Analyzing the impact of environmental factors on vegetation and agriculture using Decision tree

Rashmika Seela
University of Florida
Group 19

Munish Tanwar
University of Florida
Group 19

***Abstract*—This project addresses the intricate relationships between vegetation, rainfall, and land temperature, aiming to unravel their profound impact on agriculture and ecosystem health. The significance of this work lies in optimizing agricultural practices for a sustainable food supply and guiding conservation efforts. The technical challenges involve deciphering non-linear interactions among environmental variables and developing models capable of capturing these complexities. Our approach integrates decision tree methodologies, focusing on vegetation dynamics, to provide a holistic analysis. This project also analyzes the importance of land temperature and rainfall over each month to capture the seasonal impact on vegetation. This project's importance is underscored by its potential to offer actionable insights for policymakers, agriculturists, and conservationists, contributing to the development of sustainable practices in the face of evolving climates.**

I. Introduction

In the realm of environmental dynamics and agriculture, our project embarks on a quest to unravel the intricate relationships between crucial factors that shape ecosystems. Focused on geospatial data analysis, we delve into the profound influences of vegetation, rainfall, and land surface temperature on the health and productivity of our vital ecosystems. By harnessing spatial raster data encompassing vegetation index, precipitation, and temperature, along with geographic features like lakes and rivers, we aim to construct a comprehensive decision tree model to decipher the correlations that underpin the vulnerability of specific regions to environmental changes. The

significance of our exploration lies not only in the scientific intricacies it unravels but also in its profound societal applications. As we navigate the labyrinth of environmental factors, our insights hold the potential to revolutionize agricultural practices, ensuring a resilient and abundant food supply for our growing global population. The meticulous analysis of rainfall, land temperature, and vegetation provides a nuanced understanding of regions susceptible to environmental shifts, thereby guiding strategic initiatives to fortify global food security. Moreover, our findings contribute to conservation endeavors, offering data-driven insights for reforestation plans that combat deforestation in vulnerable areas. Beyond agriculture, policymakers can leverage these revelations to craft eco-friendly land use and environmental policies, fostering a sustainable planet and an improved quality of life for all. In

examining the intricate relationship between vegetation and surface temperature, recent studies have provided valuable insights, particularly in the context of the Tokyo region during winter [1]. Utilizing remote sensing data and ground-based measurements, the authors of the first study employed regression analysis to unravel the impact of vegetation on both surface temperature and composition. Complementing this, the second study delved into crop yield prediction, incorporating weather data and NDVI time series through a time series analysis-based model [2]. These studies, while illuminating, are region-specific and season-bound, underscoring the need for a broader, more adaptable approach. Building upon these

foundations, the current approach to predicting vegetation involves an exhaustive data collection process, as evidenced by decision Trees [3]. Drawing from long-term crop yield data, along with qualitative and quantitative environmental factors such as tillage systems and land temperature [3][4], decision Trees offer a comprehensive analysis. However, this method's potential limitation lies in its ability to capture complex, non-linear relationships between environmental factors and crop yield. While embracing various decision Tree methods, including Classification and Regression Tree (CART) and Conditional Inference (CI) [3], there is room for exploring alternative machine learning models, such as ensemble methods or neural networks, to address the intricacies of this relationship more effectively. This introduction sets the stage for a deeper exploration of current methodologies, their merits, and the evolving landscape of predictive modeling in the realm of vegetation and environmental dynamics. Existing

methods for predicting vegetation index based on environmental factors often fall short when facing the intricacies of seasonal variations. Our novel Decision Tree Regression (DTR) approach tackles this challenge head-on by tailoring its models to each month of the year. Unlike static models that overlook the shifting dynamics between factors like land surface temperature, rainfall, and NDVI, DTR dynamically adapts its decision rules, capturing the unique seasonal interactions that influence vegetation health. This granular approach not only promises significantly more accurate predictions but also unveils the deeper interplay between environment and vegetation throughout the year, enriching our understanding of ecosystems and empowering informed decision-making in resource management.

II. Problem definition

A. Basic Concepts

NDVI (Normalized Difference Vegetation Index): A dimensionless satellite-derived metric ranging from -1 to 1, quantifying vegetation photosynthetic activity and greenness. Higher NDVI signifies denser vegetation and active photosynthesis [5]. Land Temperature (LST): Surface temperature of the land, directly measured or derived from satellite data, impacting evapotranspiration, plant growth, and ecosystem functioning [6]. Rainfall: Precipitation reaching the ground, influencing soil moisture, water availability for plants, and vegetation growth [7].

B. Formal Definition

This research aims to develop a robust and accurate model for predicting NDVI values in a specific region, based on historical and/or real-time data on land temperature and rainfall.

1) Input

We will use the dataset which will consist of real-time climate snapshots of the world in csv format of size 3600x1800. We will focus on monthly records from 2022.

NDVI Data:

- This map shows where and how much green vegetation is grown
- A map will have a region colored with an index value from -0.1 (light) to 0.9 (dark)

Land Temperature Data:

- This map depicts the current global distribution of land surface temperature in Celsius.
- Each region on the map is color-coded, with shades ranging from deep blue (-25°C) to vibrant yellow (+45°C). The color legend on the right illustrates the temperature range corresponding to each color.
- (Include code to embed your map image here)

Rainfall data:

- The map shows where and how much precipitation fell around the world in millimeters
- A map will have a colored region ranging from 1.0 to 2000 mm of rainfall.

2) Output

Predicted NDVI values for the target region Object:

The focus is on model development for prediction, not simulating the intricate biophysical processes governing NDVI (Wang et al., 2015).

C. Example

Consider a forest manager in the Amazon rainforest aiming to monitor deforestation and assess forest health. Ground-based NDVI measurements are impractical for vast areas, but land temperature and rainfall data are readily available from satellites. This research addresses the problem of building a model that accurately predicts NDVI in the rainforest based on these readily available data sources, enabling the forest manager to track deforestation, monitor forest health, and inform sustainable forest management practices.

III. Proposed solution

A. Overview

To understand the relationships between environmental factors and vegetation, we propose utilizing a Decision Tree Regression (DTR) model. This model will predict vegetation index (NDVI) based on land surface temperature (LST) and rainfall data. By analyzing the decision tree structure, we will gain insights into the specific rules and thresholds governing these relationships.

We will analyze the DTR structure for each month of 2022, allowing us to investigate:

- Temporal Variations: How the decision rules and feature importance change across different months, revealing seasonal impacts on vegetation.
- Correlation Analysis: How the relationships between LST, rainfall, and NDVI evolve over time, potentially identifying long-term trends or year-on-year fluctuations.

B. Major Steps

1) Data Acquisition and Preprocessing

- Read monthly NDVI, LST, and rainfall data globally in 2022.
- Handle missing values and ensure spatial alignment of dataframes.
- Flatten dataframes and combine LST and rainfall features into a feature matrix (X).
- Extract NDVI values as the target vector (y).

2) Decision Tree Training

next line Initialize a DTR model. Train the DTR model on the preprocessed data (X, y) for each month of 2022.

3) Model Evaluation and Analysis

- Evaluate the DTR performance on the test data for each month using Mean Squared Error (MSE) and R-squared R^2 metrics.
- Analyze the decision tree structure for each month to understand the specific rules and thresholds influencing NDVI.
- Identify the relative importance of LST and rainfall features in each month based on splitting rules.
- Investigate the temporal variations in the decision tree structure and feature importance across months, uncovering seasonal and year-on-year trends.

C. Improved Understanding with Seasonal Trends

While current methods effectively predict vegetation health based on environmental factors, they often overlook the crucial role of seasonal trends. Our proposed DTR approach will take these dynamics into account, enriching our understanding of how LST, rainfall, and NDVI interact throughout the year.

D. Applications and Outcomes

The interpretable insights from the DTR model will provide a clear understanding of the relationships between LST, rainfall, and NDVI. This knowledge can be used for:

- Commercial farmers: By analyzing DTR predictions tailored to specific months and seasons, farmers can optimize water resource allocation. Consider summer months, where land surface temperature maps pinpoint areas prone to heat stress. Armed with this knowledge, farmers can prioritize irrigation in these sections, ensuring optimal crop growth while minimizing water waste. Similarly, winter maps can identify orange groves at risk of frost damage, allowing proactive measures like frost covers or irrigation to be implemented.
- Vegetation monitoring: Seasonal variations in DTR predictions can highlight areas experiencing unexpected NDVI decline. For instance, winter NDVI plummeting in a typically stable evergreen forest could signal a potential disease outbreak or disturbance event. By focusing conservation efforts on such areas during specific seasons, proactive interventions can be implemented before widespread damage occurs.
- Climate change assessment: Monitoring how seasonal relationships between LST, rainfall, and NDVI shift over time can provide valuable insights into the long-term impact of climate change on vegetation health.
- Adaptive land management: By understanding how environmental factors influence vegetation across different seasons, we can develop strategies for adapting land management practices to changing environmental conditions and ensuring the long-term sustainability of our ecosystems.

This DTR-based approach offers a transparent and informative way to understand the complex relationships between environmental factors and vegetation. By analyzing the decision tree structure and temporal variations, we can gain valuable insights for sustainable land management and conservation practices in the study area.

E. Example

We conducted the analysis for each month of 2022. Here we will discuss the analysis for January 2022.

Dataset: We start with csv files for NDVI, Land Surface Temperature, and Rainfall. Each file is of size 3600x1800

Data pre-processing: We preprocess these data by converting them into a flattened array. We then created a data frame with columns NVDI, Land temperature, and rainfall. We dropped data where the values were undefined and the final data frame consisted of 1,132,116 data points. We used land temperature and rainfall columns for features and NDVI for the target

Number of Samples: The data is split into training and testing sets using an 80-20 ratio. With a training set of size 905,692 and a testing set of size 226,424.

Training the model: We train a decision tree regression model on the training data set.

IV. Evaluations

A. Goal

This experimental evaluation aims to assess the efficacy of a decision tree regression model in predicting Normalized Difference Vegetation Index (NDVI) based on land temperature and rainfall data. We further investigate the influence of seasonality on both the model's performance and the importance of environmental factors. Our specific research questions are:

- Model Accuracy: How accurately can the decision tree model capture the relationship between NDVI and land temperature and rainfall?
- Predictive Power: Can the model effectively predict NDVI values based on land temperature and rainfall data?
- Feature Importance: Which environmental factors play the most significant roles in influencing NDVI, as determined by feature importance analysis?
- Seasonal Variations: How do the relative contributions of land temperature and rainfall to NDVI vary across different seasons?

B. Method

We utilize one year (January-December 2022) of monthly NDVI, land temperature, and rainfall data. For each month, we:

- Preprocessing: Create a DataFrame with land temperature and rainfall as features and NDVI as the target variable.
- Data Split: Divide the DataFrame into training (80%) and testing (20%) sets.
- Model Training: Train the decision tree model on the training set.
- Performance Evaluation: Evaluate the model's performance on the testing set using:
  - Mean Squared Error (MSE): Measures the average squared difference between predicted and actual NDVI values.
  - R-squared: Represents the proportion of variance in NDVI explained by the model.
- Seasonal Comparison: Store the evaluation metrics and feature importance results for each month in separate arrays for further inter-monthly comparisons.
- Analysis: Conduct correlation analysis to examine the relationships between NDVI and environmental factors, and analyze feature importance values to identify the key influential factors for each month.

C. Results

We used violin plots of visualize the relationship between NDVI and features like land temperature and Rainfall. We created a heatmap to show the correlation matrix. These violin plots would allow you to visually assess how the NDVI values vary across different Land temperature and Rainfall conditions. We collected violin plots for January, April, July, and October. We also plotted the feature importance of the decision model for each month over the year. We also collected the mean-squared error and r-squared for each month and presented them in a table. We calculated the average mean-squared error and r-sqaured.

- Land temperature and NDVI Violin Plots for January (Figure 1), April (Figure 2), July (Figure 3), and October (Figure 4)

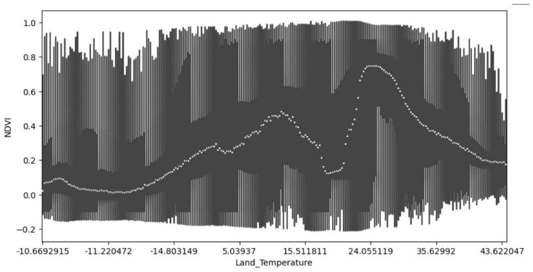


Fig. 1. January Land temperature Vs NDVI

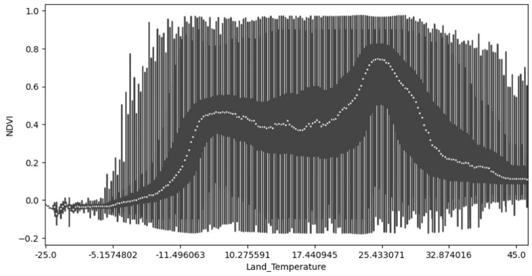


Fig. 2. April Land temperature Vs NDVI

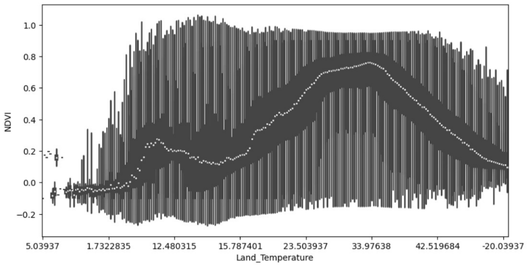


Fig. 3. July Land temperature Vs NDVI

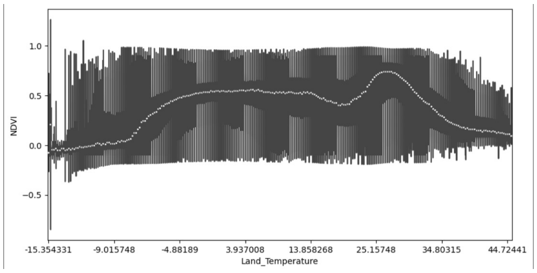


Fig. 4. October Land temperature Vs NDVI

- Rainfall and NDVI Violin Plots for January (Figure 5), April (Figure 6), July (Figure 7), and October (Figure 8)

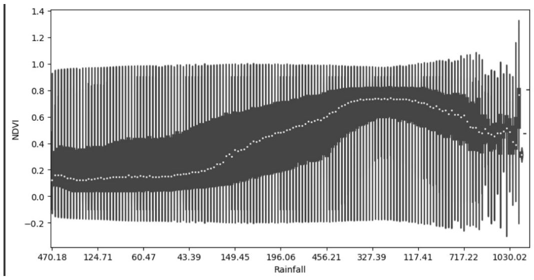


Fig. 5. January Rainfall Vs NDVI

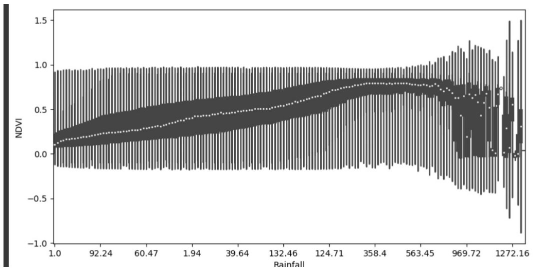


Fig. 6. April Rainfall Vs NDVI

- Land temperature features importance over months (Figure 9)
- Rainfall features importance over months (Figure 10)
- Evalution metric for each month (Figure 11)
- Average mean squared error: 0.024 and Average r-squared score: 0.64

D. Analysis of the Experimental Results

The first set of violin plots suggests a non-linear correlation between NDVI and land temperature. Initially rising with temperature, NDVI declines after reaching approximately 25 degrees Celsius. The non-linear relationship between land temperature also consists of data from January, April, July, and October The second violin plot indicates a positive correlation between NDVI and rainfall. As rainfall increases, so does the vegetation index, suggesting a positive effect on the vegetation index. The correlation is observed throughout the year. The average MSE of 0.024 suggests accurate predictions on average. Additionally, the average R-squared value of 0.64 indicates that approximately 64% of the variance in NDVI is explained by the model, highlighting its effectiveness in capturing the underlying patterns. The land temperature feature importance plots show that land temperature plays more importance from May to September. The rainfall feature importance plot suggests that rainfall plays more importance in vegetation from October to April. This suggests that during generally warmer seasons land temperature plays more importance in predicting vegetation and for cooler seasons rainfall plays more importance in predicting vegetation.

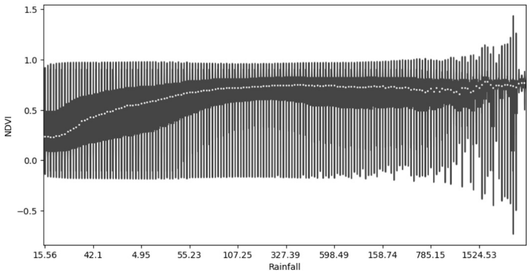


Fig. 7. July Rainfall Vs NDVI

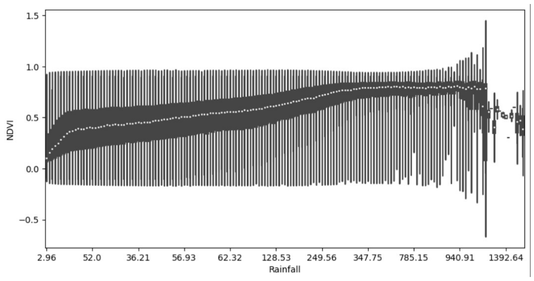


Fig. 8. October Rainfall Vs NDVI

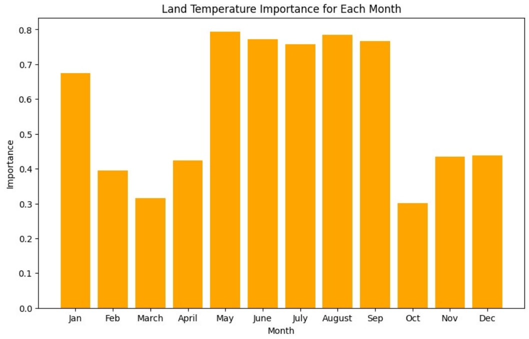


Fig. 9. Land Temparature importance over months

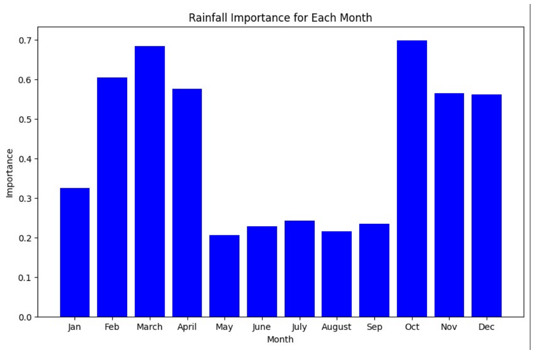
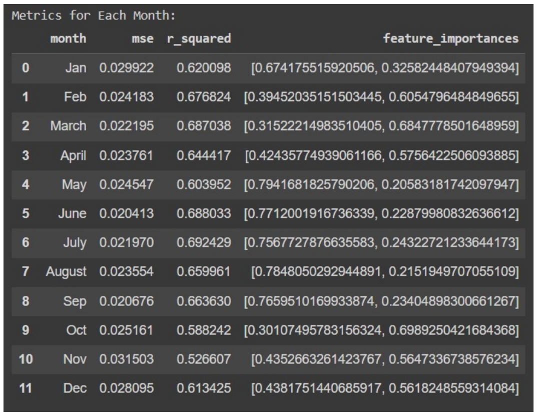


Fig. 10. Rainfall importance over months



```
Metrics for Each Month:
```

	month	mse	r_squared	feature_importances
0	Jan	0.029922	0.620098	[0.674175515920506, 0.32582448407949394]
1	Feb	0.024183	0.676824	[0.39452035151503445, 0.6054796484849655]
2	March	0.022195	0.687038	[0.31522214983510405, 0.6847778501648959]
3	April	0.023761	0.644417	[0.42435774939061166, 0.5756422506093885]
4	May	0.024547	0.603952	[0.7941681825790206, 0.20583181742097947]
5	June	0.020413	0.688033	[0.7712001916736339, 0.22879980832636612]
6	July	0.021970	0.692429	[0.7567727876635583, 0.24322721233644173]
7	August	0.023554	0.659961	[0.7848050292944891, 0.2151949707055109]
8	Sep	0.020676	0.663630	[0.7659510169933874, 0.23404898300661267]
9	Oct	0.025161	0.588242	[0.30107495783156324, 0.6989250421684368]
10	Nov	0.031503	0.526607	[0.4352663261423767, 0.5647336738576234]
11	Dec	0.028095	0.613425	[0.4381751440685917, 0.5618248559314084]

Fig. 11. Evaluation Metrics over months

V. Conclusion

In conclusion, this study aimed to analyze the correlation between vegetation index and key environmental factors, such as temperature and rainfall. Our contribution to existing literature lies in the comprehensive exploration of how seasonality impacts these correlations. Utilizing global data on NDVI, Land Surface Temperature, and rainfall for each month of 2022, we meticulously preprocessed and partitioned the data into training and testing sets. Through the implementation of a decision tree regression model, our evaluation metrics, including mean squared error and r-squared, showcased the model's robust performance.

The experimental findings reveal a nuanced relationship between land temperature and NDVI, characterized by a non-linear association, while rainfall exhibited a positive linear relationship with NDVI. Notably, the study identified distinct periods of significance, with land temperature playing a crucial role from May to September and rainfall exerting influence from October to April. The low mean squared error signifies the decision tree model's proficiency in predicting NDVI, and the substantial r-squared value of 0.64 indicates that 64

In terms of future research directions, expanding our analysis to consider factors like proximity to water bodies could enhance the precision of vegetation predictions for specific regions. Additionally, employing clustering models may prove valuable in identifying hotspots where vegetation is particularly susceptible to changes in environmental factors. These avenues hold the potential to deepen our understanding and refine predictions in the dynamic interplay between vegetation and environmental conditions.

References

[1] "Relation between vegetation, surface temperature, and surface composition in the tokyo region during winter," Remote Sensing of Environment, vol. 50, no. 1, pp. 52–60, Oct. 1994, doi: https://doi.org/10.1016/0034-4257(94)90094-9.

[2] M. Prerana, D. Gayke, S. Monika, and Rokade, "CROP YIELD PREDICTION USING WEATHER DATA AND NDVI TIME SERIES," 1637.

[3] V. K. Kalichkin, O. K. Alsova, and K. Yu Maksimovich, “Application of the decision tree method for predicting the yield of spring wheat,” IOP Conference Series: Earth and Environmental Science, vol. 839, no. 3, p. 032042, Sep. 2021, doi: https://doi.org/10.1088/1755-1315/839/3/032042.

[4] M. Shripathi Rao, A. Singh, N. V. Subba Reddy, and D. U. Acharya, “Crop prediction using machine learning,” Journal of Physics: Conference Series, vol. 2161, no. 1, p. 012033, Jan. 2022, doi: https://doi.org/10.1088/1742-6596/2161/1/012033.

[5] J. R. Jensen, Remote Sensing of the Environment. Pearson, 2007.

[6] Mildron, A., Entekhabi, D., and Ungar, S. (2016). Land Surface Temperature. In R. S. Kalluri and V. V. Lakshmi (Eds.), Remote sensing of water resources (pp. 507-545). CRC Press.

[7] 7. Nicholson, S.E., Funk, C. and Fink, A.H. (2018) Rainfall over the African Continent from the 19th through the 21st Century. Global and Planetary Change, 165, 114-127. https://doi.org/10.1016/j.gloplacha.2017.12.014

[8] Y. Wang, Z. Zhang, L. Feng, Q. Du, and T. Runge, “Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States,” Remote Sensing, vol. 12, no. 8, p. 1232, Apr. 2020, doi: https://doi.org/10.3390/rs12081232.