days between actual data points, transforming the 8-day composite satellite measurements into daily averages. The resulting dataset now provided a seamless integration of MODIS data and daily weather data, using the 'Date' column as a common key.

**Attributes of the data:**

Table 1 & 2 shows all the columns of the two datasets used for the first objective and its description:

Table : Weather and Soil Variables

|  |  |
| --- | --- |
| Date | Date |
| TMIN, TMAX, TAVG | Maximum, Min, and Average Temperature |
| HAVING | Avg Relative Humidity |
| VDEF | Avg Daily Vapor Deficit |
| HDEG | Heating Degree Days |
| CDEG | Cooling Degree Days |
| WCMN | Min Wind Chill Temp |
| WSPD | Avg Wind Speed |
| ATOT | Solar Radiation |
| Rain | Rainfall |
| SAVG | Avg Soil Temp 10 cm under sod |
| BAVG | Avg Soil Temp 10cm under bare soil |
| TR05, TR25, TR60 | Soil Moisture at 5cm, 25cm, 60cm depth. |

Table : Satellite Measurements data

|  |  |
| --- | --- |
| Date | Date |
| EVI | Enhanced Vegetation Index |
| LSWI | Land Surface Water Index |

# Predictive Modeling of ER:

For this objective, we uploaded the data provided by the USDA onto Google Colab. I did not encounter any complexity while performing the data ingestion.

**Attributes of the data:**

Table 3 shows all the columns of the dataset used for the second objective and its description.

Table : Environment variables

|  |  |
| --- | --- |
| Date | Date |
| Rg | Solar Radiation |
| Hair | Air Temperature |
| Tsoil | Soil Temperature |
| RH | Relative Humidity |
| VPD | Vapor Pressure Deficit |
| UStar | Turbulence |
| SWC | Soil Water Content |
| GPP | Gross Primary Production measured using sensor 1 |
| NEE | Net Ecosystem Exchange measured using sensor 2 |
| ET | Evapotranspiration measured using sensor 3 |
| ER | Ecosystem Respiration |

# 

# 3.2 Data Preparation:

# Predictive Modeling of EVI and LSWI:

In the data preparation phase of this first objective, it became evident that certain columns, such as minimum air temperature, average air temperature, average relative humidity, average daily vapor deficit, and heating degree days, contained noisy readings. We addressed this issue by employing linear interpolation methodologies to replace the noisy data. Additionally, we identified outliers in specific columns. To mitigate this, we applied linear interpolation based on the z-score method. Furthermore, some columns exhibited highly positive skewness, such as Cooling Degree Days, Rainfall, Soil moisture at depths of 5 cm and 25 cm, and Rolling sum of rain for 7 days, 14 days, 21 days, and 28 days. These were identified through histogram analysis. To mitigate their effects on model performance, we performed a logarithmic transformation on these columns.

Given that our first dataset encompasses two target variables, namely "EVI" and "LSWI," and our project objectives require separate predictions for these targets, we partitioned the main dataset into two distinct sub-datasets. One sub-dataset focuses on predicting the "EVI" target variable, while the other focuses on predicting the "LSWI" target variable.

We employed feature selection techniques to identify the most important variables for our predictive models. Specifically, we utilized Recursive Feature Elimination (RFE) in conjunction with a linear regression model to extract the top 8 features that carry significant weight in predicting our target variables. Following the RFE process, we further applied Principal Component Analysis (PCA) to determine the number of components that collectively account for 95% of the variance within these top 8 features.

Figure 1 provides a flowchart providing an overview from Data preparation to the stage of Dimensionality Reduction.

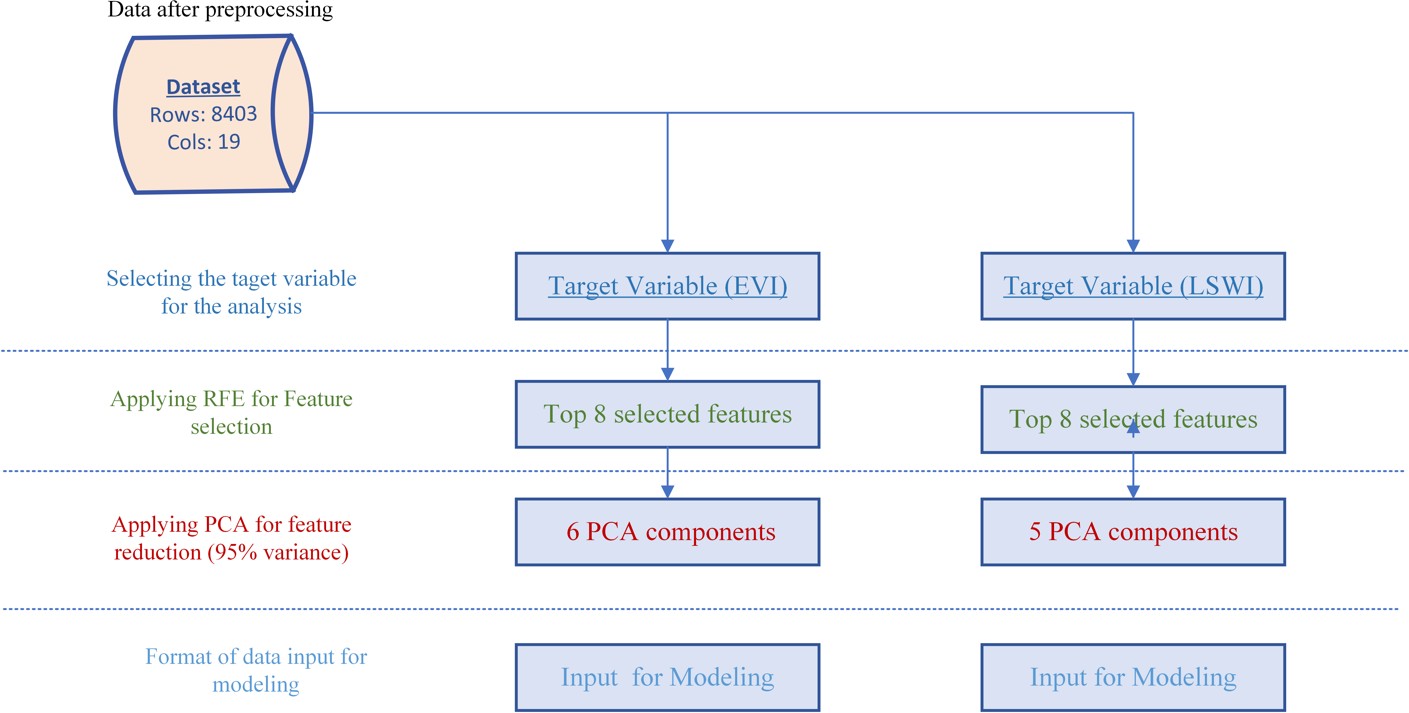


Figure : Flowchart from Data Preparation to Dimensionality Reduction

After selecting the top 8 features, we apply PCA to the dataset containing the target variable “EVI”, resulting in the plot depicted in Figure 2.

A graph with a line

Description automatically generated

Figure : No. of Exponents vs Explained Variance for EVI

From Figure 2, it is clear that six components are necessary for predicting the target variable “EVI”.

Now, we aim to elucidate the contribution of each independent variable to the first two principal components by examining the absolute values associated with each feature within these components, as illustrated in the figures below. For the dataset containing only the target variable 'EVI’, the resultant dataframe for the first two components of PCA is presented in Table 4.

|  |  |
| --- | --- |
|  |  |

Table : Features contributing to the first two principal components for target variable EVI.

|  |  |  |
| --- | --- | --- |
| index | PC1 | PC2 |
| VDEF | 0.891955 | 0.061236 |
| ATOT | 0.733901 | 0.291028 |
| RAIN | -0.15941 | 0.32068 |
| SAVG | 0.802611 | 0.371849 |
| TR25 | 0.756584 | -0.27963 |
| TR60 | 0.543174 | -0.3263 |
| RAIN\_21\_Days | -0.07677 | 0.921135 |
| RAIN\_28\_Days | -0.06011 | 0.931888 |

In Table 4, a detailed examination of the absolute coefficients reveals the distinct contributions of various features to Principal Component 1 (PC1) and Principal Component 2 (PC2). Notably, features such as Average Daily Vapor Deficit (VDEF), Average Soil Temp under 10cm sod (SAVG), and Soil temperature at 25cm depth (TR25) exert substantial influences on PC1. In contrast, the rolling sum of Rain for 21 days (Rain\_21\_days), the Rolling sum of Rain for 28 days (Rain\_28\_days), and the Rain feature demonstrate comparatively lower contributions to PC1. Turning attention to PC2, a conspicuous pattern emerges where the rolling sum of Rain for 21 days (Rain\_21\_days) and the rolling sum of Rain for 28 days (Rain\_28\_days) play pivotal roles, making substantial contributions to the variation captured by this principal component. Conversely, features such as Average Daily Vapor Deficit (VDEF) and Soil temperature at 25cm depth (TR25) exhibit diminished contributions to PC2.

Similarly, if we apply PCA to the dataset containing the target variable “LSWI”, we get the below plot as shown in Figure 3.

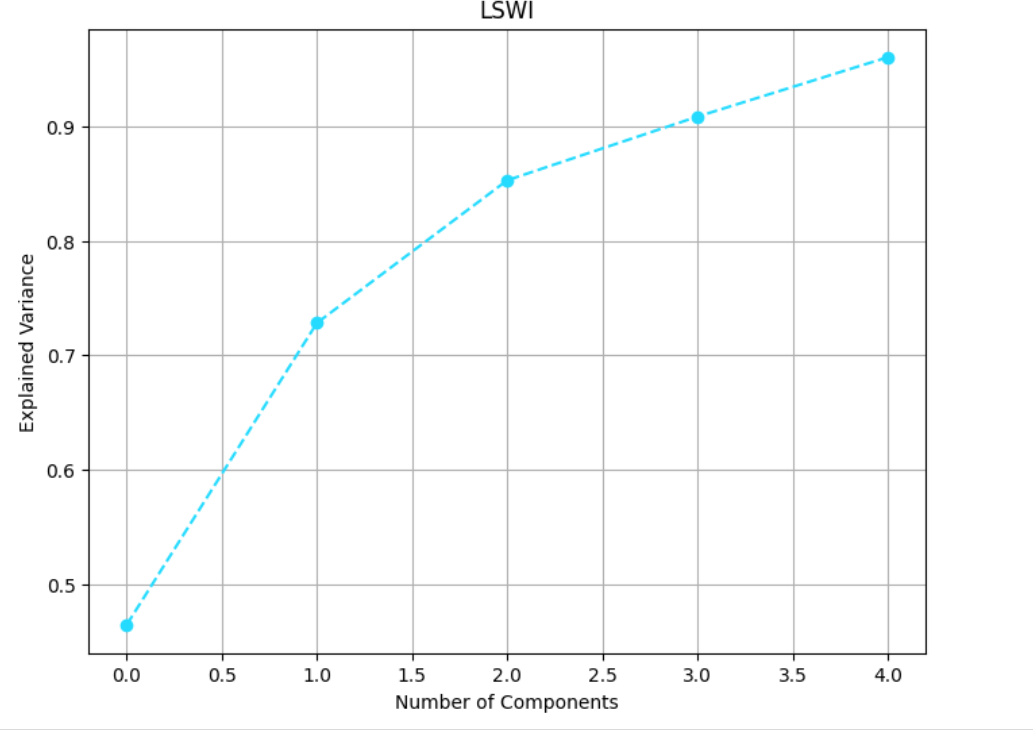


Figure : No of Components vs Explained Variance for LSWI

From Figure 3, it is also quite clear that five components are necessary for predicting the target variable “LSWI. For the dataset containing only the target variable “LSWI”, the resultant data frame for the first two components of PCA is presented in Table 5 as shown below.

Table : Features contributing to the first two principal components for target variable LSWI.

|  |  |  |
| --- | --- | --- |
| index | PC1 | PC2 |
| TMAX | 0.866123 | 0.307121 |
| SAVG | 0.885417 | 0.386243 |
| BAVG | 0.893355 | 0.351949 |
| TR05 | 0.708446 | -0.50787 |
| TR25 | 0.768765 | -0.37423 |
| TR60 | 0.532296 | -0.39037 |
| RAIN\_7\_Days | -0.07717 | 0.703913 |
| RAIN\_28\_Days | -0.0153 | 0.83664 |

As evident from Table 5, a thorough examination of the absolute values of the coefficients sheds light on the distinctive impacts of various features on Principal Component 1 (PC1) and Principal Component 2 (PC2). Notably, features such as Maximum Temperature (TMAX), Average Soil Temp under 10cm sod (SAVG), and Average Soil Temp under 10cm bare soil (BAVG) exert substantial influences on PC1. In contrast, the Rolling sum of Rain for 7 days (Rain\_7\_days) and the Rolling sum of Rain for 28 days (Rain\_28\_days) exhibit markedly diminished contributions to PC1. Directing attention towards PC2, a salient pattern emerges where the Rolling sum of Rain for 7 days (Rain\_7\_days) and Rolling sum of Rain for 28 days (Rain\_28\_days) assume pivotal roles, making significant contributions to the variance encapsulated by this principal component. In contrast, Maximum Temperature (TMAX) and Average Soil Temp under 10cm bare soil (BAVG) demonstrate comparatively diminished impacts on PC2. Now, if we don’t consider any feature selection algorithm to select the top 8 features and just apply PCA to all of the independent variables of our overall dataset to identify the components that carry 95% variance of the data, we will get the below plot as shown in Figure 4.

|  |  |
| --- | --- |
|  |  |
| Figure : No of Exponents vs Explained Variance for the overall dataset |  |

Figure 4 elucidates a notable observation: specifically, that 10 principal components effectively encapsulate 95% of the total variance within our overall dataset. It is imperative to highlight that, in this instance, the PCA algorithm is applied to the same dataset for both target variables yielding an identical number of components.

Similarly, for the overall dataset, the resultant dataframe for the first two components of PCA is presented in Table 6.

Table : Features contributing PC1 and PC2 for overall Dataset.

|  |  |  |
| --- | --- | --- |
| index | PC1 | PC2 |
| TMAX | 0.961206 | 0.038039 |
| TMIN | 0.921902 | 0.1934 |
| TAVG | 0.97066 | 0.111499 |
| HAVG | -0.27703 | 0.52615 |
| VDEF | 0.855258 | -0.26989 |
| HDEG | -0.8803 | -0.18798 |
| CDEG | 0.855303 | -0.04227 |
| WSPD | 0.010463 | -0.12408 |
| ATOT | 0.720965 | -0.02469 |
| RAIN | -0.01851 | 0.424282 |
| SAVG | 0.939973 | 0.162433 |
| BAVG | 0.957035 | 0.109716 |
| TR05 | 0.526245 | -0.59093 |
| TR25 | 0.567206 | -0.39348 |
| TR60 | 0.323391 | -0.32383 |
| RAIN\_7\_Days | 0.026528 | 0.768373 |
| RAIN\_14\_Days | 0.089541 | 0.874164 |
| RAIN\_21\_Days | 0.122476 | 0.882479 |
| RAIN\_28\_Days | 0.1489 | 0.843402 |

|  |  |
| --- | --- |
|  |  |

As observed in Table 6, a thorough examination of the absolute coefficients reveals the distinct contributions of various features to Principal Component 1 (PC1) and Principal Component 2 (PC2). Notably, features such as Maximum Temperature (TMAX), Minimum Temperature (TMIN), and Average Temperature (TAVG) exert substantial influences on PC1. Conversely, features like the Rolling sum of Rain for 7 days (Rain\_7\_days) and Average Wind Speed (WSPD) exhibit minimal impact on PC1. Turning attention to PC2, a noticeable pattern comes to the fore. Here, the rolling sum of Rain for 14 days (Rain\_14\_days) and the rolling sum of Rain for 21 days (Rain\_21\_days) play pivotal roles, making significant contributions to the variance captured by this principal component. In contrast, Cooling Degree Days (CDEG) and Solar Radiation (ATOT) distinctly showcase minimal influences on PC2.

# Predictive Modeling of ER:

For this second objective, the data preparation stage closely mirrors that of the first objective, particularly in addressing outliers, handling missing values, and feature scaling methodologies in the independent variables. However, a distinct aspect arises concerning the target variable ER, where certain missing values were identified and subsequently addressed through the removal of corresponding rows.

Additionally, notable attention was directed towards data integrity, specifically in columns such as Solar Radiation (Rg), vapor pressure deficit (VPD), relative humidity (rH), Turbulence (UStar), and Soil water content (SWC). In a single row, where these columns contained the value -9999 were identified as noisy readings, aligning with stakeholder specifications. Consequently, rows associated with such occurrences were systematically excluded from the dataset to ensure the robustness and reliability of the data for the pursuit of the second objective.

We have studied the analysis for the second objective by considering independent variables such as GPP, NEE, and ER at a time, then we apply PCA separately by taking each of these variables at a time along with the other independent variables.

When applying PCA to all independent variables and considering Gross Primary Production (GPP), Net Ecosystem (NEE), and Evapotranspiration (ET) separately, we obtain the same plot as depicted in Figure 5 below each time.

A graph with a dotted line

Description automatically generated

Figure : No of Exponents vs Explained Variance for GPP, NEE, ET at a time.

From Figure 5, it is quite evident that if we apply PCA while considering each of these independent variables such as GPP, ET, and NEE at a time along with the other independent variables, it is clear that 6 components carry 95% variance of the data. Now, we want to see how all the independent variables contribute to the first two components by examining their absolute values associated with each feature in these principal components. So, when considering only the GPP along with other variables, we get the below dataframe as shown in Table 7.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table : Features contributing to the principal components while considering GPP.   |  |  |  | | --- | --- | --- | | index | PC1 | PC2 | | Rg | 0.711054 | -0.11718 | | Tair | 0.864654 | 0.245397 | | Tsoil | 0.689912 | 0.477476 | | RH | -0.48888 | 0.654031 | | VPD | 0.916432 | -0.25599 | | Ustar | 0.284836 | 0.030151 | | SWC | -0.39378 | 0.412343 | | GPP | 0.684684 | 0.246691 | |

As discerned from Table 7, a meticulous analysis of the absolute values of the coefficients illuminates the distinct contributions of various features to Principal Component 1 (PC1) and Principal Component 2 (PC2). Notably, features such as Vapor Pressure Deficit (VPD) and air temperature (Tair) exhibit pronounced influences on PC1, while Turbulence (UStar), and Soil water content (SWC) demonstrate comparatively modest contributions to PC1. Shifting focus to PC2, a noteworthy pattern emerges, wherein relative humidity (rH) and soil temperature (Tsoil) play pivotal roles, contributing significantly to the variation captured by this principal component. In contrast, Turbulence (UStar) and Solar Radiation (Rg) exhibit diminished contributions to PC2.

When specifically incorporating NEE alongside other independent variables, the resultant dataframe is presented in Table 8 as shown below.

Table : Features contributing to the principal components while considering NEE.

|  |  |  |
| --- | --- | --- |
| index | PC1 | PC2 |
| Rg | 0.722576 | 0.411889 |
| Tair | 0.846754 | -0.4388 |
| Tsoil | 0.655997 | -0.71236 |
| rH | -0.52572 | -0.64303 |
| VPD | 0.930678 | 0.125667 |
| Ustar | 0.288377 | 0.287517 |
| SWC | -0.39807 | 0.152722 |
| NEE | -0.57323 | -0.30572 |

As depicted in Table 8, a thorough examination of the absolute coefficients sheds light on the distinct influences of various features on Principal Component 1 (PC1) and Principal Component 2 (PC2). Despite variations in the absolute coefficient values compared to Figure 8, consistent patterns are observed. Notably, features such as Vapor Pressure Deficit (VPD) and air temperature (Tair) prominently contribute to the variance captured by PC1. Concurrently, Turbulence (UStar), and Soil water content (SWC) exhibit more subdued impacts on PC1. Shifting the focus to PC2, a discernible trend emerges. Soil temperature (Tsoil) and relative humidity (rH) emerge as pivotal contributors, playing substantial roles in shaping the variability encapsulated by this principal component. In contrast, Vapor Pressure Deficit (VPD) and Soil water content (SWC) display diminished contributions to PC2.

Top of Form

When integrating Evapotranspiration (ET) into the model alongside other independent variables, the resulting dataframe is meticulously presented in the subsequent table.

Table : Features contributing to the principal components while considering ET.

|  |  |  |
| --- | --- | --- |
| index | PC1 | PC2 |
| Rg | 0.726318 | -0.1485 |
| Tair | 0.848983 | 0.27913 |
| Tsoil | 0.666211 | 0.516566 |
| rH | -0.51083 | 0.655268 |
| VPD | 0.922505 | -0.23464 |
| Ustar | 0.303662 | 0.0079 |
| SWC | -0.36636 | 0.372802 |
| ET | 0.772574 | 0.138594 |

As delineated in Table 9, a meticulous examination of the absolute coefficients unveils distinct influences exerted by various features on Principal Component 1 (PC1) and Principal Component 2 (PC2). While there exist variations in the absolute coefficient values compared to those presented in Figures 8 and 9, discernible and consistent patterns persist. Notably, features such as Vapor Pressure Deficit (VPD), and Air Temperature (Tair) prominently contribute to the variance encapsulated by PC1. Simultaneously, Turbulence (UStar) and Soil Water Content (SWC) exhibit more subdued impacts on PC1. Shifting the analytical focus to PC2 reveals a noteworthy trend. Relative Humidity (rH) and Soil temperature (Tsoil) emerge as pivotal contributors, playing substantial roles in shaping the variability encapsulated by this particular principal component. In contrast, Turbulence (UStar) and Evapotranspiration (ET) display diminished contributions to PC2, underscoring their comparatively reduced influence in shaping the observed variability.

# 3.3 Data Exploration:

# Predictive Modeling of EVI and LSWI:

For the data Exploration part of the 1st Objective, we take the rolling sum of the “Rain” variable of 7 days, 14 days, 21 days, and 28 days and draw the correlation plot of all the variables to see how much each variable correlates to each of the variables. Figure 6 shows the correlation analysis plot.

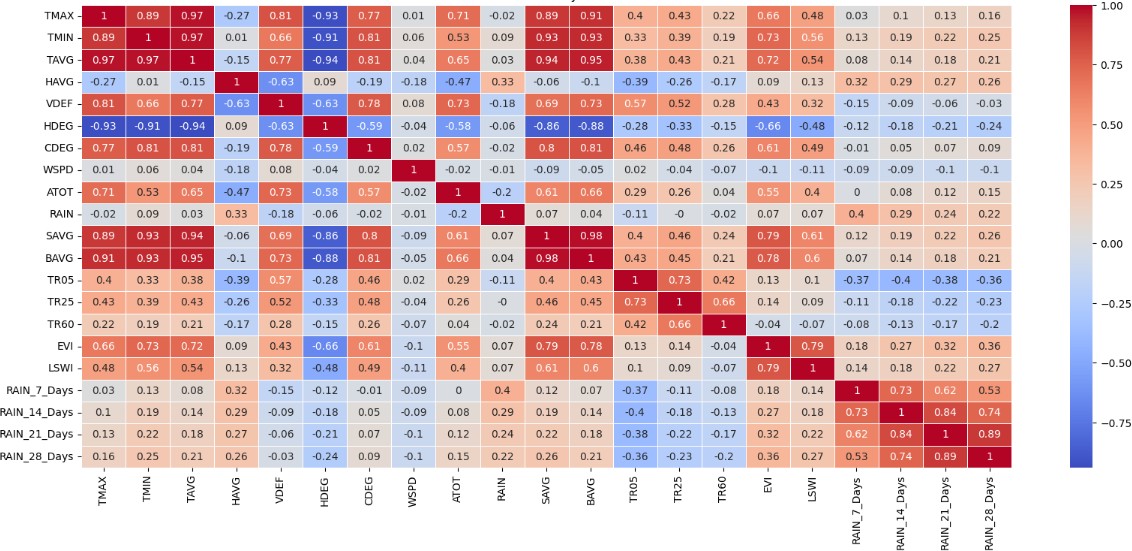


Figure :Correlation Analysis Plot for objective 1

From this correlation plot of Figure 6, it can be seen that even though variables such as Heating Degree days (HDEG) form moderate to strong correlations with other variables, surprisingly “Rain” does not form a lot of correlations with other variables, especially considering that this is not an irrigated field.

Now, Let’s look at exploratory Data Analysis of some important variables.

A group of blue graphs

Description automatically generated

Figure : Histogram of independent variables for objective 1

Figure 7 shows the histogram of independent variables of objective 1 where it is quite evident that columns such as Cooling Degree Days, Rainfall, Soil moisture at depths of 5 cm and 25 cm, and Rolling sum of rain for 7 days, 14 days, 21 days, 28 days has been identified as highly positively skewed columns as they have the skewness values of more than 1, which could impact the performance of the models. This issue has been handled in the data preparation stage.

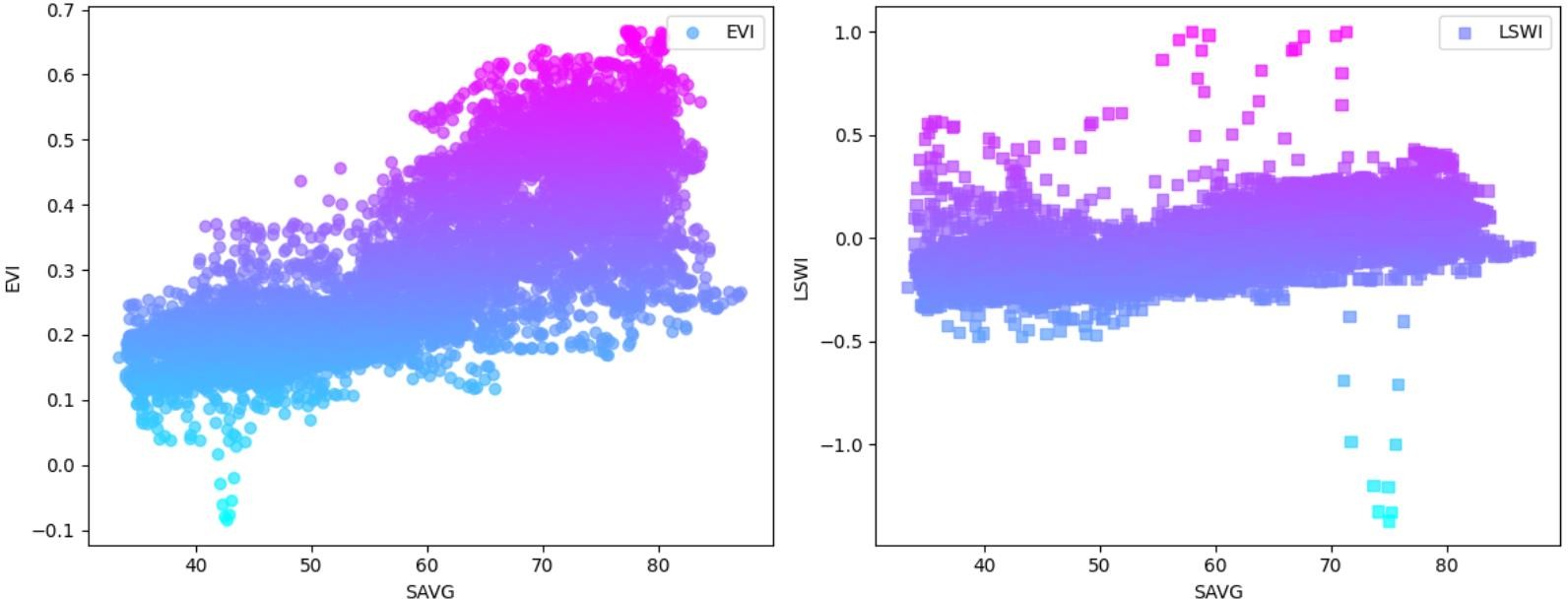


Figure : Scatter plot of SAVG for EVI and LSWI

Figure 8 shows how the average soil temperature of 10cm under sod (SAVG) affects the target variables enhanced vegetation index (EVI) and Land surface water index (LSWI) over the years. It is evident that when the average soil temperature of 10cm under sod increases, then EVI increases linearly, and LSWI increases monotonically.

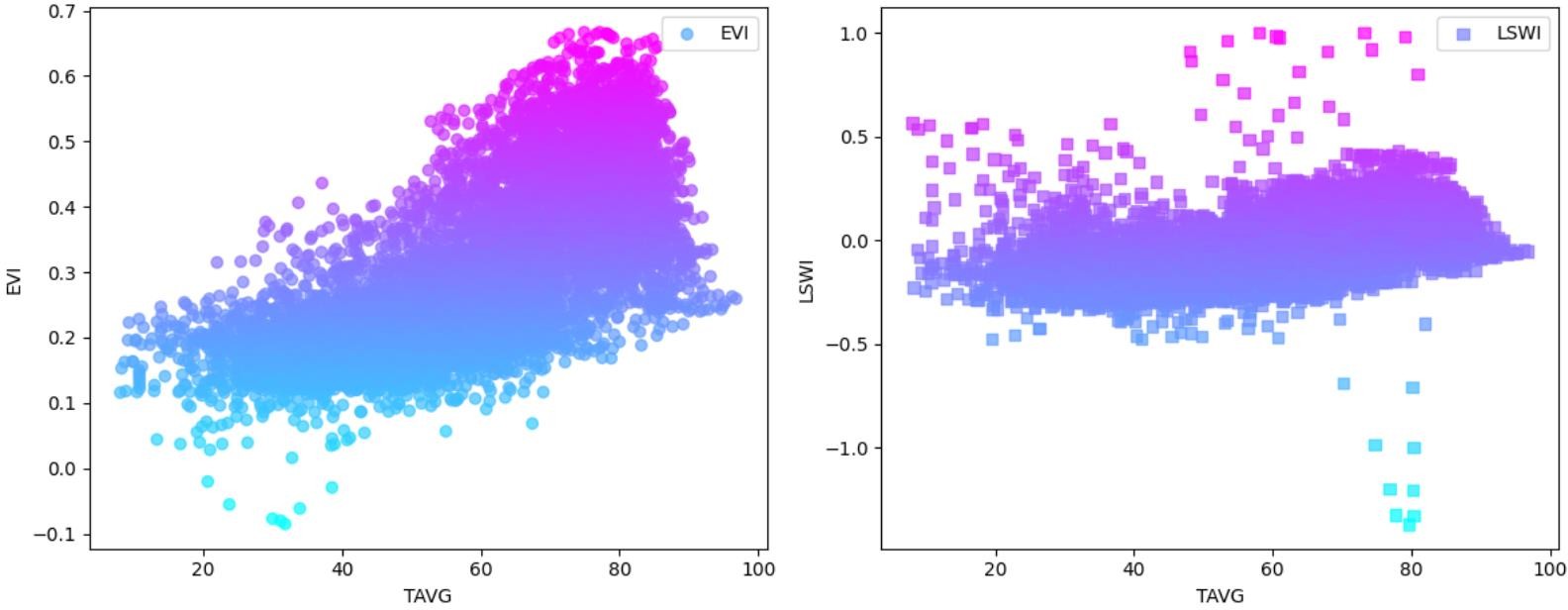


Figure : Scatter Plot of TAVG for both EVI and LSWI

Figure 9 shows how average temperature (TAVG) affects the target variables' enhanced vegetation index (EVI) and Land surface water index (LSWI) over the years. Likewise, as we have seen in the case of average soil temperature, it is also evident that when average temperature increases, then EVI increases linearly, and LSWI increases monotonically.

# Predictive Modeling of ER:

A screenshot of a graph

Description automatically generated

Figure : Correlation Plot for the objective 2

From Figure 10, it can be seen that Relative Humidity forms a moderate positive correlation with two different sensor measurements of Gross Primary Production and Evapotranspiration while it forms a moderate negative correlation with Net Ecosystem Exchange. Similarly, Air temperature also forms a moderate positive correlation with Vapor Pressure Deficit, while it forms a very strong positive correlation with Soil temperature.

A group of blue bars

Description automatically generated

Figure : Histogram Analysis for Objective 2

Figure 11 illustrates the histogram of independent variables about objective 2. Notably, discernible skewness is observed in columns such as Vapor Pressure Deficit, three distinct sensor measurements of Gross Primary Production, Net Ecosystem Exchange, and evapotranspiration. These columns exhibit a skewness value exceeding 1, indicating a highly positive skewness. Recognizing the potential impact of such skewness on model performance, we addressed this issue in the same as the approach employed during the data preparation phase of objective 1.

A screenshot of a graph

Description automatically generated

Figure : Scatter Plot of ER with both rH and VPD

Figure 12 shows how Relative Humidity and Vapor Pressure Deficit (TAVG) affects the target variables' Ecosystem Respiration (ER). It is quite evident from the scatter plot of Relative Humidity and Ecosystem Respiration that even though on average, the value of relative humidity increases ecosystem respiration remains almost or monotonically increasing, and similarly from the scatter plot of Vapor Pressure Deficit and Ecosystem Respiration, it is noticeable that as the amount of vapor pressure deficit increases, ecosystem respiration increases almost in a linear fashion.

# Methodology:

# 4.1 Techniques:

We employ the same methodology for our objectives 1 and 2. After all the data preparation is done, both sub-divided datasets have been further partitioned into training and validation sets, with an 80% portion allocated for training and 20% for validation. We have tuned the hyperparameter for each of the models through 5-fold cross-validation using GridSearchCV on the training datasets. Hyperparameter tuning generally involves adjusting the hyperparameters of a model to optimize its performance. It is essential for the model training because it helps to improve the model's ability to generalize to new, unseen data by finding the best set of hyperparameter values that minimize errors or maximize accuracy. Also, cross-validation is a technique used in machine learning to assess how well a predictive model will generalize to an independent dataset. It involves partitioning the available data into subsets, training the model on some of these subsets, and evaluating its performance on the remaining data. It is used to avoid overfitting and underfitting by providing a more robust estimate of a model's performance and ensuring it generalizes well to unseen data.

Specifically, the models employed in this project include Linear Regression, Support Vector Regressor, Random Forest Regressor, and XGBoost Regressor. These algorithms are chosen because of their strong interpretability, versatility, and capability to capture interactions and complex relationships in the data. Linear Regression is chosen because of its simplicity as well as interpretability and it is strongly recommended to use this model as a baseline model specifically for agricultural data, while Support Vector Regressor is known for its versatility and its ability to handle high-dimensional data, finding optimal hyperplanes for separating data points. Random Forest and XGBoost are very effective ensemble models that can capture interactions and complex relationships in the data.

We will utilize the validation datasets to assess these models' performance. In this evaluation, we will employ a range of performance metrics, including R-squared, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). It's important to highlight that, among these metrics, R-squared has been designated as the primary evaluation metric aligned with our project's overarching objectives. This choice underscores the significance of the R-squared metric in gauging the predictive power and goodness-of-fit of the models under consideration. The results of the model’s performance are shown in the next section.

# Procedures:

Below Figure 13 provides a flowchart showing the overall process.

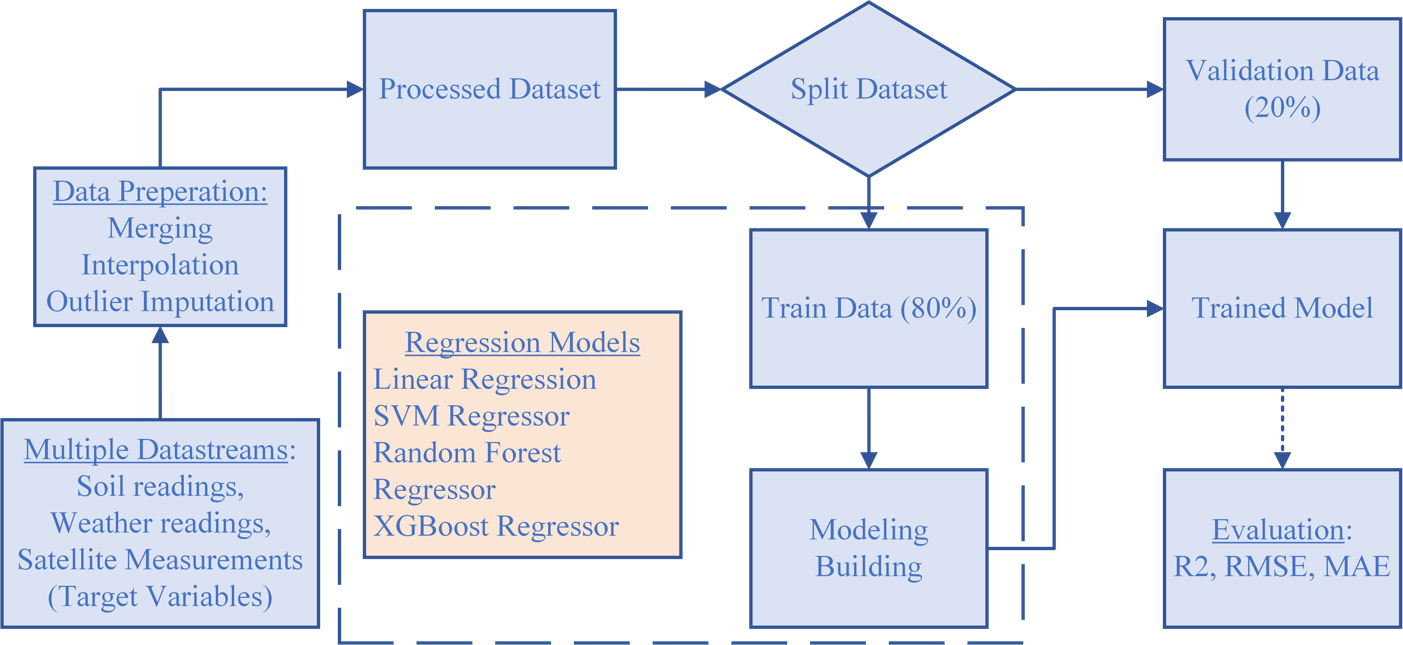


Figure : Flowchart of Overall Process

The process employed in this project was meticulously crafted to systematically address the critical issue of global food security through the lens of sustainable agriculture, with the explicit aim of accomplishing the predefined objectives. Throughout the project lifecycle, the acquired skills from core courses in Data Science and Analytics (DSA) were judiciously applied, leveraging a comprehensive understanding of data-driven methodologies. This deliberate integration of theoretical knowledge into practical applications served as the backbone for the successful execution of the project, contributing to a robust and informed approach to tackling the complex challenges associated with global food security.

The project commenced with data collection, where two datasets were obtained from the United States Department of Agriculture(USDA) for the first objective. The first dataset contains the weather and soil variables taken from a single pasture, whereas the second dataset contains the satellite measurements of 8-day intervals along with the target variables such as Enhanced Vegetation Index(EVI) and Land surface water index(LSWI).

For objective 1 of the data ingestion part, first, all the weather and soil variables are combined to have one dataset, and then in the other dataset, 8-day composite satellite measurements are transformed into daily averages by performing Linear interpolation methodologies. Then both of these datasets are merged using the right join to have a complete dataset.

In the data preparation phase of objective 1, it was quite evident that there were some columns which have values more than -50, and those values are considered noisy readings according to our stakeholders. If the column has more than 50% of values containing -50, those columns have been dropped, and if the column has less than 50% of values containing -50 values, those noisy readings have been imputed by Linear Interpolation methodologies to get the more reliable estimates. Also, outliers were present in some of the columns, which were identified through the Z-score methodologies. Even though we detected some outliers we did not remove those outliers as they were not extreme outliers, and we wanted to preserve as much as information possible to make the machine learning models more robust. After handling the outliers, through the histogram analysis, it also became quite clear that there were some columns present with highly positive skewness, which could impact the performance of the models. To mitigate this issue, logarithmic transformations were performed on those columns so that those columns could have a more normalized distribution, stabilize the variance, and have much more reduced skewness values. In addition to this, according to our stakeholders, we have introduced four new features by taking a rolling sum of Rain variables for the 7 days, 14 days, 21 days, and 28 days respectively. We follow two approaches for our analysis part and in the first approach the dataset is divided into two parts one contains the target variable EVI and the other contains the target variable LSWI. We then incorporated a feature selection algorithm such as the Recursive Feature Elimination algorithm for both of the datasets individually to select the top 8 features which would be essential for the analysis of our models as well as for our sponsors. After selecting the top 8 features, standardization methodologies were also introduced on the independent variables for both of the datasets to re-scale the features to a uniform scale, values ranging from 0 to 1. Furthermore, principal components analysis (PCA) was also applied to both of the datasets to see the number of components that contain 95% variance of the data. In the second approach, the Feature selection algorithm is not included in selecting the top 8 features, we just apply PCA on all of our independent variables of the overall dataset to see the number of components that carry 95% variance of the data. In both of the approaches, these selected components are fed into our models.

In the context of Objective 2, on data ingestion, it is noteworthy that the environmental variables encompassing weather and soil attributes differ from those utilized in Objective 1. Importantly, the data integration process transpired seamlessly as the relevant datasets were directly imported into the Google Colab Notebook without encountering any complications.

In the data preparation phase of Objective 2, we observed missing values in the target variables, prompting the removal of corresponding rows to maintain data integrity. Additionally, certain columns exhibited noisy readings, specifically denoted by the value -9999 in individual rows. To address this issue, rows with such outliers were excluded from the dataset. The data preprocessing procedures adhered to the methodologies employed in Objective 1, encompassing the handling of missing values, outlier detection, and feature scaling using standardized techniques. We have considered each of the three sensor measurements individually along with others independently, where PCA was applied separately to reveal the number of components required to capture 95% of the data variance. A detailed exploration of the first two principal components provided insights into the contributions of each feature. The resultant components from the distinct PCA analyses played a pivotal role in subsequent predictive modeling. These components were used as input for predictive models, each constructed by considering one of the three sensor measurements at a time.

To construct predictive models for both objectives 1 & 2, supervised regression techniques were applied to the training data to predict continuous values in both target variables. Specifically, implementations of Linear Regression, Support Vector Machine (SVM), Random Forest, and XGBoost from the Scikit-learn library were employed. To ensure robust model performance, cross-validation was integrated during the training phase to mitigate the risks of underfitting and overfitting. The project embraced a methodical approach, characterized by iterative experimentation and refinement. This included the crucial step of hyperparameter tuning to optimize the efficacy of the machine-learning models. Multiple iterations were meticulously executed, engaging in a process of fine-tuning the models to discern the most effective approach.

The comprehensive workflow of the supervised machine learning algorithms, elucidated above, is visually depicted in Figure 10. Evaluation of each algorithm's performance was conducted employing a spectrum of regression metrics, notably including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R²). For this project, according to our stakeholders, R-squared has been chosen as the primary metric. This rigorous assessment framework served as a basis for discerning the strengths and weaknesses of each model, ultimately facilitating the selection of the most adept approach for achieving the project objectives.

Throughout the project's lifecycle, we strategically applied skills acquired from core Data Science and Analytics (DSA) courses. This encompassed a comprehensive range of tasks, including adept data preprocessing, in-depth data exploration, sophisticated feature engineering, and meticulous model evaluation. The techniques assimilated from specialized courses, particularly those focusing on machine learning, and intelligent data analytics, played a pivotal role in shaping our approach.

An integral aspect of our methodology was the commitment to robust documentation. The entire process, from the application of specific techniques to the nuanced lessons learned, was diligently recorded. This documentation not only ensures transparency in our decision-making but also serves as a valuable repository of knowledge, providing insights into the intricacies of the project's evolution.

In conclusion, the project adhered to a systematic and methodical approach. It seamlessly integrated data preprocessing, feature engineering, and model training, leveraging supervised machine learning algorithms most effectively. The application of skills from core DSA courses was not only evident but played a pivotal role in the project's success. Clear delineation of responsibilities further underscored the project's organizational structure. A detailed presentation of the results and findings will be expounded upon in a subsequent section of this comprehensive project report.

# Results and Analysis:

# Predictive Modeling of EVI and LSWI:

For Objective 1, we addressed the problem through two distinct approaches for each target variable: (i) Feature Selection + PCA + Modeling, with or without hyperparameters, and (ii) PCA + Modeling, with or without hyperparameters. Model performance is assessed using the validation dataset."

Table 10 shows the different models that are implemented along with the parameters used and their options. The selection column of the table shows the best model parameters.

Table : Hyperparameter Tuning for EVI

|  |  |  |
| --- | --- | --- |
| **Model Name** | **Hyperparameter Tuning** | **Selection** |
| Linear Regression | N/A | N/A |
| Support Vector Regressor | C = [0.1, 1, 10]  gamma = [0.01, 0.1, 1]  Kernel = [‘rbf’, ‘sigmoid’] | C = 1, gamma = 0.1  Kernel = rbf |
| Random Forest Regressor | n\_estimators = [50,100,200,300,500],  max\_depth = [3,4,5,6],  min\_samples\_split = [2,3,4 5, 6] | n\_estimators = 200  max\_depth = 6  min\_samples\_split = 2 |
| XGBoost Regressor | n\_estimators = [50,100,200,300,500],  max\_depth: [2,3,4,5],  learning\_rate: [0.01, 0.1, 0.2],  colsample\_bytree: [0.8, 0.9, 1.0] | n\_estimators = 300  max\_depth = 5  learning\_rate = 0.1  colsample\_bytree = 1.0 |

Table 11 shows the results of the target variable EVI without hyperparameter tuning.

Table : Results of the target variable EVI without hyperparameter tuning.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Approach – 1 (PCA) | | | Approach – 2 (FS+PCA) | | |
| Model Name | RMSE | MAE | R-Squared | RMSE | MAE | R-Squared |
| Linear Regression | 0.07 | 0.06 | 0.71 | 0.08 | 0.06 | 0.64 |
| Support Vector Regressor | 0.06 | 0.05 | 0.76 | 0.07 | 0.05 | 0.73 |
| Random Forest Regressor | 0.06 | 0.04 | 0.79 | 0.06 | 0.04 | 0.77 |
| XGBoost Regressor | 0.06 | 0.05 | 0.77 | 0.07 | 0.06 | 0.76 |

Table 12 shows the results of the target variable EVI with hyperparameter tuning.

Table : Results of the target variable EVI with hyperparameter tuning.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Approach – 1 (PCA) | | | Approach – 2 (FS+PCA) | | |
| Model Name | RMSE | MAE | R-Squared | RMSE | MAE | R-Squared |
| Linear Regression (Without Tuned) | 0.07 | 0.06 | 0.71 | 0.08 | 0.06 | 0.64 |
| Support Vector Regressor | 0.06 | 0.05 | 0.78 | 0.07 | 0.05 | 0.75 |
| Random Forest Regressor | 0.06 | 0.04 | 0.80 | 0.05 | 0.04 | 0.78 |
| XGBoost Regressor | 0.03 | 0.02 | 0.81 | 0.04 | 0.03 | 0.79 |

Table 13 shows the results of the target variable LSWI without hyperparameter tuning.

Table : Results of the target variable LSWI without hyperparameter tuning.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Approach – 1 (PCA) | | | Approach – 2 (FS+PCA) | | |
| Model Name | RMSE | MAE | R-Squared | RMSE | MAE | R-Squared |
| Linear Regression | 0.10 | 0.08 | 0.56 | 0.11 | 0.08 | 0.48 |
| Support Vector Regressor | 0.09 | 0.07 | 0.65 | 0.10 | 0.09 | 0.61 |
| Random Forest Regressor | 0.09 | 0.06 | 0.66 | 0.10 | 0.06 | 0.64 |
| XGBoost Regressor | 0.10 | 0.07 | 0.64 | 0.09 | 0.08 | 0.62 |

Table 14 shows the results of the target variable LSWI with hyperparameter tuning.

Table : Results of the target variable LSWI with hyperparameter tuning.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Approach – 1 (PCA) | | | Approach – 2 (FS+PCA) | | |
| Model Name | RMSE | MAE | R-Squared | RMSE | MAE | R-Squared |
| Linear Regression (Without Tuned) | 0.10 | 0.08 | 0.56 | 0.11 | 0.08 | 0.48 |
| Support Vector Regressor | 0.08 | 0.06 | 0.67 | 0.08 | 0.09 | 0.63 |
| Random Forest Regressor | 0.06 | 0.04 | 0.72 | 0.07 | 0.05 | 0.70 |
| XGBoost Regressor | 0.04 | 0.03 | 0.69 | 0.08 | 0.04 | 0.67 |

In the context of predicting the target variable EVI, the XGBoost emerges as the superior model, showcasing heightened performance relative to its counterparts. Conversely, in predicting the target variable LSWI, the Random Forest Regressor takes precedence, exhibiting superior performance compared to alternative models. Now we are plotting the regression plot for the best models of each of the target variables.

**Regression plot Analysis for the target variable EVI:**

While predicting the target variable EVI, it is noticeable that the XGBoost Regressor model works better than the other two models. So, in the case of EVI, we are plotting the regression plot for the XGboost Regressor, as shown in Figure 14.

A diagram of a plot

Description automatically generated

Figure : Regression plot of XGBoost for target variable EVI

Figure 14 illustrates a noteworthy observation: the R-squared value for the best XGBoost Model attaining 0.81 underscores the close alignment of data points with the diagonal line. This substantial correlation coefficient serves as a quantitative indicator of the model's adeptness in capturing the variance present in the data. The proximity of the points to the diagonal line, as indicated by the high R-squared value, signifies a robust fit and underscores the XGBoost Model's efficacy in explaining the variability observed in the target variable EVI.

**Regression plot Analysis for the target variable LSWI:**

In the prediction of the target variable LSWI, a distinct superiority is observed in the performance of the Random Forest Regressor when compared to the other models. Consequently, for the case of LSWI, we are plotting the regression plot for the Random Forest Regressor, as shown in Figure 15.

A diagram of a graph

Description automatically generated

Figure : Regression plot of Random Forest for target variable LSWI

Figure 15 distinctly reveals a compelling observation: the R-squared value of 0.72 for the optimal Random Forest Model attests to the proximity of the data points to the diagonal line. This alignment signifies a robust fit of the model to the observed data, underscoring the effectiveness of the Random Forest Model in capturing the underlying patterns associated with the target variable LSWI.

**Feature Importance:**

The extraction of the top 5 features serves a dual purpose. Primarily, this endeavor seeks to discern the features that wield the greatest influence in predicting the target variables, providing valuable insights into the underlying dynamics of the predictive models. Furthermore, this succinct selection of prominent features serves the practical objective of facilitating stakeholders in identifying and understanding the specific variables that exert a substantial impact on the predicted outcomes. Such clarity is integral for informed decision-making and a comprehensive understanding of the factors driving the predicted values in our models.

A graph of a graph

Description automatically generated with medium confidence

Figure : Top 5 features extracted from XGBoost for target variable EVI.

With XGBoost established as the optimal model for predicting the target variable EVI, a focused analysis on feature importance reveals that the top 5 contributing features are Average Daily Vapor Deficit (VDEF), Solar Radiation (ATOT), Average Soil Temperature measured at 10 cm under sod (SAVG), Soil Moisture at a depth of 25cm (TR25), and 60cm (TR60). The associated feature importance is visually depicted in Figure 16. Notably, within this set of influential features, the Average Daily Vapor Deficit (VDEF) emerges as the most critical determinant, bearing the highest feature importance in the prediction of the target variable EVI. This insight underscores the significance of considering these specific environmental variables in the context of EVI prediction and affords valuable guidance for understanding the underlying dynamics driving the model's predictive performance.

A graph of blue rectangular bars

Description automatically generated

Figure : Top 5 features extracted from Random Forest for target variable LSWI.

As, the best model for predicting the target variable LSWI is identified as the Random Forest, so, an exploration into the top 5 features extracted from this model reveals key contributors, namely Cooling Degree Days (CDEG), Average Soil Temperature measured at 10 cm under sod (SAVG), Avg Soil Temp 10 cm Under Bare Soil (BAVG), Average temperature(TAVG), and minimum temperature (TMIN), as delineated in Figure 17. Also, it is noticeable that within this subset of features, the average daily Vapor Deficit emerges as the most influential determinant in predicting the target variable LSWI. This discernment underscores the pivotal role of specific climatic factors, particularly vapor deficit, in shaping the predictive prowess of the Random Forest model for LSWI.

# Predictive Modeling of ER:

For objective 2, We have predicted the ER using the PCA+modeling approach with or without hyperparameter tuning by considering 3 different sensor measurements(GPP, NEE, ET) one at a time.

**Modeling for GPP:**

Table 15 shows the different models that are implemented along with the parameters used and their options while considering only GPP. The selection column of the table shows the best model parameters.

Table : Hyperparameter Tuning for GPP Modeling.

|  |  |  |
| --- | --- | --- |
| **Model Name** | **Hyperparameter Tuning** | **Selection** |
| Linear Regression | N/A | N/A |
| Support Vector Regressor | C = [0.1, 1, 10]  gamma = [0.01, 0.1, 1]  Kernel = [‘rbf’, ‘sigmoid’] | C = 10, gamma = 1  Kernel = rbf |
| Random Forest Regressor | n\_estimators = [50,100,200,300,500,1000],  max\_depth = [3,4,5,6],  min\_samples\_split = [2,3,4 5,6] | n\_estimators = 1000  max\_depth = 6  min\_samples\_split = 2 |
| XGBoost Regressor | n\_estimators = [50,100,200,300,500,1000],  max\_depth: [2,3,4,5],  learning\_rate: [0.01, 0.1, 0.2],  colsample\_bytree: [0.8, 0.9, 1.0] | n\_estimators = 1000  max\_depth = 5  learning\_rate = 0.1  colsample\_bytree = 1.0 |

Table 16 shows the results of the target variable ER without hyperparameter tuning using GPP.

Table : Results of ER without using hyperparameter tuning considering GPP.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | RMSE | MAE | R-Squared |
| Linear Regression | 1.97 | 1.34 | 0.53 |
| Support Vector Regressor | 1.75 | 0.97 | 0.64 |
| Random Forest Regressor | 1.59 | 0.81 | 0.69 |
| XGBoost Regressor | 1.73 | 1.03 | 0.65 |

Table 17 shows the results of the target variable ER without hyperparameter tuning using GPP.

Table : Results of ER using hyperparameter tuning considering GPP.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | RMSE | MAE | R-Squared |
| Linear Regression(Without Tuned) | 1.97 | 1.34 | 0.53 |
| Support Vector Regressor | 1.55 | 0.72 | 0.71 |
| Random Forest Regressor | 1.48 | 0.76 | 0.73 |
| XGBoost Regressor | 1.70 | 0.96 | 0.67 |

**Modeling for NEE:**

Table 18 shows the different models that are implemented along with the parameters used and their options while considering only NEE. The selection column of the table shows the best model parameters.

Table : Hyperparameter Tuning for NEE Modeling.

|  |  |  |
| --- | --- | --- |
| **Model Name** | **Hyperparameter Tuning** | **Selection** |
| Linear Regression | N/A | N/A |
| Support Vector Regressor | C = [0.1, 1, 10]  gamma = [0.01, 0.1, 1]  Kernel = [‘rbf’, ‘sigmoid’] | C = 10, gamma = 0.1  Kernel = rbf |
| Random Forest Regressor | n\_estimators = [50,100,200,300,500,1000],  max\_depth = [3,4,5,6],  min\_samples\_split = [2,3,4 5,6] | n\_estimators = 1000  max\_depth = 6  min\_samples\_split = 3 |
| XGBoost Regressor | n\_estimators = [50,100,200,300,500,1000],  max\_depth: [2,3,4,5],  learning\_rate: [0.01, 0.1, 0.2],  colsample\_bytree: [0.8, 0.9, 1.0] | n\_estimators = 1000  max\_depth = 5  learning\_rate = 0.1  colsample\_bytree =0.9 |

Table 19 shows the results of the target variable ER without hyperparameter tuning using NEE.

Table : Results of ER without using hyperparameter tuning considering NEE.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | RMSE | MAE | R-Squared |
| Linear Regression | 1.93 | 1.32 | 0.57 |
| Support Vector Regressor | 1.14 | 0.70 | 0.84 |
| Random Forest Regressor | 1.02 | 0.61 | 0.87 |
| XGBoost Regressor | 1.11 | 0.74 | 0.85 |

Table 20 shows the results of the target variable ER with hyperparameter tuning using NEE.

Table : Results of ER using hyperparameter tuning considering NEE.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | RMSE | MAE | R-Squared |
| Linear Regression(Without Tuned) | 1.93 | 1.32 | 0.57 |
| Support Vector Regressor | 1.07 | 0.66 | 0.86 |
| Random Forest Regressor | 0.85 | 0.43 | 0.91 |
| XGBoost Regressor | 0.98 | 0.62 | 0.88 |

**Modeling for ET:**

Table 21 shows the different models that are implemented along with the parameters used and their options while considering only ET. The selection column of the table shows the best model parameters.

Table : Hyperparameter Tuning for ET Modeling.

|  |  |  |
| --- | --- | --- |
| **Model Name** | **Hyperparameter Tuning** | **Selection** |
| Linear Regression | N/A | N/A |
| Support Vector Regressor | C = [0.1, 1, 10]  gamma = [0.01, 0.1, 1]  Kernel = [‘rbf’, ‘sigmoid’] | C = 10, gamma = 1  Kernel = rbf |
| Random Forest Regressor | n\_estimators = [50,100,200,300,500,1000],  max\_depth = [3,4,5,6],  min\_samples\_split = [2,3,4 5,6] | n\_estimators = 500  max\_depth = 5  min\_samples\_split = 2 |
| XGBoost Regressor | n\_estimators = [50,100,200,300,500,1000],  max\_depth: [2,3,4,5],  learning\_rate: [0.01, 0.1, 0.2],  colsample\_bytree: [0.8, 0.9, 1.0] | n\_estimators = 1000  max\_depth = 4  learning\_rate = 0.01  colsample\_bytree = 0.9 |

Table 22 shows the results of the target variable ER without hyperparameter tuning using ET.

Table : Results of ER without using hyperparameter tuning considering ET.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | RMSE | MAE | R-Squared |
| Linear Regression | 1.99 | 1.35 | 0.52 |
| Support Vector Regressor | 1.77 | 1.01 | 0.63 |
| Random Forest Regressor | 1.63 | 0.84 | 0.67 |
| XGBoost Regressor | 1.72 | 0.99 | 0.64 |

Table 23 shows the results of the target variable ER without hyperparameter tuning using ET.

Table : Results of ER using hyperparameter tuning considering ET.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | RMSE | MAE | R-Squared |
| Linear Regression(Without Tuned) | 1.99 | 1.35 | 0.52 |
| Support Vector Regressor | 1.74 | 0.97 | 0.66 |
| Random Forest Regressor | 1.57 | 0.74 | 0.70 |
| XGBoost Regressor | 1.70 | 0.96 | 0.65 |

The **above-presented** results underscore a discernible enhancement in model performance following the application of hyperparameter tuning. Notably, across three distinct sensor measurements, the Random Forest Regressor consistently outperforms alternative models, as evidenced by the superior R-squared values— **the primary** metric in our evaluation. Consequently, the Random Forest Regressor is identified as the optimal model for the given task.

Further analysis reveals a noteworthy observation: when leveraging the sensor measuring Net Ecosystem Exchange (NEE), the model exhibits superior predictive p**ower** in **predicting** Ecosystem Respiration (ER). This performance disparity is conspicuous when contrasted with the outcomes derived from the sensors measuring Gross Primary Production (GPP) and Evapotranspiration (ET).

**Regression plot Analysis for the target variable ER:**

In predicting the target variable ER, a notable superiority is evident in the performance of the Random Forest Regressor relative to other models and using sensors measuring Net Ecosystem Exchange (NEE). Accordingly, for the case of ER, a regression plot is presented exclusively for the Random Forest Regressor, considering the influence of NEE in the modeling process, as illustrated in Figure 18.

A graph with blue dots and red line

Description automatically generated

Figure : Regression plot of Random Forest for target variable ER using NEE.

In forecasting the target variable ER, a notable superiority is evident in the performance of the Random Forest Regressor relative to other models while sensors measuring Net Ecosystem Exchange (NEE). Accordingly, for the case of ER, a regression plot is presented exclusively for the Random Forest Regressor, considering the influence of NEE in the modeling process, as illustrated in Figure 18.

**Feature Importance:**

As the sensor measuring NEE shows better results and the Random Forest outperforms other models, we have extracted the top 5 features from the Random Forest Regressor so that our stakeholders could understand which particular features are beneficial towards predicting the target variable ER.

A graph of blue squares

Description automatically generated

Figure : Top 5 features extracted from Random Forest Regressor for target variable ER using NEE.

The optimal model for predicting the target variable ER has been identified as the Random Forest. Upon delving into the top 5 features extracted from this model, significant contributors emerge, notably Turbulence (UStar), Solar Radiation (Rg), Vapor Pressure Deficit (VPD), an interactive term represented by Tair\*VPD, and Soil Water Content (SWC), as elucidated in Figure 18. Noteworthy is the observation that within this subset of features, Turbulence stands out as the most influential determinant in predicting the target variable ER, leveraging the sensor measurements of NEE. This discernment accentuates the pivotal role of specific climatic factors, particularly Turbulence, in shaping the predictive efficacy of the Random Forest model for ER.

# Deliverables:

In pursuit of the first objective, we have successfully developed predictive models to predict the target variables—Enhanced Vegetation Index and Land Surface Water Index. The predictive power of these models holds significant implications for the United States Department of Agriculture (USDA), offering a robust mechanism to monitor crop conditions and estimate potential yields. This predictive insight is invaluable for informed agricultural planning, optimal resource allocation, and the initiative-taking management of risks associated with crop production. The utility extends further as the models predicting the Land Surface Water Index empower the USDA to monitor drought conditions and assess water availability. Such critical information serves as the cornerstone for the implementation of effective water management strategies, especially in regions prone to drought. By leveraging these predictive models, the USDA can cultivate sustainable agricultural practices, ensuring resilience in the face of evolving environmental dynamics. Additionally, a pivotal outcome of our endeavor is the identification of optimal features from the predictive models. This feature selection provides stakeholders with key insights, elucidating the variables most instrumental in predicting the Enhanced Vegetation Index and Land Surface Water Index. This knowledge facilitates a nuanced understanding of the contributing factors, empowering stakeholders to make informed decisions based on the most impactful features. Transitioning to the second objective, we have crafted predictive models aimed at forecasting Ecosystem Respiration using data from three distinct sensors. Noteworthy from our findings is the superior performance of the sensor measuring Net Ecosystem Exchange in predicting Ecosystem Respiration. This discernment holds practical significance as it allows us to confidently recommend the adoption of this specific sensor, presenting a cost-effective alternative to utilizing all three sensors. This recommendation is poised to generate substantial cost savings for stakeholders. Moreover, the prediction of Ecosystem Respiration plays a pivotal role in estimating the quantity of carbon released by ecosystems. This insight is foundational for carbon budgeting, offering stakeholders valuable information to navigate and optimize their environmental stewardship initiatives. The combination of accurate predictions and cost-effective recommendations aligns with the overarching goal of fostering sustainability and informed decision-making within the realm of ecosystem monitoring and management.

# References:

* <https://www.mdpi.com/2072-4292/6/10/10193>
* <https://www.tandfonline.com/doi/full/10.1080/22797254.2020.1786466>
* <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
* <https://builtin.com/machine-learning/pca-in-python>
* <https://machinelearningmastery.com/rfe-feature-selection-in-python/>
* <https://en.wikipedia.org/wiki/Vapour-pressure_deficit#:~:text=For%20this%20reason%20the%20ideal,between%200.8%20and%200.95%20kPa>.
* <https://stats.stackexchange.com/questions/104622/what-does-an-actual-vs-fitted-graph-tell-us>
* <https://www.geeksforgeeks.org/xgboost-for-regression/>

# Self-Assessment:

My learning objectives were --

* To utilize data science methodologies for solving real-world problems.
* To work with datasets that are related to agriculture.
* To improve my problem-solving and critical-thinking skills.

I am delighted to report the achievement of my learning objectives. The journey commenced with the identification of a real-world problem amenable to resolution through the application of data science principles. My project, which focused on predicting the Enhanced Vegetation Index, Land Surface Water Index, and Ecosystem Respiration based on weather and soil variables, served as a practical manifestation of these objectives.

To attain this goal, I harnessed crucial Data Science and Analysis (DSA) skills. An initial phase involved a meticulous dataset analysis, facilitating a comprehensive grasp of the information embedded within. Furthermore, I engaged in feature engineering, creating new variables to assess their potential impact on enhancing the accuracy of my predictive models.

The culmination of my efforts materialized through the application of various supervised machine learning regression techniques; a process instrumental in realizing the outlined project objectives. This endeavor provided me with valuable hands-on experience in deploying machine learning algorithms, contributing significantly to my proficiency in working with diverse ML models.

The accomplishment of the stipulated individual learning objectives necessitated a multifaceted skill set in Data Science and Analysis (DSA). Primarily, the ability to identify and articulate a real-world problem as a regression challenge underscored a deep understanding of regression algorithms. Additionally, navigating through the intricacies of the dataset required proficiency in essential DSA skills, including data cleaning, pre-processing, analysis, visualization, featurization, and modeling.

On a personal development front, I recognized the need to acquire specific skills tailored to the project's demands. Consequently, I undertook a focused learning path, delving into Principal Component Analysis (PCA) for feature dimensionality reduction, exploring Feature Selection algorithms such as Recursive Feature Elimination, and gaining a comprehensive understanding of agriculture-related terminology relevant to the project's domain. This dedicated pursuit of knowledge further enriched my skill set and contributed significantly to the successful completion of the project.