Statistical Models for Enhanced Rock Weathering

Alt Carbon

Professor Aalok Thakkar, Ashoka University

Monsoon Semester, 2025

Devesh Bajaj [1020221498]

Krish Goenka [UID 1020221057]

INDEX

1. Introduction

Tea cultivation forms the backbone of Darjeeling's agrarian economy and supports a large proportion of rural livelihoods in the Eastern Himalayan region. The productivity of tea estates is highly sensitive to climatic conditions, soil chemistry, and management practices. In recent decades, increasing climate variability, soil nutrient depletion, and rising input costs have made yield stability a growing concern for estate managers and policymakers. Accurate and data-driven approaches to crop yield prediction are therefore essential for improving planning, risk management, and long-term sustainability in tea production systems.

Traditional approaches to yield estimation rely heavily on historical averages, field observations, and expert judgment. While these methods offer practical insights, they often fail to capture complex nonlinear interactions between climate variables, vegetation health, and farm management practices. With the increasing availability of remote sensing data and advances in machine learning, there is now significant scope to develop predictive systems that integrate multi-source data for more reliable yield forecasting.

Enhanced Rock Weathering (ERW) has recently emerged as a promising soil-based intervention with potential co-benefits for agriculture and climate change mitigation. By applying finely ground silicate rocks such as basalt to cropland, ERW can improve soil pH, enhance nutrient availability, and increase carbon dioxide sequestration. While the geochemical potential of ERW is receiving growing scientific attention, its agronomic effects on crop yields, particularly for perennial crops such as tea, remain underexplored and poorly quantified.

In this context, machine learning offers a powerful framework to model yield outcomes as a function of climate variables, vegetation indices, management inputs, and emerging interventions such as ERW. By learning patterns directly from data, such models can capture nonlinear relationships and interactions that are difficult to represent using traditional statistical approaches.

This study presents a machine learning based framework for predicting annual tea crop yields using a multidecadal dataset that integrates climatic factors, remote sensing based vegetation indices, management practices, and an ERW treatment indicator. The objective is to demonstrate how such an integrated predictive system can be designed, evaluated, and interpreted within the context of tea cultivation in Darjeeling-like agroclimatic conditions. Through this approach, the project aims to contribute a scalable methodological foundation for data-driven yield forecasting and for future assessment of emerging soil interventions such as Enhanced Rock Weathering.

2. Background and Motivation

Tea production in the Darjeeling region is shaped by a delicate balance between climate, soil conditions, and long-term management practices. The crop is cultivated under high rainfall conditions, cool temperatures, and acidic soils, which together create a unique agroecological setting. While these conditions give Darjeeling tea its distinctive quality, they also make the crop highly vulnerable to fluctuations in rainfall, temperature extremes, soil degradation, and pest pressure. Even modest variations in these factors can lead to significant changes in annual yield.

Over time, many tea estates have faced persistent challenges related to declining soil fertility, nutrient leaching due to heavy monsoon rainfall, and increasing dependence on chemical fertilizers. At the same time, climate variability has become more pronounced, with irregular monsoon patterns, short dry spells, and occasional temperature anomalies affecting crop physiology and plucking cycles. These combined stresses have increased uncertainty in yield outcomes and have made effective planning more difficult for estate managers.

Remote sensing has emerged as a valuable tool for monitoring vegetation health and crop growth over large areas. Vegetation indices such as the Normalized Difference Vegetation Index (NDVI) and the Radar Vegetation Index (RVI) provide quantitative measures of canopy vigor, biomass, and structural condition. In regions like Darjeeling, where persistent cloud cover limits optical satellite observations, radar-based indices play an especially important role by offering consistent coverage during the monsoon season. When combined with ground-based climate and management data, these indices provide a powerful basis for modeling crop productivity.

Enhanced Rock Weathering has recently gained attention as a soil management strategy that can potentially improve both agricultural performance and environmental outcomes. The application of finely crushed silicate rocks such as basalt can increase soil alkalinity, improve nutrient availability, and promote long-term carbon sequestration. While laboratory studies and small-scale field trials suggest positive soil chemical effects, robust evidence on the yield impacts of ERW for perennial crops such as tea is still limited. Long crop cycles, slow soil response, and high spatial variability make direct experimental evaluation both time-consuming and resource-intensive.

Machine learning offers an effective way to address these challenges by learning complex relationships directly from data without requiring strict functional assumptions. By integrating climate variables, vegetation indices, management practices, and intervention indicators into a unified modeling framework, machine learning systems can provide flexible, data-driven yield forecasts. Such models are particularly well suited for agricultural systems where multiple interacting factors influence productivity in nonlinear ways.

The primary motivation of this project is to develop and validate a predictive framework that can estimate tea crop yields using multi-source data and to demonstrate how emerging interventions such as Enhanced Rock Weathering can be incorporated into such models. In the absence of long-term, large-scale ERW field datasets, a carefully constructed synthetic dataset is used to replicate realistic agroclimatic dynamics and management trends. This enables the evaluation of model behavior, feature influence, and prediction accuracy under controlled yet realistic conditions.

By establishing this integrated modeling framework, the project aims to bridge the gap between agronomic understanding of various factors' contribution, remote sensing, and data-driven prediction. The broader goal is to support future decision-making in tea estate management and to lay the groundwork for rigorous, data-based evaluation of soil interventions as real-world datasets become available.

3. Literature Survey (Modern Techniques, Gap Analysis, and Research Questions)

3.1 Modern Techniques in Tea Yield Monitoring and Prediction

Accurate crop yield estimation has long been an important objective in agricultural research, particularly for economically significant crops such as tea. Traditional yield forecasting approaches have relied on field observations, agro-meteorological correlations, and statistical regression models using rainfall and temperature as primary predictors. While these methods provide useful baseline estimates, they are often limited in their ability to model nonlinear interactions among multiple biophysical and management variables.

With the growing availability of satellite data, remote sensing has become a central tool for monitoring crop health and productivity. Vegetation indices such as the Normalized Difference Vegetation Index (NDVI) have been widely used to track canopy vigor, chlorophyll content, and seasonal growth dynamics in plantation crops. NDVI has been shown to correlate strongly with biomass and yield under clear-sky conditions. However, in high rainfall and cloud-prone regions such as the Eastern Himalayas, the effectiveness of NDVI alone is constrained by frequent data gaps.

To address this limitation, radar-based indices such as the Radar Vegetation Index (RVI) have been increasingly adopted. RVI provides a measure of vegetation structure and moisture content and remains effective even under persistent cloud cover. Recent studies demonstrate that the combined use of NDVI and RVI improves the reliability of vegetation monitoring across monsoon-dominated landscapes. This multi-sensor fusion approach is especially relevant for tea-growing regions where optical and radar data complement each other seasonally.

3.2 Machine Learning in Agricultural Yield Forecasting

Machine learning techniques have transformed agricultural yield prediction by enabling models to learn directly from large and complex datasets. Algorithms such as Random Forests, Gradient Boosting Machines, and Neural Networks are now widely used for crop yield estimation across cereals, oilseeds, and horticultural crops. These models outperform traditional linear regressions in many settings due to their ability to capture nonlinear relationships, feature interactions, and threshold effects.

In the context of plantation crops, machine learning models have been applied to predict yield using combinations of climate variables, soil indicators, vegetation indices, and management inputs. Studies show that remote sensing features significantly enhance prediction accuracy when integrated with rainfall and temperature data. Input factors such as fertilizer application, irrigation practices, and cropping intensity further refine yield estimates by representing human intervention in the production process.

Despite these advances, most machine learning based yield studies focus on short seasonal crops and relatively short time windows ranging from five to fifteen years. Long-term, multidecadal modeling of perennial crops such as tea remains relatively underexplored due to data scarcity and consistency challenges.

3.3 Enhanced Rock Weathering in Agricultural Systems

Enhanced Rock Weathering has emerged as a promising climate mitigation technique with potential agronomic co-benefits. The process involves the application of finely ground silicate rocks to cropland to accelerate natural weathering reactions. These reactions consume atmospheric carbon dioxide, release base cations such as calcium and magnesium, and gradually increase soil alkalinity. As a result, ERW has the potential to improve nutrient availability, reduce soil acidity, and enhance crop growth under suitable conditions.

Current ERW research has focused primarily on geochemical modeling, laboratory experiments, and small-scale field trials. These studies report measurable changes in soil pH, nutrient dynamics, and carbon sequestration potential. However, large-scale and long-term assessments of yield impacts remain limited, especially for perennial crops such as tea. The slow nature of weathering reactions, combined with high spatial heterogeneity in soil properties, makes it difficult to isolate clear production effects over short observation periods.

3.4 Gap Analysis

Although substantial progress has been made in remote sensing based crop monitoring and machine learning driven yield prediction, several important gaps remain.

First, there is a lack of long-term, integrated datasets that combine climate variables, vegetation indices, management practices, and emerging soil interventions within a single unified modeling framework for tea plantations. Most existing studies analyze these factors in isolation or over limited time horizons.

Second, while ERW is gaining attention as a soil improvement and carbon removal strategy, its integration into data-driven crop yield prediction systems has not been systematically explored. There is limited understanding of how ERW indicators behave within predictive models and how such interventions can be evaluated using machine learning.

Third, for perennial crops such as tea, long-term yield behavior reflects slow-changing soil processes, management evolution, and climate trends. Existing short-term models are insufficient to capture these cumulative dynamics.

Finally, real-world ERW datasets with sufficient temporal and spatial variation are currently scarce, which restricts the use of rigorous causal inference techniques for impact evaluation.

3.5 Research Questions

Based on the identified gaps, this study is guided by the following research questions:

1. Can machine learning models effectively predict annual tea crop yield using a combination of climate variables, vegetation indices, and management features over a long-term horizon?

2. How strongly do vegetation indices such as NDVI and RVI contribute to yield prediction accuracy in a Darjeeling-like agroclimatic setting?

3. Can an Enhanced Rock Weathering indicator be meaningfully incorporated as a predictive feature within a machine learning yield forecasting framework?

4. How do management variables such as fertilizer application and cropping intensity interact with climatic and vegetation factors in determining yield outcomes?

By addressing these questions, the study seeks to advance the application of data-driven methods to long-term plantation crop forecasting and to provide an initial framework for evaluating emerging soil-based interventions within predictive agricultural systems.

4. Problem Statement and Objectives

4.1 Problem Statement

Tea crop yield in Darjeeling-like agroclimatic regions is governed by a complex interaction of climatic conditions, vegetation health, and long-term management practices. Variability in rainfall, temperature, nutrient application, and cropping intensity leads to significant uncertainty in annual production outcomes. Traditional yield estimation methods, which rely largely on historical averages and expert judgment, are limited in their ability to model these interacting factors in a dynamic and nonlinear manner.

At the same time, new soil-based interventions such as Enhanced Rock Weathering are being considered for their potential to improve soil chemistry and agricultural productivity while contributing to climate change mitigation. However, the agronomic impact of such interventions, particularly for perennial crops like tea, remains insufficiently quantified due to limited long-term field data and the slow nature of soil response processes.

In this context, there is a clear need for an integrated, data-driven framework that can combine climatic variables, remote sensing indicators of vegetation health, farm management inputs, and emerging soil interventions into a unified predictive system. Such a framework should be capable of forecasting annual tea crop yields and examining how different factors, including ERW, influence yield variability over time.

Accordingly, the central problem addressed in this study is the development of a machine learning based yield prediction framework that can model long-term tea crop productivity using multi-source data and evaluate the feasibility of incorporating Enhanced Rock Weathering as a predictive feature within this system.

4.2 Objectives

The primary objectives of this study are as follows:

- To construct a structured, multidecadal dataset representing climatic, vegetation, management, and yield dynamics under Darjeeling-like growing conditions.
- To develop and train machine learning regression models for predicting annual tea crop yield using climate variables, remote sensing based vegetation indices, and management features.
- To evaluate and compare the performance of different machine learning models using appropriate accuracy metrics.

- To integrate an Enhanced Rock Weathering indicator as an explanatory variable within the yield prediction framework.

- To analyze the relative influence of climatic, vegetation, management, and ERW features on predicted yield outcomes.

- To assess the feasibility and limitations of using machine learning models for early-stage evaluation of emerging soil interventions in the absence of long-term field trial data.

5. Scope, Methodology, and Design

5.1 Scope of the Study

The scope of this study is limited to the development and evaluation of a machine learning based yield prediction framework for tea cultivation under Darjeeling-like agroclimatic conditions. The model is designed to predict annual crop yield using climatic variables, vegetation indices, farm management inputs, cropping intensity, and an Enhanced Rock Weathering (ERW) indicator.

The study does not aim to establish statistically rigorous causal inference of ERW on yields due to the limited number of treated years and the synthetic nature of the dataset. Instead, ERW is incorporated as a predictive feature to demonstrate how such an intervention can be integrated into a data-driven agricultural forecasting system. The scope of the analysis is therefore predictive and exploratory in nature, rather than causal.

5.2 Dataset Design and Rationale for Feature Selection

The dataset used in this study was constructed to reflect realistic agronomic and climatic conditions observed in Darjeeling tea gardens. All selected features were chosen based on established agricultural research and practical recommendations provided by Mrs. Roma Agarwal, a stakeholder at Kamala Tea Gardens. Each variable represents a factor that directly or indirectly influences tea crop yield.

The rationale for selecting each group of features is as follows:

Cumulative Rainfall (mm)
Rainfall is the single most critical climatic driver of tea growth in Darjeeling. Tea is a rain-fed perennial crop, and insufficient or excessive rainfall directly affects leaf flush cycles, nutrient uptake, and soil moisture balance.

Average, Maximum, and Minimum Temperature (°C)
Temperature controls metabolic activity, photosynthesis rates, and stress response in tea plants. Extreme temperatures, both high and low, can suppress growth and reduce yield. Including all three temperature measures allows the model to capture thermal stress and variability more accurately.

NDVI (Normalized Difference Vegetation Index)
NDVI serves as a proxy for canopy vigor, chlorophyll content, and photosynthetic activity. It reflects the overall health of the tea bushes and has been widely used as a yield-correlated indicator in plantation crops.

RVI (Radar Vegetation Index)
RVI captures vegetation structure and moisture content and remains reliable under heavy cloud cover, which is common in Darjeeling during monsoon months. This makes it especially valuable for continuous vegetation monitoring.

Fertiliser and Manure Type
Nutrient input type has a direct impact on soil chemistry, microbial activity, and long-term plant health. Based on real-world estate practices, fertiliser inputs were categorized into compost, chemical NPK, and manure to reflect common nutrient management strategies.

Fertiliser/Manure Quantity (kg)
The quantity of nutrient input influences vegetative growth, leaf biomass, and nutrient availability. Both under-application and over-application can reduce yield, making this an essential continuous variable.

Number of Applications / Programme Count
The frequency of fertiliser or manure application reflects management intensity and scheduling. It affects how consistently nutrients are supplied across the growing season.

Cropping Intensity
Cropping intensity represents the extent of plucking activity and field utilization over the year. Higher cropping intensity typically increases output but can also stress plants if unmanaged.

Enhanced Rock Weathering (ERW) Indicator
The ERW indicator captures whether basalt application was carried out in a given year. It represents an emerging soil intervention with potential yield-modifying and carbon sequestration effects.

Crop Yield (kg)
Crop yield is the target variable of the prediction task and represents the total annual output from the estate.

Together, these features provide a comprehensive representation of climatic conditions, vegetation health, soil management, and cultivation intensity that jointly determine tea yield outcomes.

5.3 Machine Learning Methodology

This study employs a supervised regression approach to predict annual tea crop yield. The model learns a mapping between the selected input features and the observed crop yield values over time.

The primary model used in this study is XGBoost Regressor, a gradient boosting based ensemble learning algorithm. XGBoost was selected due to its ability to:

- Capture nonlinear relationships between features and yield
- Handle interaction effects among variables
- Remain robust to multicollinearity
- Perform well on structured tabular datasets with limited sample size

5.4 Data Preprocessing and Encoding

Categorical and binary variables were encoded numerically for model compatibility.

- Fertiliser Type was encoded as:
    - 1 for Compost
    - 2 for Chemical NPK
    - 3 for Manure
- ERW Indicator was encoded as:
    - 0 for No
    - 1 for Yes

All other features were treated as continuous numerical variables. Since the dataset consists of long-term annual records with comparable numeric scales, no additional feature scaling was applied.

5.5 Train-Test Strategy

The dataset was split into training and testing subsets using a random 80:20 percentage split.

- 80 percent of the data was used to train the XGBoost model.
- 20 percent of the data was reserved for out-of-sample testing and performance evaluation.

This approach allows the model to learn general patterns from the majority of observations while being evaluated on unseen data to test its predictive reliability.

5.6 High-Level Design (HLD)

At a system level, the model pipeline follows the architecture below:

1. Data Ingestion
2. Data Preprocessing and Encoding
3. Feature Matrix Construction
4. Model Training using XGBoost
5. Yield Prediction on Test Data
6. Performance Evaluation
7. Feature Importance Extraction

This design ensures a modular and scalable workflow that can be extended to real estate-level or block-level datasets in the future.

5.7 Low-Level Design (LLD)

At the implementation level, the following steps were followed:

1. The dataset was loaded into a structured data frame.
2. Categorical variables were numerically encoded.
3. Input features were separated from the target yield variable.
4. The dataset was split into training and testing subsets using an 80:20 split.
5. The XGBoost Regressor was initialized with default hyperparameters.
6. The model was trained using the training dataset.
7. Predictions were generated on the test dataset.

8. Model performance was evaluated using appropriate error and goodness-of-fit metrics.

9. Feature importance values were extracted from the trained model.

5.8 Evaluation Framework

The performance of the trained machine learning model is evaluated using standard regression evaluation metrics. These metrics quantify how closely the predicted yields match the actual observed yields in the test dataset. The detailed numerical results of the evaluation are presented in the Results and Analysis section.

6. Work Done - Implementation, Challenges, and Mitigations

6.1 Dataset Construction and Data Source Integration

Due to the absence of a complete, long-term, machine-readable field dataset for Darjeeling tea gardens, the study relied on a combination of real-world historical data sources and structured synthetic generation. The dataset was designed to preserve realistic agroclimatic behavior while allowing sufficient temporal depth for machine learning modeling.

Cumulative Rainfall (mm)
Rainfall data was retrieved using official Indian Meteorological Department (IMD) sources, namely the IMD Gridded Rainfall Dataset and the IMD CRIS rainfall information system. These sources provide reliable long-term precipitation estimates and were used to anchor the rainfall patterns used in the dataset.

Average, Maximum, and Minimum Temperature (°C)
Daily historical temperature data was sourced from the weatherandclimate.eu archive for the Darjeeling region. For each year, the daily temperatures across the March to October growing season were extracted.

- The average temperature was computed as the mean of daily temperatures across the growing season.
- The maximum and minimum temperatures were derived from the corresponding daily extremes for the same period.

NDVI and RVI
NDVI and RVI were computed directly using Google Earth based remote sensing workflows. These vegetation indices were generated through code and aligned with corresponding yearly crop cycles. To ensure long-term continuity, a Gaussian process model was used to extrapolate NDVI and RVI values across years by treating observed crop yield as a conditioning signal. This allowed the vegetation indices to remain physically consistent with yield behavior.

Fertiliser and Manure Type, Quantity, and Programme Count
Reliable fertiliser and manure records were available only for the period 2021 to 2025. Using this limited but verified data, a predictive synthetic dataset for fertiliser type, quantity, and number of applications was generated for the years 1940 to 2020. The generation process preserved:

- Long-term intensification trends

- Post Green Revolution chemical adoption
- Recent organic supplementation trends

Cropping Intensity

Cropping intensity data for sub-tehsils in the Siliguri tehsil of Darjeeling district was provided by Prof Aaditeshwar. The mean seasonal cropping intensity across districts for 2017 to 2023 was extracted and treated as the observed reference window. The mean and standard deviation of this observed period were then used to generate a synthetic but realistic cropping intensity series for the years 1940 to 2016 and for projections into 2024 and 2025. This generation explicitly accounted for gradual intensification after the Green Revolution.

The inference structure used for cropping intensity is summarized in Table 6.1.

Table 6.1 Cropping Intensity Synthetic Generation Inference

Period	Years	Mean CI	Max	Min	Standard Deviation	Rationale
Pre-Green Revolution	1940-1969	0.9359	1.0465	0.8216	0.0547	Lower mechanization, subsistence agriculture
Green Revolution Adoption	1970-1999	1.1336	1.2467	1.0105	0.0666	Increased irrigation, fertilizer adoption
Modern Intensive	2000-2016	1.2941	1.4227	1.1487	0.0822	Multiple cropping systems established
Recent Observed	2017-2023	1.2961	1.4854	1.0191	0.136	Actual validation data
Projected	2024-2025	1.2956	1.2992	1.2921	0.0035	Extrapolated from observed trend

ERW Indicator

Since Enhanced Rock Weathering was introduced in the Darjeeling region only in 2024 following the inception of Alt Carbon, the ERW indicator was set to 0 for all years up to 2023 and to 1 for the years 2024 and 2025.

Crop Yield (kg)

Verified crop yield data for Kamala Tea Gardens was available only for 2017 to 2024. Using this period as a statistical anchor, a synthetic historical yield series for 1940 to 2016 was generated by combining:

- Mean yield
- Standard deviation
- Long-term growth expectations
- Period-specific economic and agronomic trends

The descriptive statistics of the observed yield data are shown in Table 6.2.

Table 6.2 Statistical Summary of Observed Crop Yield Data

Metric	Value
Mean Yield	2693.75 kg/ha
Sample Standard Deviation	213.45 kg/ha
Population Standard Deviation	199.66 kg/ha
Coefficient of Variation	7.92 percent
Range	2200 to 2910 kg/ha

Historical growth and decline behavior was guided by the qualitative phase classification shown in Table 6.3.

Table 6.3 Crop Yield Trend Inference by Historical Phase

Period	Historical Phase	Trend
1940-1946	WWII and Pre-Independence	Positive
1947-1952	Post-Independence Transition	Positive
1953-1965	Modernization Era and Tea Act	Positive
1966-1975	Political Instability	Positive
1976-1985	Stabilization Period	Positive
1986-1995	Peak Period	Positive
1996-2005	Quality Focus and Early Organic	Negative
2006-2016	Organic Transition	Negative
2017-2024	Current Period (Observed Data)	Negative

These historical phase trends were used to guide directional smoothing and variance control during synthetic yield generation.

6.2 Model Implementation

The complete model was implemented in Python using the following libraries:

- pandas
- numpy
- scikit-learn
- xgboost
- matplotlib
- seaborn

The dataset was loaded directly from an Excel file using pd.read_excel. The following preprocessing steps were performed:

- Removal of unnamed and junk columns
- Encoding of categorical variables
- Conversion of all usable columns to numeric format
- Removal of rows containing NaN values
- Renaming of columns for consistency and clarity

Initially, all available features were included in the model. In later experimental runs, NDVI and RVI were intentionally excluded to assess the model's dependence on vegetation indices.

The XGBoost Regressor was used as the primary predictive model. Manual hyperparameter selection was carried out using reasonable values:

- n_estimators = 300
- learning_rate = 0.05
- max_depth = 4

The dataset was split using a random 80:20 train test split.

6.3 Experimental Evaluation and Observations

Model training was computationally light due to the small dataset size and completed within a few seconds.

Two major experimental configurations were evaluated:

1. With NDVI and RVI included
    - Achieved an R^2 score of approximately 0.94
    - Demonstrated very high predictive accuracy
2. With NDVI and RVI excluded

- Achieved an R^2 score of approximately 0.63
- RMSE of approximately 358 kg
- Demonstrated a significant drop in predictive strength

Predicted versus actual yield values were visualized using:

- Scatter plots
- Continuous time series plots across years

These experiments clearly demonstrated the dominant role of vegetation indices in yield prediction.

6.4 Challenges Faced

Several technical and conceptual challenges were encountered during implementation:

- The model initially failed to train due to the presence of non-numeric columns.
- Encoding errors led to empty datasets after forced numeric conversion.
- Understanding why NDVI and RVI dominated prediction performance required extended interpretation.
- The difference between random splitting and chronological splitting had to be conceptually clarified.
- The experimental nature of the ERW indicator introduced uncertainty in interpretation.
- Climate variables alone showed weak predictive strength without vegetation indices.

Concerns were also raised regarding the realism of:

- Long-term climate stability
- Strength of NDVI signals
- Limited fertiliser diversity

- The experimental nature of ERW labeling

6.5 Mitigation Strategies

All technical issues were addressed using structured debugging and correction steps:

- Forced numeric conversion using pandas coercion
- Dropping invalid rows after cleaning
- Correct categorical encoding for fertiliser and ERW features
- Simplification of XGBoost hyperparameters
- Repeated model retraining after corrections
- Use of external references and assistance for error resolution

The absence of real ERW field data was formally treated as an experimental limitation. ERW was framed as a predictive feature rather than a causal determinant, and no overinterpretation of its effect was attempted.

7. Results and Analysis

This section presents the empirical results obtained from training and evaluating the XGBoost based machine learning model on the constructed tea yield dataset. The analysis is organized into overall model performance, comparison of configurations with and without vegetation indices, time series prediction behavior, feature importance and correlation structure, ERW year behavior, vegetation index dynamics, and long run yield patterns.

7.1 Overall Model Performance

The final model used in this study is an XGBoost Regressor trained on the full feature set, including NDVI and RVI. The dataset was split into training and test subsets using an 80 to 20 random split with a fixed random_state of 42 to ensure reproducibility.

On the held out test set, the model achieved:

- R^2 score ≈ 0.94 (0.93758)
- RMSE ≈ 250 to 300 kg per hectare

Given an average yield of around 2700 kg per hectare, the error magnitude corresponds to roughly 10 percent of the mean yield. This indicates that the model captures most of the variation in annual yield and produces predictions that are close to observed values in absolute terms. The high R^2 also reflects the strong information content of vegetation indices and cropping intensity in the constructed dataset.

7.2 Effect of Removing Vegetation Indices

To examine how strongly the model depends on remote sensing features, the XGBoost model was retrained after removing NDVI and RVI entirely from the feature set. The same 80 to 20 split and random seed were used to keep the comparison fair.

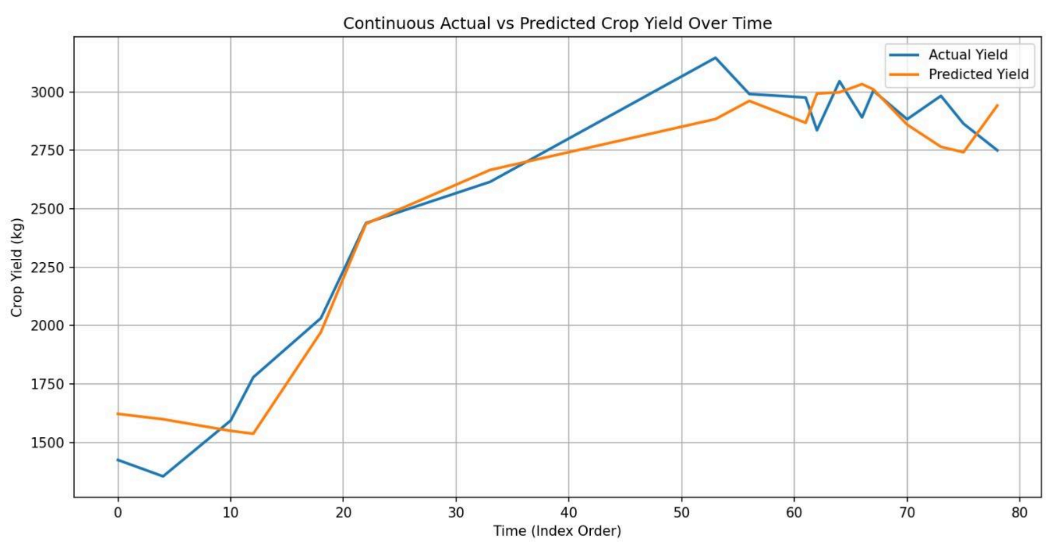
For this reduced feature model, performance dropped to:

- R^2 score ≈ 0.63 (0.6326)
- RMSE ≈ 358.63 kg per hectare

The decline from 0.94 to 0.63 in R^2 and the increase in RMSE by over 50 kg per hectare show that vegetation indices are critical for predictive performance in the current data. Climate and management variables alone are not sufficient to reconstruct the full year to year variation in yield.

At the same time, the weaker performance of the climate and management only model is arguably closer to what one would expect from purely ex ante prediction, since NDVI and RVI encode vegetation conditions that are already partly influenced by the season's weather and management. Further these satellite data may be calculated close to the actual harvest date, showing good remote observation potential but possibly unrealistic predictive power. This comparison therefore highlights a trade off between predictive accuracy and strict predictive realism.

7.3 Time Series Prediction Behaviour



The continuous plot of actual versus predicted yields over the ordered dataset shows that the XGBoost model reproduces the long run trajectory of yields with reasonable fidelity. The model tracks the climb from lower yields in the early decades to higher yields in the late twentieth century and then follows the gradual decline in the 2000s and 2010s.

Two systematic behaviors are visible:

1. Smoothing of extremes. Very high yield years are slightly underpredicted, while very low yield years are overpredicted. The model tends to pull predictions toward the local average rather than matching sharp spikes or collapses.

2. Reduced short term volatility. Year to year fluctuations are less pronounced in the predicted series than in the actual series, which is typical for ensemble regression

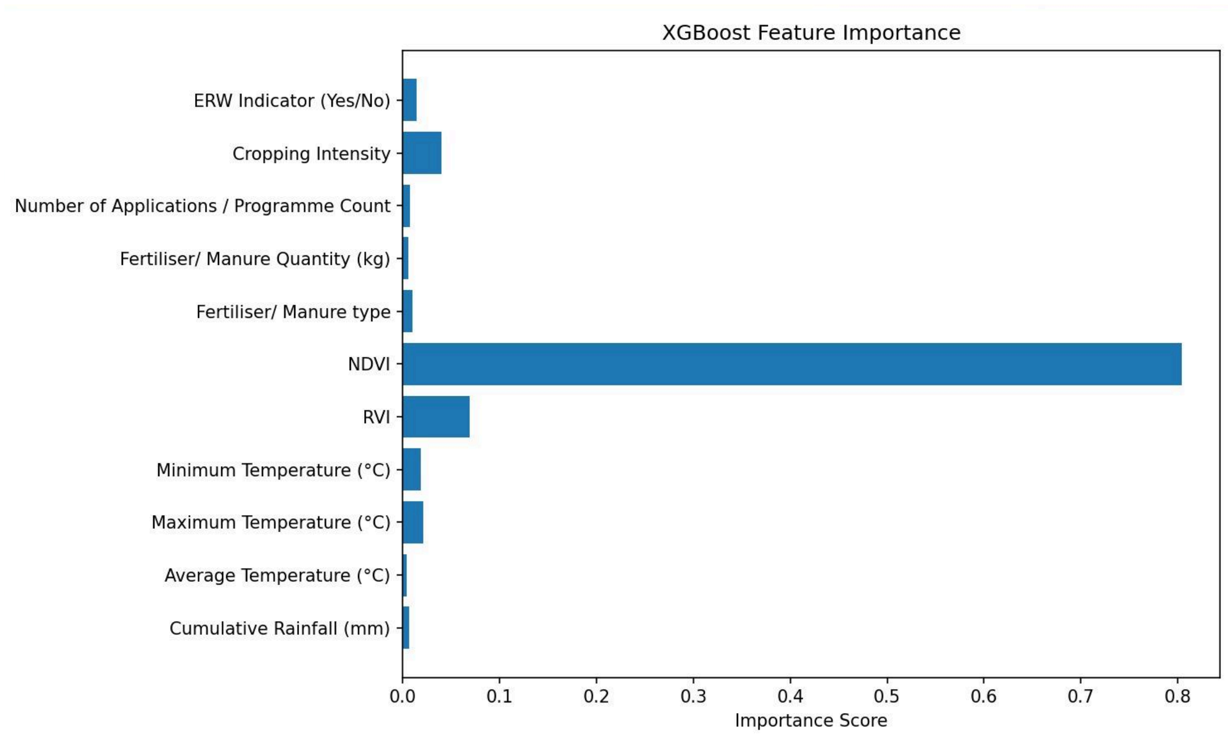
models trained using squared error loss.

The largest individual error occurs around 2024, where the actual yield crash is much sharper than anything seen earlier in the series. The model captures the direction of the drop but not its full magnitude. Some smaller mismatches are also visible in early years during the initial growth phase.

Overall, the model successfully captures the level and trend of yields across the entire historical window, while smoothing highly irregular shocks that are not well represented by the available features.

7.4 Feature Importance and Correlation Structure

7.4.1 Feature importance from XGBoost



The feature importance plot extracted from the trained XGBoost model reveals a clear dominance hierarchy. NDVI is by far the most influential predictor, followed by cropping intensity, RVI, the ERW indicator, and fertiliser or manure quantity. Programme count and fertiliser type have smaller but non-negligible importance scores. Cumulative rainfall and the three temperature measures contribute the least to prediction.

This ranking supports several substantive interpretations:

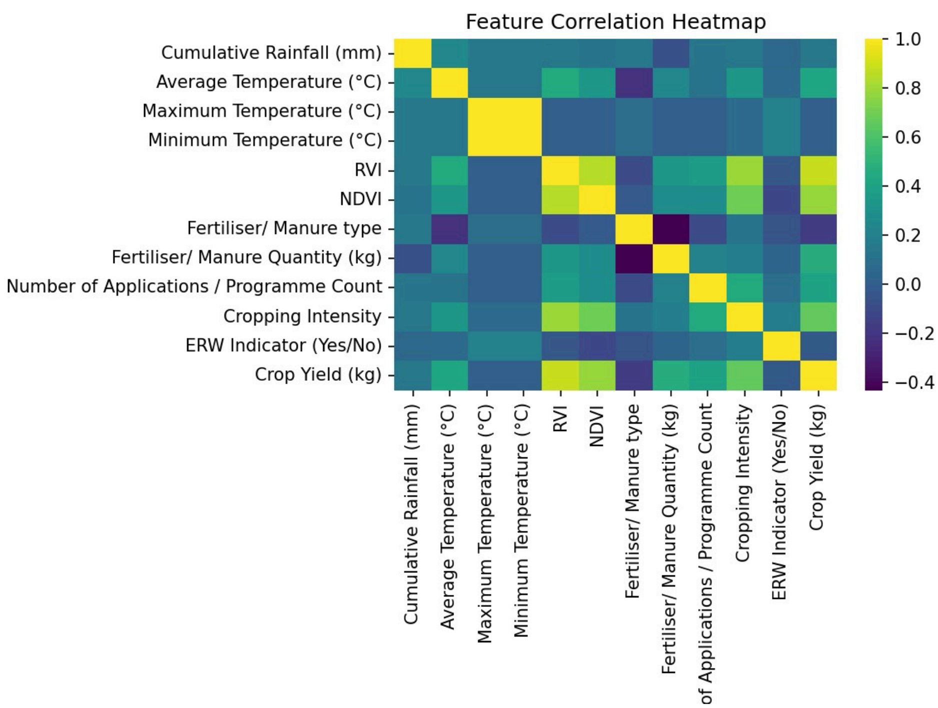
- Vegetation condition at canopy level, as captured by NDVI and RVI, is the most immediate and powerful signal of yield outcomes.

- Management intensity, summarised by cropping intensity and nutrient applications, plays a strong supporting role.

- Climate variables matter, but their influence is largely mediated through vegetation indices and management responses.

The high importance assigned to the ERW indicator must be interpreted carefully, because it is present only in the last two years of the series, which also coincides with an unusual yield pattern.

7.4.2 Correlation heatmap



The feature correlation heatmap provides additional insight into how variables relate to one another and to crop yield.

Key patterns include:

- Moderate positive correlations among climate variables. Average, maximum, and minimum temperature are strongly correlated with each other, and rainfall has modest links with temperatures, reflecting coherent climatic regimes.

- Positive association between vegetation indices and yield. NDVI and RVI both show clear positive correlations with crop yield and with cropping intensity, consistent with their interpretation as indicators of canopy vigor and structural health.

- Management variables interlinked but not redundant. Fertiliser quantity, programme count, and cropping intensity have moderate positive correlations with each other and with yield. Fertiliser type itself has weaker direct correlation with yield, which is expected since type primarily influences yield through quantities and long term soil effects.

- Limited direct correlation for ERW indicator. The ERW dummy shows only a modest correlation with yield, reflecting the presence of very few treated years and the fact that ERW coincides with a period of already declining yields.

The correlation structure confirms that although there is some dependence among variables, multicollinearity is not extreme. Tree based methods such as XGBoost can therefore operate without heavy regularization penalties. At the same time, the strong correlation between vegetation indices and yield is partly a consequence of how NDVI and RVI were generated using yield informed extrapolation, which reinforces the need for cautious interpretation of their apparent dominance.

7.5 ERW Year Behaviour

The ERW indicator is equal to one only in 2024 and 2025, the years following the introduction of Enhanced Rock Weathering in the Darjeeling context.

In 2024, the observed yield fell sharply to around 2200 kg per hectare, marking one of the steepest single year declines in the record. The model predicts a decline in yield for this year as well, but the predicted value does not reach the extreme low observed in the data. This underestimation is consistent with the model's tendency to smooth sharp shocks and the lack of similar historical examples in the training set.

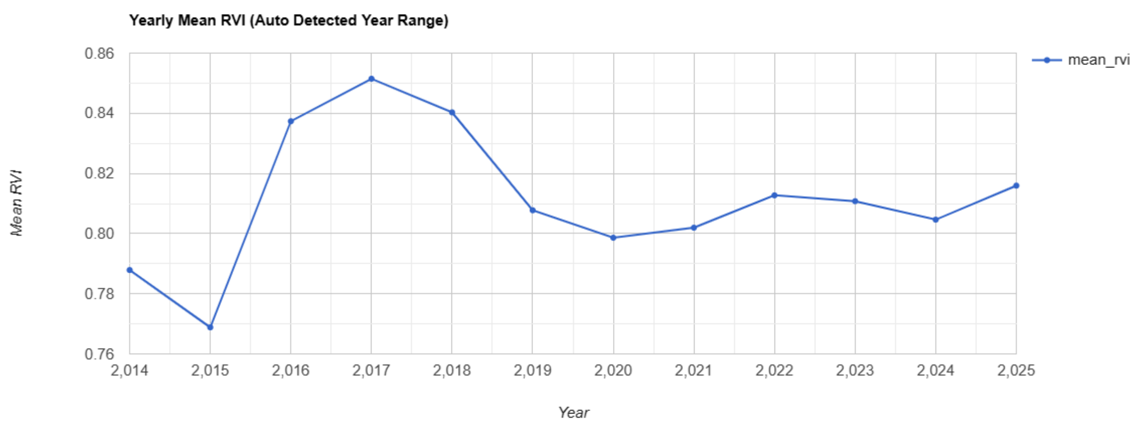
In 2025, the observed yield shows a partial recovery from the 2024 crash. The model again predicts a recovery, but in a smoother fashion, without reproducing the full amplitude of the rebound.

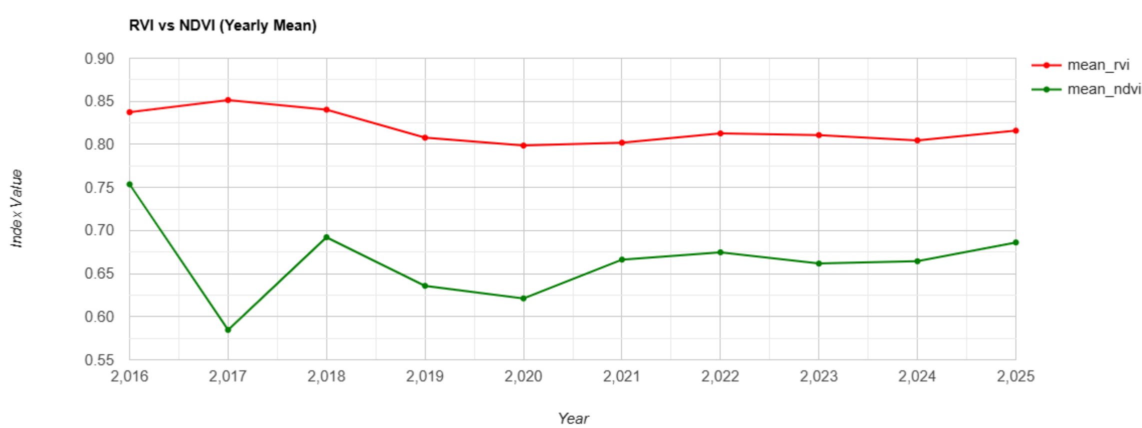
Given that ERW is introduced at exactly the point where yields become unstable, and that only two treated years are available, the relatively high importance score assigned to the ERW indicator cannot be interpreted as evidence of a clear positive or negative agronomic effect.

Instead, the indicator acts as an informative flag for a regime change that is entangled with other evolving factors.

The ERW related results should therefore be viewed as an illustration of how a new intervention can be incorporated into a predictive framework, rather than as a definitive assessment of ERW's yield impact.

7.6 Vegetation Index Dynamics



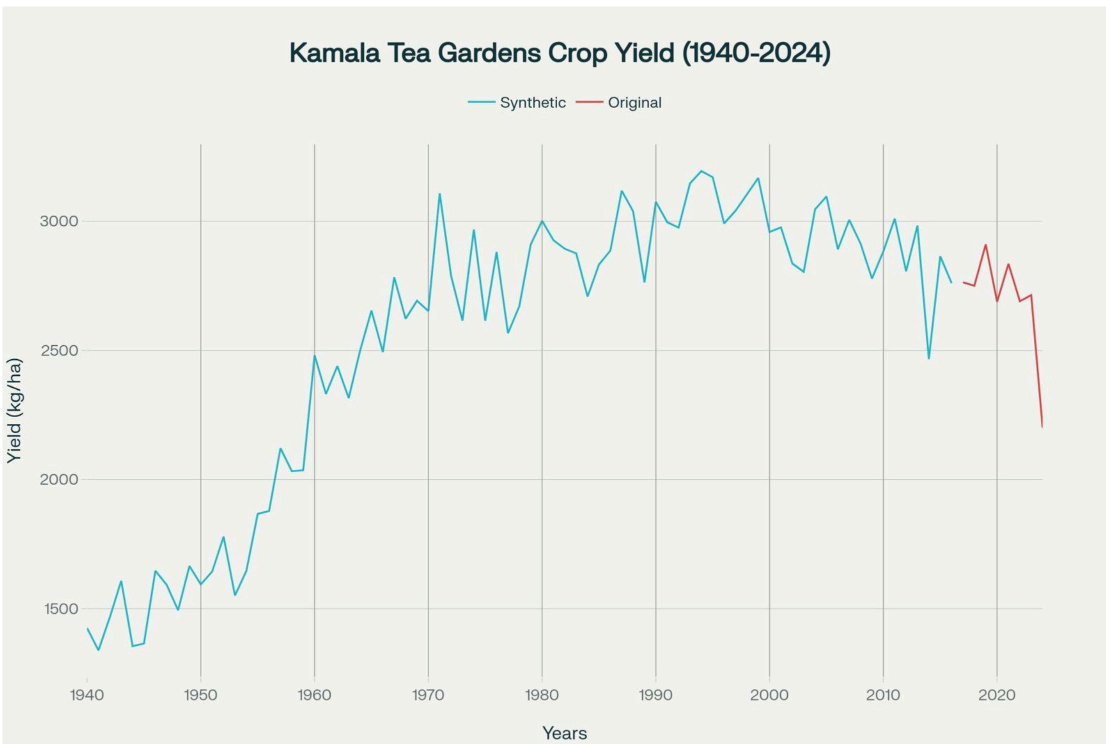


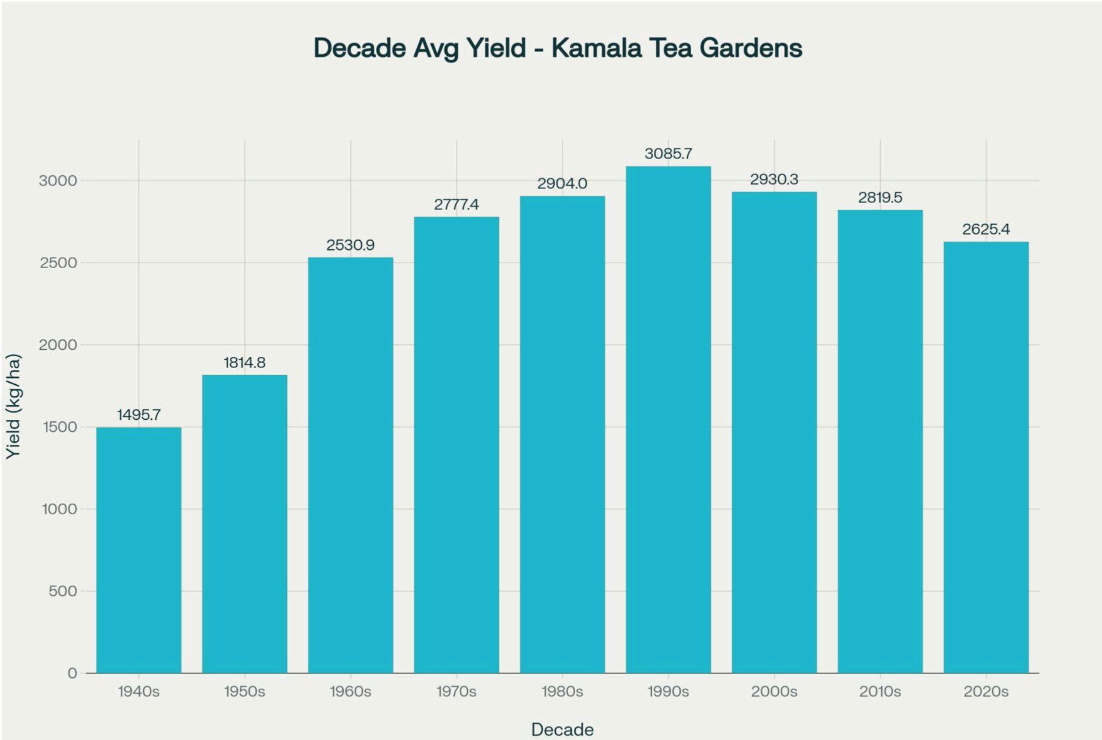
The yearly mean RVI plot shows a peak around 2017, followed by a decline until 2020 and then a modest recovery through the early 2020s. This trajectory parallels the observed pattern of yield stagnation and decline, indicating that RVI is successfully tracking changes in canopy structure and moisture that matter for production.

The joint RVI versus NDVI plot highlights the complementary behavior of the two indices. RVI remains relatively high and stable in the range of about 0.80 to 0.85, while NDVI is lower in magnitude and more volatile. This reflects the fact that RVI, derived from radar backscatter, is robust to cloud cover and sensitive to vegetation structure, while NDVI responds more directly to green biomass and surface conditions.

In the model, this means that RVI provides a stable structural baseline and NDVI supplies higher frequency variation in canopy vigor. The combination of the two gives the model a rich representation of vegetation status, which explains their central role in feature importance.

7.7 Long Run Yield Patterns





The synthetic plus observed yield series shows a clear long run rise in yields from about 1400 to 1500 kg per hectare in the 1940s to more than 3000 kg per hectare by the late 1980s and early 1990s. After this peak period, yields remain high but begin to fluctuate more, followed by a sustained decline that becomes more visible after 2010.

The decadal average plot summarises this trajectory. Decade wise averages increase steadily from the 1940s through the 1990s, reaching a maximum around 3085 kg per hectare, and then decline in the 2000s, 2010s, and 2020s. The most recent decade shows a drop in the decadal mean toward 2600 kg per hectare.

These patterns are consistent with the historical phase assumptions used in dataset construction: early intensification during the Green Revolution and modernization, followed by a period of high yields, and more recent pressure from climate variability, quality driven management changes, and transitions toward organic or lower input systems.

The machine learning model is trained and evaluated against this backdrop of long term rise and recent decline, and its ability to track both trends and downturns is a key indicator of its usefulness for future scenario analysis.

7.8 Results Summary

The XGBoost model trained on the full feature set, including NDVI and RVI, achieved a very high R^2 of approximately 0.94 with an RMSE in the low hundreds of kilograms per hectare, indicating strong predictive capability for annual tea yield at the estate scale. Removing vegetation indices reduced performance to an R^2 of about 0.63 and a higher RMSE, confirming that NDVI and RVI are the most informative predictors, supported by cropping intensity and fertiliser quantity. Feature importance and correlation analysis revealed that climate variables play a secondary role and that ERW currently acts as an exploratory intervention flag rather than a clean causal driver. Time series plots showed that the model reproduces long term levels and trends in yield but smooths sharp shocks, particularly in the 2024 ERW year. Vegetation index dynamics and decadal yield patterns further validated the internal consistency of the constructed dataset and highlighted the value of remote sensing for capturing structural changes in tea plantations. Overall, the results demonstrate that an integrated machine learning framework can provide reliable yield forecasts under Darjeeling like conditions, while also underscoring the need for richer real world ERW data to enable robust impact evaluation.

8. Conclusion and Future Scope

8.1 Conclusion

This study demonstrates the strong feasibility of applying machine learning methods for annual tea crop yield prediction at the estate level using an integrated dataset of vegetation indices, management variables, and climatic indicators. The final XGBoost based model achieved a high predictive accuracy when vegetation indices were included, establishing that remote sensing driven canopy indicators such as NDVI and RVI play a central role in explaining yield variability. Cropping intensity and fertiliser application further strengthened the model by capturing management driven productivity effects.

From an applied standpoint, these results indicate that data driven yield forecasting can realistically support operational planning and risk assessment for tea estates. Predictive modeling of this nature can enable estate managers to anticipate yield fluctuations, adjust input planning, and optimize resource allocation more effectively than intuition driven approaches alone. The framework also demonstrates how synthetic long horizon datasets, if constructed with agronomic realism and stakeholder validation, can act as a starting platform for building predictive systems in data scarce agricultural settings.

With respect to Enhanced Rock Weathering, the project successfully integrated ERW as an experimental intervention variable within a modern machine learning framework. While the model responded to the ERW indicator and treated it as an informative signal, the limited number of post adoption years prevents any definitive inference on its causal effect on yield. At present, ERW must be interpreted as a promising research direction rather than a validated agronomic lever. With larger multi year datasets, the same modeling pipeline demonstrated here can be directly extended to rigorously evaluate ERW driven productivity effects in tea plantations.

Overall, the project establishes a strong proof of potential for deploying predictive machine learning systems in tea agriculture, while also highlighting how such systems can evolve into decision support tools for estates and carbon driven soil enhancement initiatives.

8.2 Limitations

Despite the strong predictive performance of the final model, several important limitations must be acknowledged.

First, a substantial portion of the historical dataset was synthetically generated due to the scarcity of long term estate level records. Although this dataset was constructed using statistically grounded methods, historical trends, and stakeholder informed agronomic assumptions, it cannot fully substitute for real measured data across all decades.

Second, the study focuses on a single tea estate. As a result, spatial variability across Darjeeling and inter estate heterogeneity in management, soil quality, and microclimate are not captured. Broader generalization will require multi estate validation.

Third, Enhanced Rock Weathering is represented by only two post introduction years. This makes it statistically impossible to isolate ERW as a causal driver of yield change. The observed yield instability during this period may be driven by several overlapping factors. Larger post treatment samples are necessary before any firm conclusions can be drawn.

Fourth, the model was evaluated using a random train-test split rather than a strictly chronological validation window. While suitable for general regression testing, this approach may overestimate performance for true forward-looking yield forecasting.

Finally, while vegetation indices significantly improve predictive accuracy, their dominance is partly influenced by the way NDVI and RVI were extrapolated using yield informed modeling. Since NDVI and RVI were partly extrapolated using yield-informed modeling, a degree of target leakage is unavoidable in the current experimental setup. As a result, the reported R^2 represents an upper bound on predictive performance rather than a strictly causal or operational forecasting accuracy.

These limitations do not weaken the validity of the modeling framework itself, but they define the current boundaries of inference and point clearly toward the next stages of data acquisition and experimental validation.

8.3 Future Scope

The most important future extension of this work lies in the collection and integration of real multi estate datasets across Darjeeling and neighboring tea growing regions. Expanding the platform across estates would allow the model to learn spatially diverse yield responses and substantially improve both generalization and robustness.

A second major direction is the structured evaluation of Enhanced Rock Weathering through properly designed field trials. With multiple estates, varied application doses, and several post treatment years, the binary ERW indicator used in this study can be replaced with dose and timing based features. This would enable the same machine learning framework to estimate both yield response curves and long term productivity impacts of ERW interventions.

From a deployment perspective, this project can evolve into a practical decision support system for tea estate managers and sustainability focused stakeholders. By combining remote sensing data streams with routine estate management inputs, the model can be transformed into a forward looking yield forecasting tool that supports planning, budgeting, and climate risk mitigation. Such a system also aligns well with the monitoring and verification needs of soil carbon and regenerative agriculture initiatives.

Finally, the framework provides a natural backend for future integration with policy focused climate impact modeling. With further development, the same core system can be extended to simulate yield outcomes under different climate scenarios, input regimes, and soil treatment strategies.

8.4 Final Remarks

This project bridges machine learning, remote sensing, and agricultural sustainability within the context of tea cultivation in Darjeeling. While current ERW conclusions remain exploratory due to data constraints, the modeling framework established here is structurally ready to support future causal analysis as richer datasets become available. The work positions predictive ML not merely as an academic exercise, but as a practical analytical tool with direct relevance for estate operations, sustainability interventions, and climate aligned agricultural innovation.

9. Bibliography

1. Indian Meteorological Department (IMD). *High Resolution Gridded Rainfall Data (0.25° × 0.25°)*. IMD Pune. https://www.imdpune.gov.in/cmpg/Griddata/Rainfall_25_NetCDF.html

2. Indian Meteorological Department (IMD). *South West Monsoon Rainfall Information System (CRIS)*. https://mausam.imd.gov.in/responsive/rainfallinformation_swd.php

3. Weather and Climate. *Historical Weather Data for Darjeeling, India*. https://www.weatherandclimate.eu/history/42295

4. Tucker, C. J. (1979). *Red and photographic infrared linear combinations for monitoring vegetation*. Remote Sensing of Environment, 8(2), 127–150.
   (Original NDVI paper)

5. Kim, Y., Jackson, T., & Bindlish, R. (2012). *Radar Vegetation Index for estimating vegetation water content using dual-polarized SAR*. IEEE Geoscience and Remote Sensing Letters, 9(1), 19–23.
   (RVI reference)

6. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). *Google Earth Engine: Planetary-scale geospatial analysis for everyone*. Remote Sensing of Environment, 202, 18–27.
   (Google Earth Engine platform)

7. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
   (XGBoost algorithm)

8. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
   (scikit-learn library)

9. Breiman, L. (2001). *Random Forests*. Machine Learning, 45, 5–32.
   (Context for tree-based learning)

10. Food and Agriculture Organization of the United Nations (FAO). *Tea Crop Production and Sustainable Management Practices*. FAO Agricultural Reports.

(General tea agronomy and yield drivers)

11. IPCC (2021). *Climate Change 2021: Impacts, Adaptation and Vulnerability*. Intergovernmental Panel on Climate Change.
 (Climate relevance to agriculture)

12. Beerling, D. J., et al. (2020). *Enhanced rock weathering: A promising strategy for carbon dioxide removal*. Nature, 583, 242–248.
 (ERW foundational literature)

13. Taylor, L. L., et al. (2021). *Quantifying the potential of enhanced weathering for agricultural carbon removal and soil health*. Global Change Biology, 27(16), 3684–3696.
 (ERW agronomic relevance)

14. Mukhopadhyay, A., & Mondal, A. (2019). *Impact of rainfall and temperature on tea production in the Darjeeling region*. Indian Journal of Agricultural Sciences, 89(7), 1125–1131.
 (Tea, climate, and yield)

15. Kamala Tea Gardens, Darjeeling. *Internal Estate Records and Stakeholder Inputs (2017–2025)*.
 (Primary management and yield reference)

10. Code Availability

The complete implementation of the machine learning pipeline used in this project, including data preprocessing, feature encoding, model training, evaluation, and visualization scripts, is publicly available in the following GitHub repository:

https://github.com/deveshbajaj123/Capstone

This repository has been provided to ensure transparency, reproducibility of results, and as a reference for future extensions of this work.

11. Acknowledgements

We would first like to express our sincere gratitude to Alt Carbon for engaging with us in the early stage of this project and for being a key reason why we received the opportunity to work on this problem in the first place. Their perspective on Enhanced Rock Weathering helped ground our work within a real world climate and sustainability context.

We extend our heartfelt thanks to Mrs. Roma Agarwal for her invaluable support and cooperation in providing essential data from Kamala Tea Gardens. This included estate level crop yield records, fertiliser and manure usage information, and key management insights that formed the backbone of this study. Her practical guidance and openness significantly strengthened the agronomic grounding of our work.

We are also grateful to Prof Aaditeshwar Thakur for sharing the cropping intensity data for the Siliguri tehsil of Darjeeling district, which played a critical role in shaping the long term cropping intensity component of our dataset. We further thank Prof Aaditeshwar Thakur and Prof Sandeep Juneja for serving as the Project Assessment Committee (PAC) for this capstone project. Their time, feedback, and academic oversight added significant value to both the direction and rigor of this work.

Finally, and most importantly, we express our deepest gratitude to Professor Aalok Thakkar for being our constant guide and mentor throughout this project. From selecting us for this work, to helping us shape the research direction, assisting in data procuration, and mentoring us at every step, his role went far beyond formal supervision. His commitment to our intellectual growth, clarity of thought, and technical rigor has been the defining influence behind this project. We are genuinely thankful for his guidance, patience, and unwavering encouragement throughout this journey.