**Crop Yield Prediction of Groundnut in Gujarat State using Machine Learning Algorithms** 

**M.Sc. Agriculture Analytics**

**Submitted By:**

**Drashti Nayakpara**

ID: 202319019 / TS90709

**Divya Prajapati**

ID: 202319024 / TS90712

**Submitted To:**

**Dr. N. R. Patel**

(Agriculture & Soils Department)

**Dr. Kamal Pandey**

(Geoweb Services, IT and Distance Learning)

**Table of Contents**

[**Abstract** 4](#_Toc164430924)

[**Acknowledgements** 5](#_Toc164430925)

[**1.** **Introduction** 6](#_Toc164430926)

[**1.1.** **Objective** 7](#_Toc164430927)

[**1.2.** **Research Questions** 7](#_Toc164430928)

[**2.** **Literature Review** 8](#_Toc164430929)

[**3.** **Data collection and Feature Engineering** 10](#_Toc164430930)

[**4.** **Study Area and Materials Used** 11](#_Toc164430931)

[**4.1.** **Study Area** 11](#_Toc164430932)

[**4.2.** **Satellite Data and Products** 12](#_Toc164430933)

[**4.3.** **Ancillary Crop Data** 14](#_Toc164430934)

[**4.4.** **Crop Mask** 14](#_Toc164430935)

[**4.5.** **Vegetation Index** 14](#_Toc164430936)

[**4.5.1.** **Normalized Difference Vegetation Index (NDVI)** 14](#_Toc164430937)

[**4.5.2.** **Enhanced Vegetation Index (EVI)** 15](#_Toc164430938)

[**4.5.3.** **Leaf Area Index (LAI)** 16](#_Toc164430939)

[**4.5.4.** **Fraction Photosynthetically Active Radiation (FPAR)** 17](#_Toc164430940)

[**4.5.5.** **Gross Primary Product (GPP)** 18](#_Toc164430941)

[**4.5.6.** **Rainfall** 19](#_Toc164430942)

[**4.5.7.** **Land Surface Temperature (LST)** 20](#_Toc164430943)

[**4.5.8.** **Soil Moisture Index (SMI)** 21](#_Toc164430944)

[**5.** **Methodology** 23](#_Toc164430945)

[**5.1. Algorithms (Models)** 24](#_Toc164430946)

[**5.1.1. Training the Model** 24](#_Toc164430947)

[**6.** **Result and Discussion** 29](#_Toc164430948)

[**6.1.** **Evaluation of the Model** 29](#_Toc164430949)

[**6.2.** **Model Assessment** 29](#_Toc164430950)

[**6.3.** **Hyperparameter Tuning** 30](#_Toc164430951)

[**6.4.** **Dashboard** 31](#_Toc164430952)

[**6.5.** **Model Deploy** 32](#_Toc164430953)

[**7.** **Conclusion** 34](#_Toc164430954)

[**References** 35](#_Toc164430955)

[**Appendix** 37](#_Toc164430956)

[**Figure 1: Correlation: Heat Map** 11](#_Toc164430957)

[**Figure 2: Study Area** 12](#_Toc164430958)

[**Figure 3: Spatial Variation of NDVI within the crop Masked Area** 15](#_Toc164430959)

[**Figure 4: Spatial Variation of EVI within the crop Masked Area** 16](#_Toc164430960)

[**Figure 5: Spatial Variation of LAI within the crop Masked Area** 17](#_Toc164430961)

[**Figure 6: Spatial Variation of FPAR within the crop Masked Area** 18](#_Toc164430962)

[**Figure 7: Spatial Variation of GPP within the crop Masked Area** 19](#_Toc164430963)

[**Figure 8: Spatial Variation of Rainfall within the crop Masked Area** 20](#_Toc164430964)

[**Figure 9: Spatial Variation of LST within the crop Masked Area** 21](#_Toc164430965)

[**Figure 10: Spatial Variation of SMI within the crop Masked Area** 22](#_Toc164430966)

[**Figure 12: Accuracy Assessment: After Hyperparameter Tunning** 31](#_Toc164430967)

[**Figure 11: Accuracy Assessment: Before Hyperparameter Tunning** 31](#_Toc164430968)

[**Figure 13: Dashboard** 32](#_Toc164430969)

[**Figure 14: Model Deploy** 33](#_Toc164430970)

[**Table 1: Satellite and Parameter** 13](#_Toc164430976)

[**Table 2: Accuracy Score** 30](#_Toc164430977)

# **Abstract**

Planning ahead and accurately anticipating crop yield is critical for farmers and the agriculture sector. This aids in strategy planning and educated agricultural import and export decisions, which have a direct bearing on farmers' income. We can anticipate crop yields more precisely by applying machine learning algorithms, which might benefit farmers and the sector as a whole.

In this study, groundnut production for the five districts of Gujarat state—Junagadh, Jamnagar, Amreli, Bhavnagar, and Rajkot—was predicted using machine learning-based algorithms ahead of the actual harvest. The Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Leaf Area Index (LAI), Fraction of Photosynthetically Active Radiation (FPAR), Gross Primary Product (GPP) with weather parameters Rainfall, Land Surface Temperature (LST), and Soil Moisture Index (SMI) are among the products that are used to train the ML models. These products are obtained from MODIS, NASA and CHIRPS. Models are assessed using parameters such as R2 Score. With an R2 Score of 0.91, the XGBoost Regressor method produced the best results out of all of other algorithms.

**Keywords**: Machine learning, Yield prediction, Groundnut, Vegetation Indices, Linear Regression, XGBoost, Random Forest

# **Acknowledgements**

On completion of the research work as part of the M. Sc degree program offered as Joint Education Program by DAIICT, IIRS and AAU, I would like to remember with great gratitude all the fascinating people who made it possible with their relentless support and insightful guidance.

This wouldn’t have been possible without the solicitous guidance of my supervisors N. R. Patel sir and Kamal Pandey sir who have been instrumental in shaping the research and guiding me towards the completion of research. They have been helpful guiding me throughout this journey whenever I needed their support.

I would like to thank my batchmates, DAIICT batchmates who have been the support system for me throughout this journey. They helped me to correct my mistakes and guided me through this journey. I am very happy to share a journey with them.

Last but not the least, the support from my family has been instrumental for me to complete the course. Their hard work and supporting words had inspired me to keep going.

# **Introduction**

Groundnut is a significant oilseed and cash crop in India. It is an excellent source of protein, iron, calcium, and vitamins in addition to high-quality edible oil. It is also known by various name like peanut, wondernut, and poor man's cashew nut[1]. Peanuts are also known by many other local names such as earthnuts, ground nuts, goober peas, monkey nuts, pygmy nuts and pig nuts. Peanut or groundnut (*Arachis hypogaea*), is a species in the legume or "bean" family. The peanut was probably first domesticated and cultivated in the valleys of Paraguay. It is an annual herbaceous plant. Groundnut is the major oil seed crop in India and it plays a major role in providing raw materials to various industries and creates employment opportunities for millions of people[2].

India is the country with the most groundnut farming area, however its output is low when compared to China and the USA. This is partly because the crop is cultivated in arid, low-fertility, rain-fed fields with no input control. The production of groundnut has increased in recent years due to improved farming technologies and optimum resource utilization[3]. The principal Indian states that cultivate groundnuts include Gujarat, Andhra Pradesh, Karnataka, Maharashtra, Tamil Nadu, Madhya Pradesh, Rajasthan, and Orissa.

Gujarat accounts for 40.5% of the total groundnut production of India — the sowing was 2.16 lakh hectares in 2020-21 whereas the production was 4.16 million tonnes, followed by 1.39 tonnes in Rajasthan. It is the major oilseed crop cultivated highest in the Saurashtra region of the Gujarat state. The practice of analyzing massive amounts of data to find relevant information is known as data mining. In today's digital environment, data mining is becoming more and more vital to turn data into knowledge. When applied to farming, Remote Sensing empowers farmers to make more informed decisions by utilizing data-driven insights[4].

Machine Learning using remote sensing data aims to increase the accuracy of crop predictions. Predicting crop production is a big challenge in agriculture because it depends on so many different factors, such weather, soil, the state of the economy, and more[5]. While farmers had to estimate production in the past based on their expertise, Machine Learning uses data-driven methods to produce projections that are more accurate. The historical data is utilized as training data to develop the ability to categorize yield estimates for the future. Precise projections of crop yields are crucial for planning procurement, distribution, storage, major agricultural import-export plans, marketing, and in some instances, taking timely decisions to increase or decrease production.

Promising outcomes have been shown when yield prediction is done using machine learning methods. Artificial neural networks, decision trees, support vector machines, random forests, and other machine learning approaches may learn from historical yield data and make predictions that are reasonably accurate[6]. The capacity of machine learning algorithms to analyze vast amounts of data and identify trends is helpful in forecasting agricultural yields. Utilizing machine learning can lead to more accurate and exact yield forecasts, which in turn can assist farmers in making better decisions.

Data from Google Earth Engine was used in this study, which was carried out in five districts of the state of Gujarat. Crop masks were used to filter out non-agricultural areas and forest cover, ensuring accuracy in vegetation index values[7]. The study's goal is to provide the best model possible for calculating agricultural output. This project aims to help agricultural planning and resource allocation decision-making processes by utilizing the most recent information available. In the end, it hopes to make a significant contribution to Gujarat's efforts towards sustainable development and food security[8].

## **Objective**

To develop a robust crop yield estimation model for groundnut cultivation in Gujarat state using remote sensing data and machine learning techniques, aiming to provide accurate predictions of crop yield to support agricultural planning and decision-making.

* + To develop different ML models for yield prediction of Groundnut Crop
  + To evaluate the performance of the models for yield prediction.
  + To develop a web-based dashboard for feature visualisation
  + To deploy the best model for yield prediction in the web platform

## **Research Questions**

* What are the main determinants of Gujarat's groundnut yield, and how can they be measured?
* Which machine learning techniques work best for estimating the link between different influencing factors and groundnut yield?
* What effect does the selection of features (soil moisture, vegetation indices, etc.) have on the yield estimating model's performance?
* What are the possible uses and advantages of precise crop yield estimation for various stakeholders, such as farmers, policymakers, and Gujarati agricultural extension services?

# **Literature Review**

* 1. **A Machine Learning-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning with Regression Models**[9]

The goal of this study was to develop a machine learning model for predicting agricultural productivity using six distinct regressor Model. After examination, the Random Forest Regressor was shown to be the most successful model, outperforming the others. The model's average predicted value is 468.16 units away from the actual value, according to its Mean Absolute Error (MAE) of 468.16. Furthermore, a reasonable degree of generalization ability is indicated by the Cross-Validation score of 0.6087, suggesting that the model can generate reasonably accurate predictions on data that has not yet been observed. The Random Forest Regressor is a promising method for optimizing agricultural yield estimates because of its superiority, which may be attributed to its capacity to manage complex interactions between different input variables and farm produce output.

* 1. **Groundnut Crop Yield Prediction Using Machine Learning Techniques**[10]

In order to ensure global food security, in particular, yield prediction in agriculture is crucial, as highlighted in the paper by Vimita Shah and Prachi Shah, who also support the use of machine learning techniques in this endeavor. In order to predict crop output based on soil, environmental, and abiotic factors, it analyzed groundnut data spanning eight years and assessed several techniques, such as linear regression, regression tree, k-nearest neighbour (KNN), and artificial neural network (ANN). According to the results, KNN is the best algorithm for predicting groundnut yield. The study of the literature emphasizes how machine learning can effectively and precisely forecast crop productivity, which motivates more research in a variety of crops and geographical areas.

* 1. **Crop Yield Prediction in Cotton for Regional Level using Random Forest Approach**[11]

In order to anticipate Cotton Crop Yield, this study uses machine learning regression models that take into account many geographical characteristics, including temperature, rainfall, NDVI, LST, SPI, VCI, GDD, and area. It makes use of ground-based data, agricultural yield data from 2001 to 2017, and long-term agromet-spectral variables from satellites. With great speed and dependability, the Random Forest (RF) algorithm forecasts cotton production at three distinct points prior to harvest. For final yield projections in September, December, and February, the RF model had coefficients of determination (R2) of 69%, 60%, and 39%, respectively. These results demonstrate the model's effectiveness in predicting cotton crop outcomes over a range of time periods.

* 1. **Remote Sensing Based Yield Estimation of Rice (Oryza Sativa L.) Using Gradient Boosted Regression in India**[12]

The study used a Gradient Boosted Regression (GBR) model to estimate rice yields in India at a spatial resolution of approximately 500 m. The Leaf Area Index (LAI) of the Moderate Resolution Imaging Spectroradiometer (MODIS) and district-level yield observations were used to calibrate the GBR model. The reported block-level yields and district-level observations were used to validate the downscaled and reaggregated yields. According to the study, the GBR model estimated rice yields in India with good accuracy in rainfed, water-limited agricultural settings. According to the study, the present methods for timely rice yield estimation needed by government and insurance organizations can be supplemented by this downscaling methodology of rice yield calculation utilizing GBR.

* 1. **Machine-Learning-Based Regional Yield Forecasting for Sugarcane Crop in Uttar Pradesh, India**[13]

The goal of this research is to forecast the yield of sugarcane in Uttar Pradesh (UP) using machine learning regression algorithms with moderate-resolution satellite photos. It trains four regression models (SVR, GBR, XGB, and RF) using standard MODIS data products as features. GBR and XGB models, with R2 values of 0.66 and 0.65, respectively, are the most accurate models, according to evaluation based on the R2 measure. The study highlights the effectiveness of using machine learning algorithms and satellite data to accurately anticipate sugarcane yield at a regional level, indicating possible applications in agricultural operations.

# **Data collection and Feature Engineering**

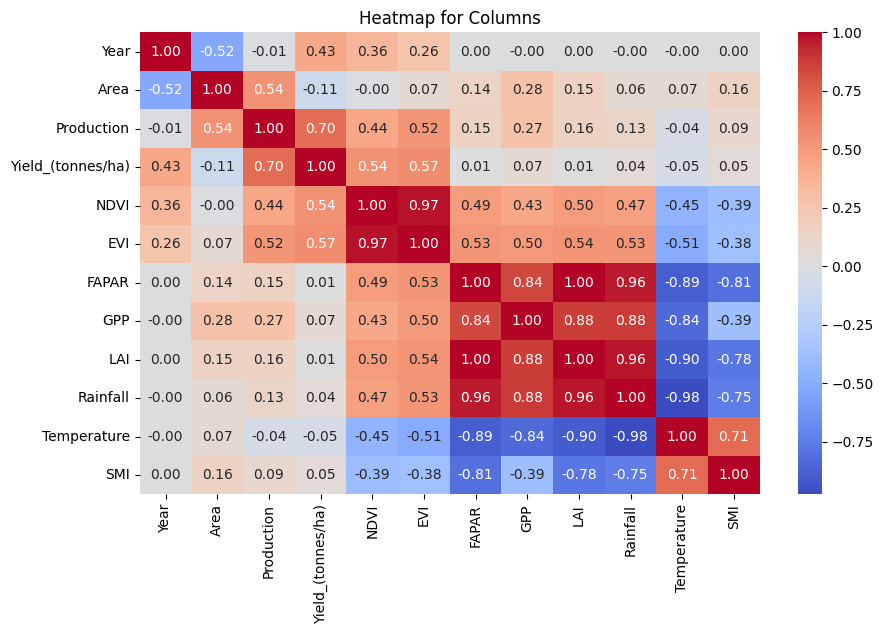
Agriculture is a multi-phase process that has three main production phases. Depending on the inherent weather patterns, farmers apply pesticides, insecticides, fertilizers, seeds, and water at certain intervals. Yield is the farm's output and has a significant impact on crop vegetation. Using vegetation indices, a Machine Learning (ML) model may predict past production and show statistical correlations between variables that impact crop productivity. The study focuses on groundnuts, a large crop that is growing and important to the agricultural ecosystem across five districts[14].

Data on different vegetation indices and meteorological parameters, such as NDVI, EVI, GPP, FPAR, LAI, LST, SMI and Precipitation were gathered for the months of June, July, August, September and mid-October from different satellite through Google Earth Engine. obtains, following the application of the crop mask, the average values for each index for each month. Data from Google Earth Engine was gathered every five months for 22 years (2000-2021) for every index. Information gathered from The Directorate of Agriculture about sowing Area, Production and Yield.

District-wise Area, Production and Yield of Important Food & Non-food Crops in Gujarat State: Directorate of Agriculture, Gujarat[15]

Data on the monthly average of a certain district for all four months during a 22-year period was immediately gathered and saved in a comma-separated values (.CSV) file from Google Earth Engine. Thus, the data is divided into 13 columns (distict\_name+8 indices Year+ Area + Production + Yield) and 110 rows (22 years \* 5dist). Data with Yield (tonne/ha) as the output and 12 characteristics as the input. Data containing additional input Numeric and Categorical variables. The Label Encoder is used to encode variables such as the district name and year; the remaining variables are Standard scalar.

It is essential to divide the dataset into training and testing sets in order to appropriately assess the performance of a machine learning model. The model learns patterns from the training data and is then assessed using test data that hasn't been seen yet, with 80% of the data allocated for training and 20% for testing[16]. This method aids in evaluating the model's capacity for generalization, guaranteeing that it functions effectively on fresh, untested data in addition to the training dataset, enhancing its dependability and efficiency (Figure 1).

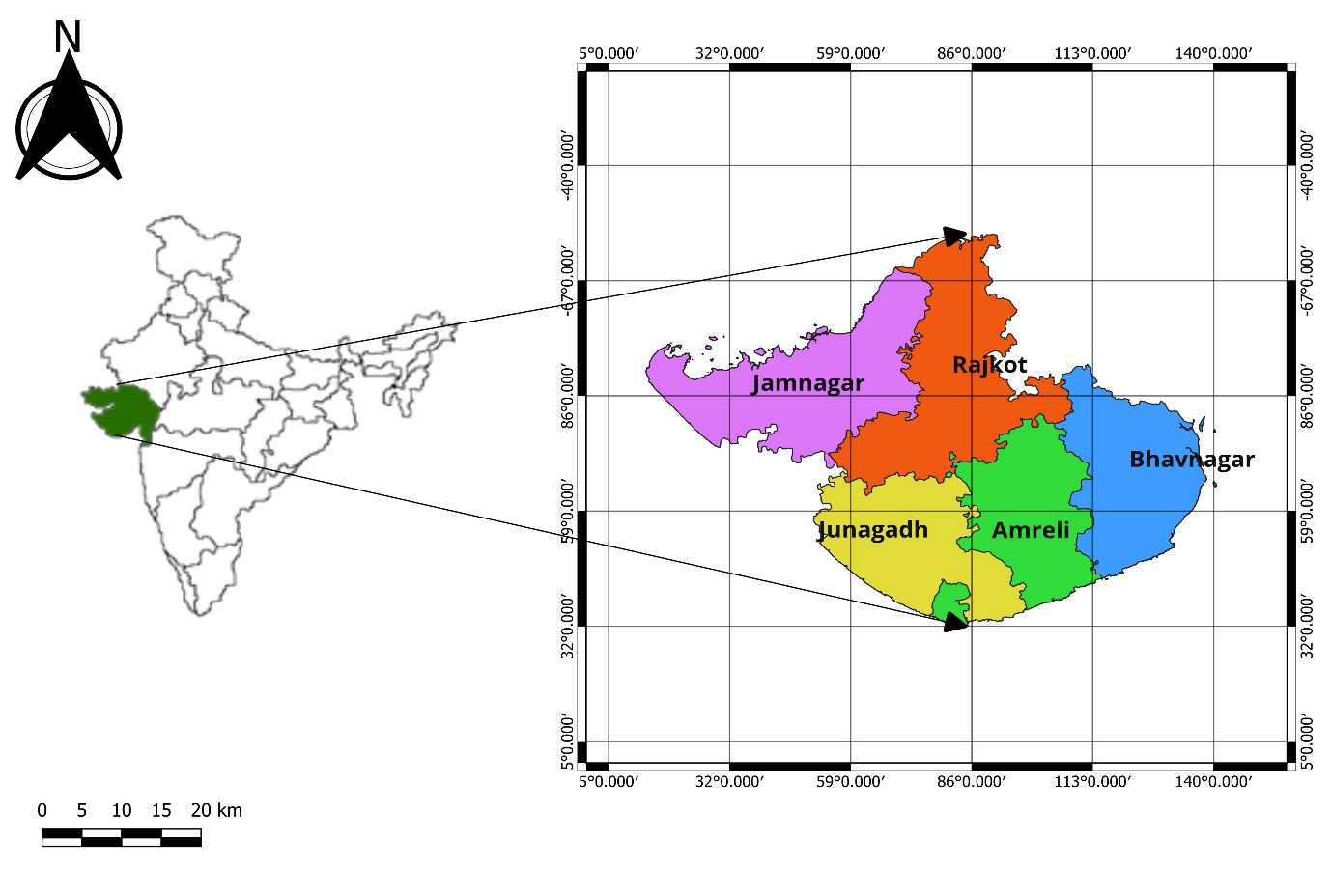


**Figure 1: Correlation: Heat Map**

# **Study Area and Materials Used**

## **Study Area**

The agricultural landscape of Gujarat's Jamnagar, Junagadh, Amreli, Rajkot, and Bhavnagar districts is shaped by a semi-arid environment with 600–900 mm of annual rainfall and temperatures between 15°C and 30°C. The economy is anchored by groundnut, which is planted from June to October and is aided by crops like cotton, castor, wheat, and pearl millet. These crops thrive in the semi-arid environments and are tolerant of the climatic challenges faced by the area. Farmers in Gujarat skill fully utilize the limited rainfall by implementing appropriate agricultural methods and water management, which greatly enhances the prosperity of the state's economy as well as the local economy (Figure 2).



**Figure 2: Study Area**

## **Satellite Data and Products**

With 36 multispectral bands and a wide 2,330 km swath width, the Moderate-Resolution Imaging Spectroradiometer (MODIS) has been carried by the Terra and Aqua spacecraft since 1999 and 2002, respectively. With its three different spatial resolutions—1000, 500, and 250 meters—MODIS enables twice-daily worldwide coverage[17]. Because of its adaptability, this instrument can record a great deal of detail and frequency in a wide range of Earth's surface characteristics and atmospheric occurrences. MODIS is essential for many applications, including as land cover monitoring, climate studies, disaster management, and agricultural assessments because of its quick revisit time and high-resolution imagery.

Approximately 44 geophysical outputs are available from MODIS, a medium-resolution sensor that can be used to analyze the atmosphere, land, and ocean. These products have been processed into high-level products for regional to global modeling and monitoring purposes. These products include vegetation indices such as the Leaf Area Index (LAI), Fraction of Photosynthetically Active Radiation (FPAR), Evapotranspiration (ET), Potential Evapotranspiration (PET) and Gross Primary Product (GPP). This satisfies end users' immediate demands and removes the need for them to process the data themselves.

The Climate Hazards Group (CHG) at the University of California, Santa Barbara (UCSB) has generated a high-resolution precipitation dataset called CHIRPS, or Climate Hazards Group InfraRed Precipitation with Station data[18]. It generates gridded precipitation estimates with precise temporal and spatial resolutions by fusing satellite data with on-the-ground measurements.

With a temporal resolution of daily data and a geographical resolution of roughly 0.05 degrees, or roughly 5 km, the CHIRPS dataset spans the whole planet. To produce precise and trustworthy estimates of precipitation, especially in areas with few ground-based observations or erratic monitoring networks, it combines infrared observations from geostationary satellites with precipitation gauge data.

CHIRPS data would be a good option for evaluating rainfall variability or carrying out research on drought, agriculture, or hydrology.

Numerous applications, such as drought monitoring, agricultural assessments, water resource management, and climate research, make extensive use of the CHIRPS dataset. It is useful for comprehending precipitation patterns, evaluating climate variability, and assisting in the decision-making processes associated with disaster preparedness and response due to its high spatial and temporal resolutions and worldwide coverage (Table 1).

|  |  |
| --- | --- |
| Satellite | Parameter |
| MODIS | NDVI, EVI, GPP, FPAR, LAI, LST |
| CHIRPS | Rainfall |
| NASA | SMI |

**Table 1: Satellite and Parameter**

## **Ancillary Crop Data**

**Groundnut** (*Arachis hypogea*) holds immense economic significance in Gujarat, with the state contributing substantially to India's overall production. During the Kharif season, Gujarat accounts for approximately 50% of India's groundnut output and cultivates around 42% of the total groundnut area in the country. Spread over roughly 2 million hectares, groundnut cultivation in Gujarat yields approximately 2.6 million tonnes annually. The Directorate of Agriculture systematically collects and maintains comprehensive data on major crops, including groundnuts, at the district level. This database encompasses information on total area, production, and productivity, offering valuable insights into the agricultural landscape. By releasing district-wise statistics on area, production, and yield of various crops, including both food and non-food crops, stakeholders such as policymakers, researchers, and farmers can make informed decisions, devise effective strategies, and monitor the performance of agricultural sectors over time.

## **Crop Mask**

Crop masking is an essential tool for index-based evaluations and yield prediction since it makes it possible to isolate and examine particular agricultural regions in satellite data[19]. Crop masks facilitate the deployment of algorithms and models specifically designed for agricultural monitoring and analysis by precisely demarcating crop fields or regions of interest.

## **Vegetation Index**

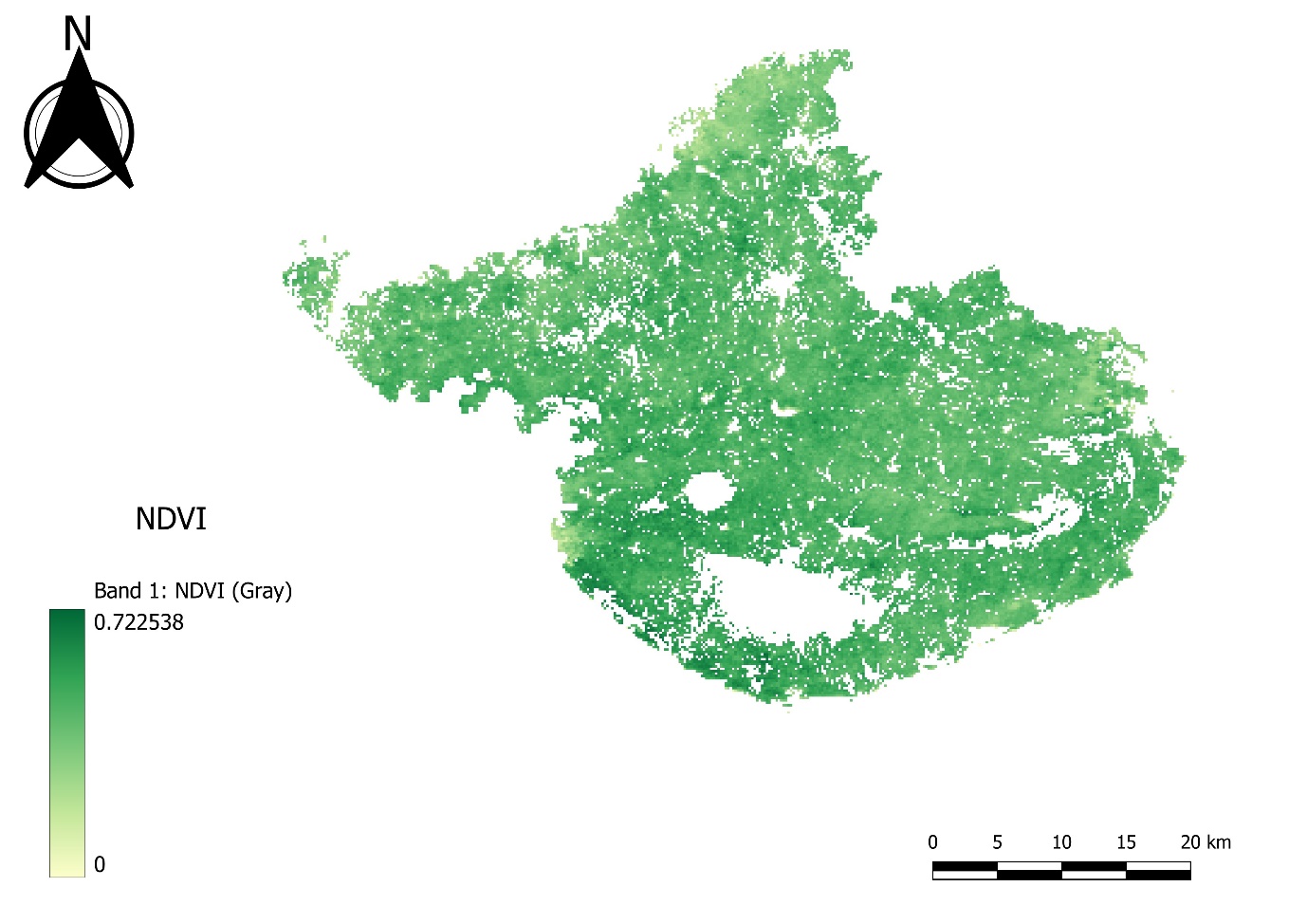
### **Normalized Difference Vegetation Index (NDVI)**

The Normalized Difference Vegetation Index is used in remote sensing to assess the health and vitality of vegetation (NDVI). It computed by comparing the amount of light reflected by plants in the near-infrared and red wavelengths. Healthy vegetation reflects more near-infrared light than red light, while damaged or sparse vegetation reflects more red light.

The ratio of the difference between near-infrared and red light to the total of near-infrared and red light is what the NDVI equation calculates. Next, a scale is applied to this ratio, ranging from -1 to +1, where higher values correspond to healthier vegetation [20](Figure 3).

**NDVI = (NIR - RED) / (NIR + RED)** [21]

NDVI is used by scientists, farmers, and government agencies to track the growth of vegetation, detect environmental stressors like drought, and make decisions about crop management and land use.



**Figure 3: Spatial Variation of NDVI within the crop Masked Area**

### **Enhanced Vegetation Index (EVI)**

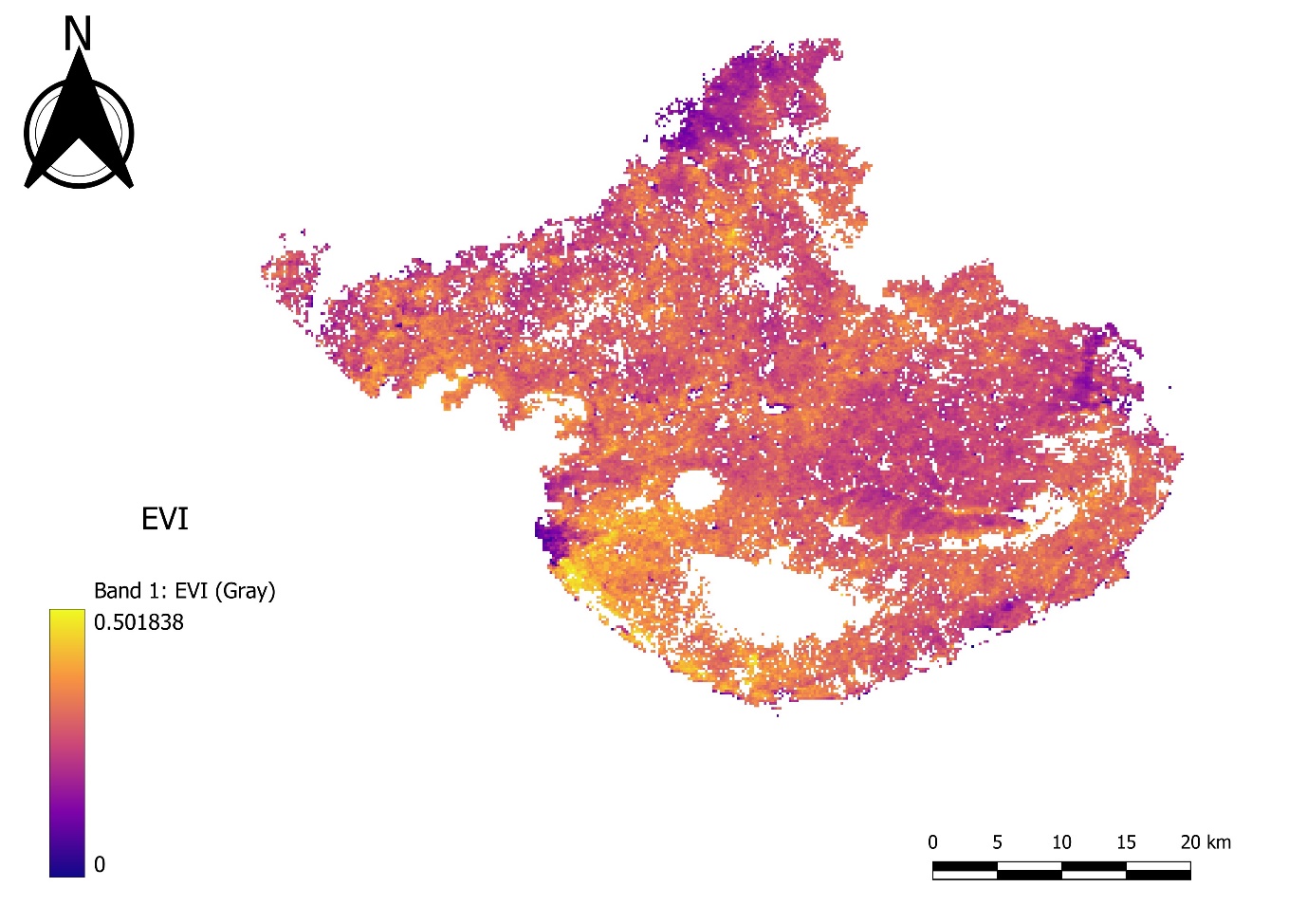
Another metric used in remote sensing to evaluate the health and vitality of plants is the Enhanced Plants Index (EVI). It was created to get around several problems with the Normalized Difference Vegetation Index (NDVI) in regions with a lot of vegetation or where the presence of soil or other atmospheric elements could distort the results.

Together with the red and near-infrared bands, the EVI equation also includes the blue band, which is susceptible to atmospheric impacts and aerosol scattering[22]. The following is the formula:

**EVI = 2.5 x ((NIR - Red) / ((NIR + 6 \* Red - 7.5 \* Blue) + 1))** [23]

The result is a value between -1 and +1, where higher numbers indicate healthier vegetation (Figure 4).

Researchers, farmers, and government agencies utilize EVI, like NDVI, to monitor vegetation growth and identify environmental stressors in places with dense vegetation or atmospheric interference. This approach can also be used to calculate plant productivity and biomass.



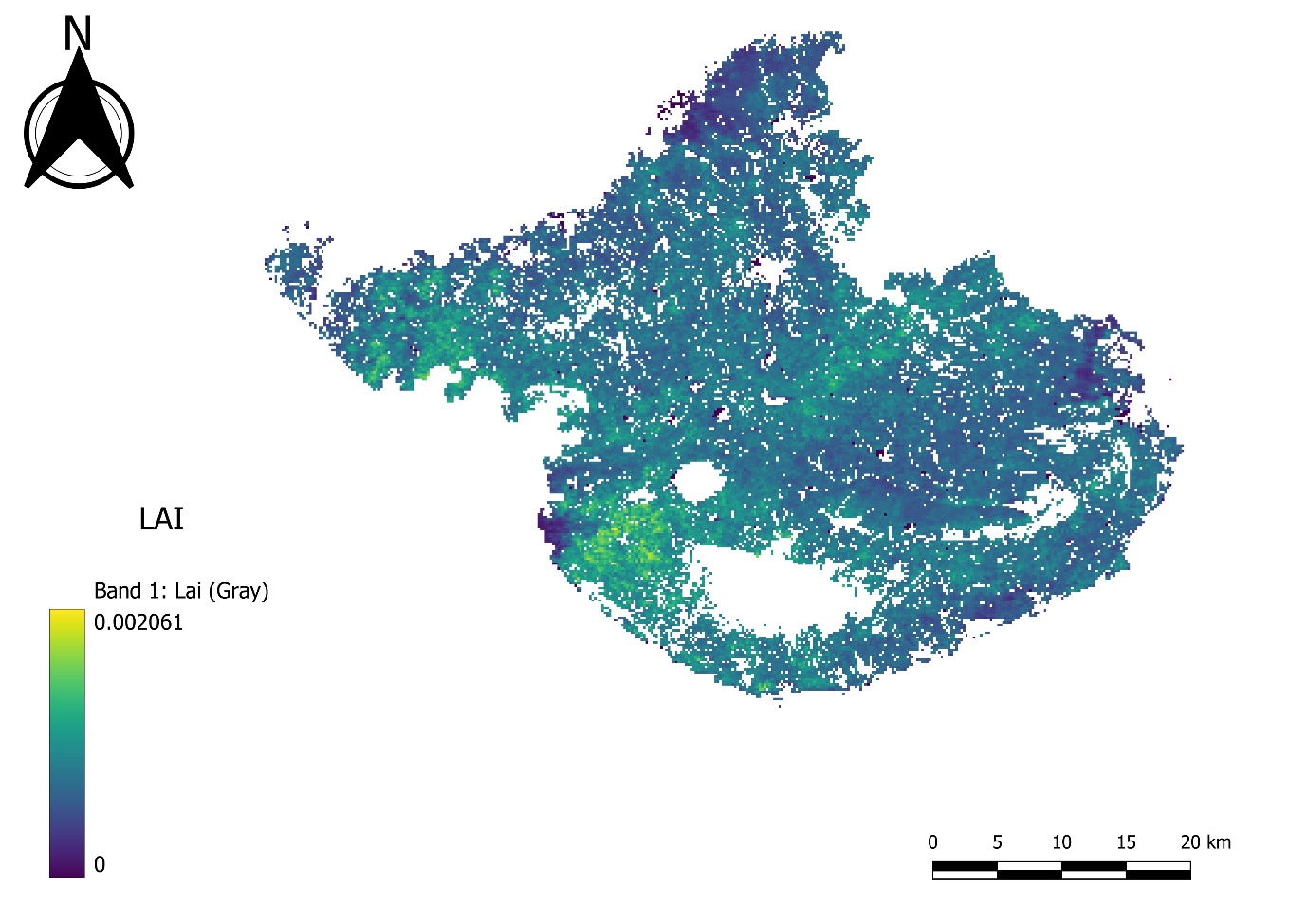
**Figure 4: Spatial Variation of EVI within the crop Masked Area**

### **Leaf Area Index (LAI)**

The measurement of green leaf area per ground unit, or Leaf Area Index (LAI), is essential for calculating the potential photosynthesis in vegetation patches. Both direct leaf area measurements and indirect approaches, such as remote sensing, are used in the computation of LAI[24]. Accurate evaluation of vegetation productivity and health is made possible by these techniques, which is essential for comprehensing ecosystem dynamics and agricultural management. When making judgments about environmental monitoring and land use planning, LAI offers insights into plant growth, carbon sequestration, and ecosystem functioning.

**LAI = Leaf Area / Ground Area** [25]

The range of the LAI scale is 0 (nothing) to more than 10. (Thick woods of conifers). Seasonality also affects LAI, with values normally peaking during the growing season and falling during senescence or dormancy (Figure 5).



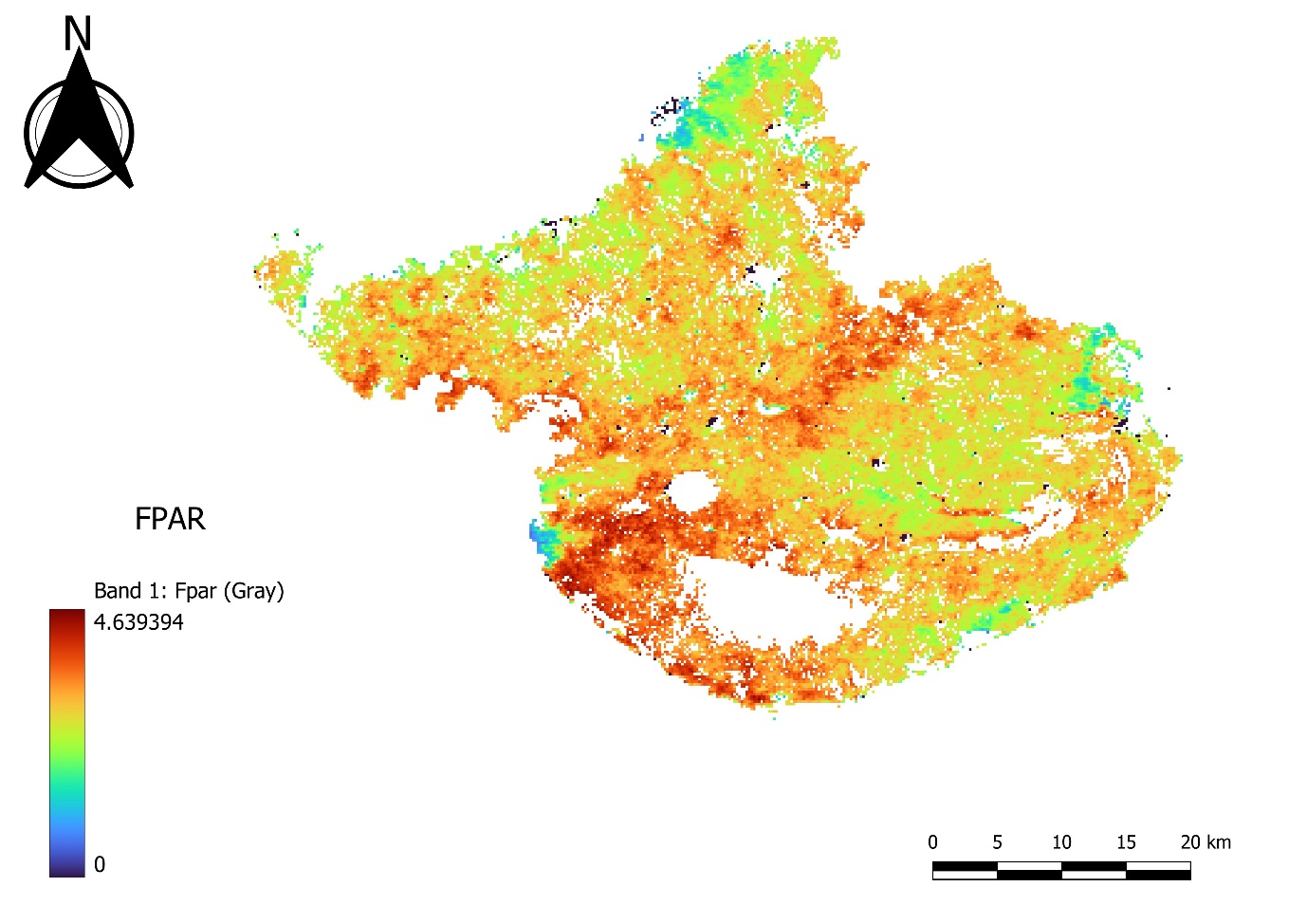
**Figure 5: Spatial Variation of LAI within the crop Masked Area**

### **Fraction Photosynthetically Active Radiation (FPAR)**

FPAR or Fraction Photosynthetically Active Radiation, is a measure of the amount of sunlight absorbed by vegetation for use in photosynthesis. This measure is important for understanding the health and productivity of vegetation. The region of the electromagnetic spectrum that plants employ for photosynthesis is known as photosynthetically active radiation (PAR), and it contains some near-infrared wavelengths as well as visible light. Usually stated as a percentage, F­PAR quantifies the portion of this PAR that is actually absorbed by vegetation.

**FPAR = (PAR absorbed by vegetation) / (incoming PAR)** [26]

There are two possible values for the Fraction of Photosynthetically Active Radiation (FPAR): 0 and 1, or 0% and 100%. When the value is 1, it means that all of the incoming PAR is absorbed by the vegetation, whereas a value of 0 means that none of it is (Figure 6).   
  
Through the detection of light reflection from vegetation, remote sensing devices are able to quantify FPAR and estimate the amount of PAR received by the plant. It is possible to watch how the productivity and health of the vegetation vary in response to climate change, deforestation, and drought.



**Figure 6: Spatial Variation of FPAR within the crop Masked Area**

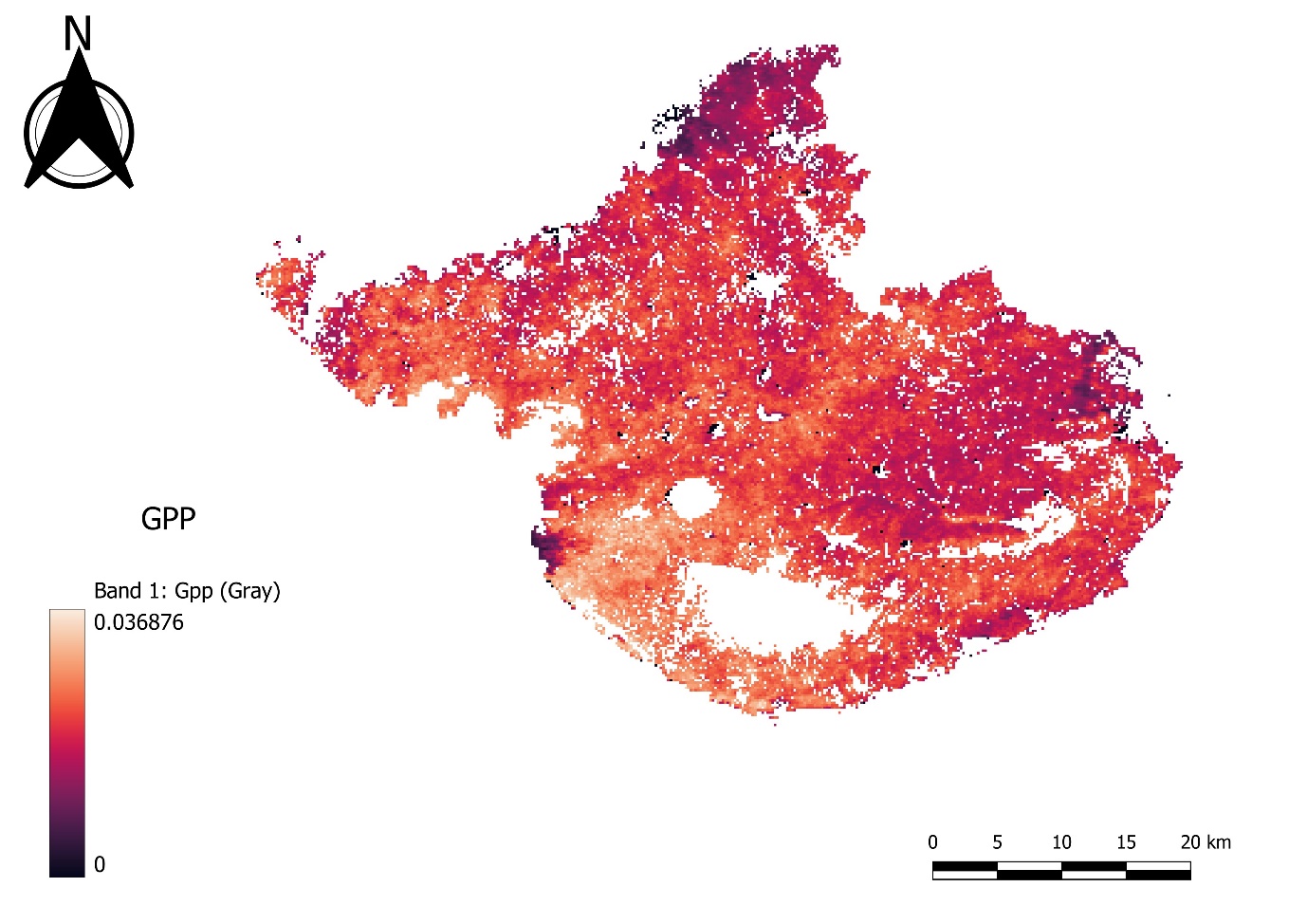
### **Gross Primary Product (GPP)**

Gross primary productivity (GPP) is the total amount of carbon fixed during photosynthesis by all producers within an ecosystem. GPP is the total quantity of organic material synthesised in a given amount of time; autotrophs use some percentage of this production.

In many ecological and agricultural models, GPP plays a key role in processes including carbon sequestration, nutrient cycling, and biomass production. Precise measurements of GPP are necessary to comprehend how vegetation and the carbon cycle on Earth are impacted by climate change [27](Figure 7).

**GPP=PAR×LAI×ϵ** [28]

The net carbon dioxide uptake by plants is the difference between the amount of carbon dioxide taken in through photosynthesis and the amount released through respiration. The entire amount of carbon dioxide absorbed during photosynthesis is known as gross primary productivity.

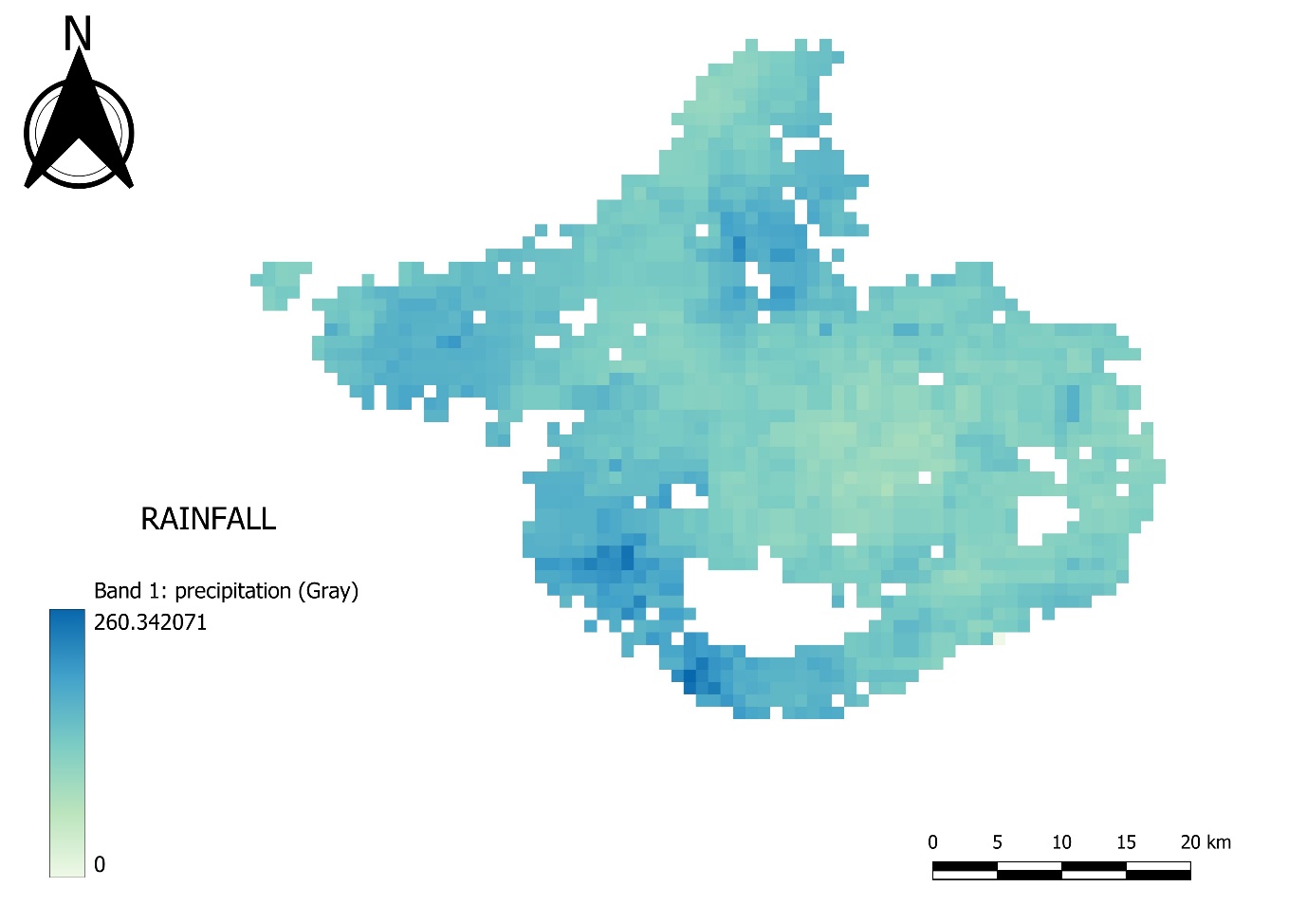


**Figure 7: Spatial Variation of GPP within the crop Masked Area**

### **Rainfall**

Rainfall has a direct impact on plant growth, development, and total productivity, making it essential for agricultural produce. Milimeter(mm) is the unit of measurement. Water availability for plant uptake, soil moisture retention, nutrient transport, and the regulation of physiological processes critical to crop development are just a few of the ways in which it is significant (Figure 8).

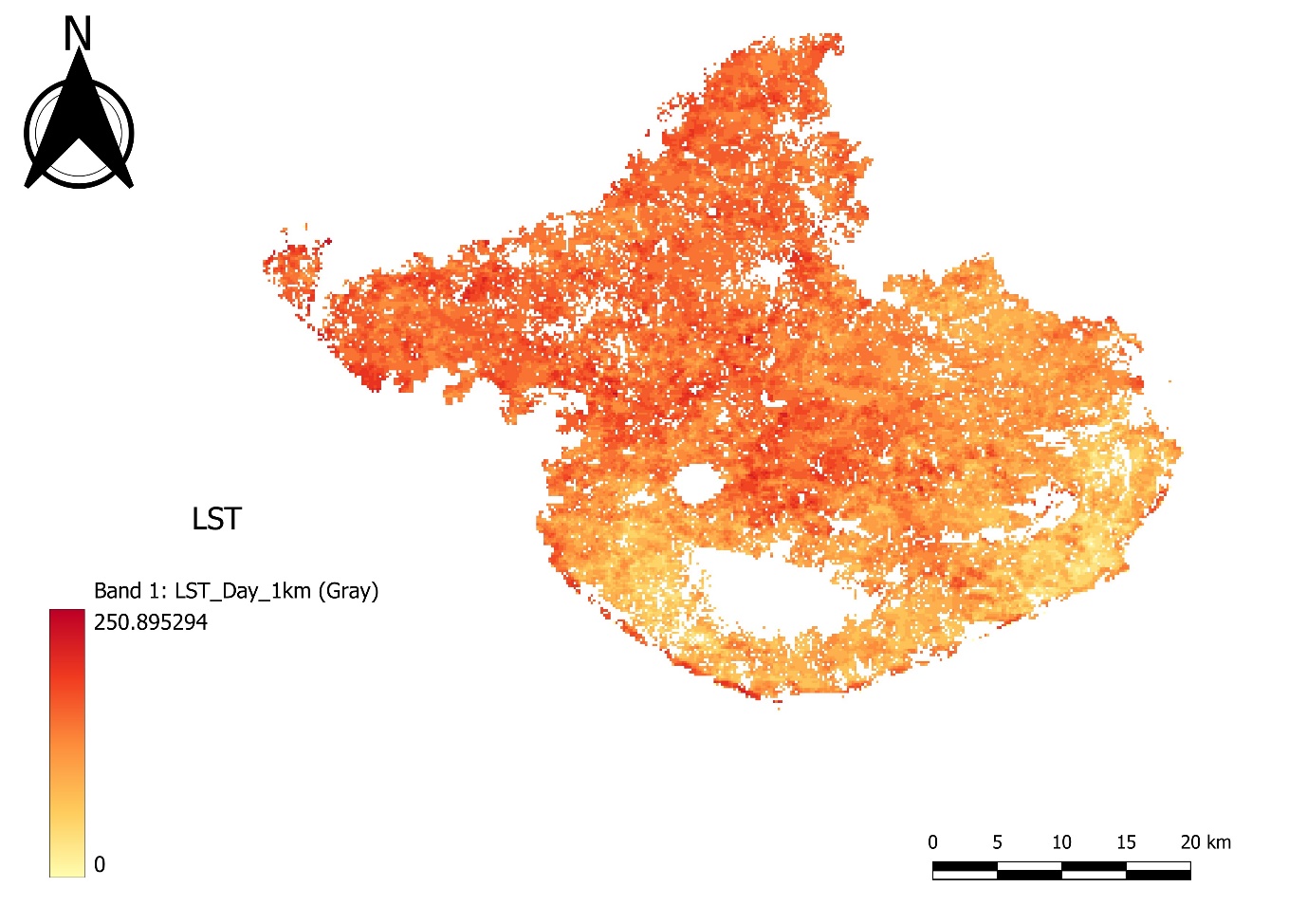
In order to provide water availability, maintain soil moisture, facilitate nutrient transport, support photosynthesis and growth, regulate temperature, control pests and diseases, and improve yield stability and resilience, rainfall is crucial to agricultural productivity. Crop planning, sustainable agricultural practices, and efficient water management measures all depend on an understanding of the significance of rainfall and how it affects crop productivity.



**Figure 8: Spatial Variation of Rainfall (mm) within the crop Masked Area**

### **Land Surface Temperature (LST)**

Since land surface temperature (LST) has a direct impact on plant growth and development, it is essential for forecasting agricultural productivity. Kelvin(K) is the unit of measurement. Elevated temperatures have the potential to cause stress to crops, resulting in lower photosynthesis, limited availability of water, and heightened vulnerability to pests and illnesses. On the other hand, too low a temperature might restrict plant development and impede metabolism. Farmers may reduce temperature-related hazards and improve productivity by monitoring LST, which assists them with crop selection, irrigation, and planting schedule optimization. LST data also helps identify areas that are affected by heat, allowing for targeted interventions like crop rotation, shade, and irrigation management to sustain productivity and guarantee food security in a changing environment (Figure 9).



**Figure 9: Spatial Variation of LST(K) within the crop Masked Area**

### **Soil Moisture Index (SMI)**

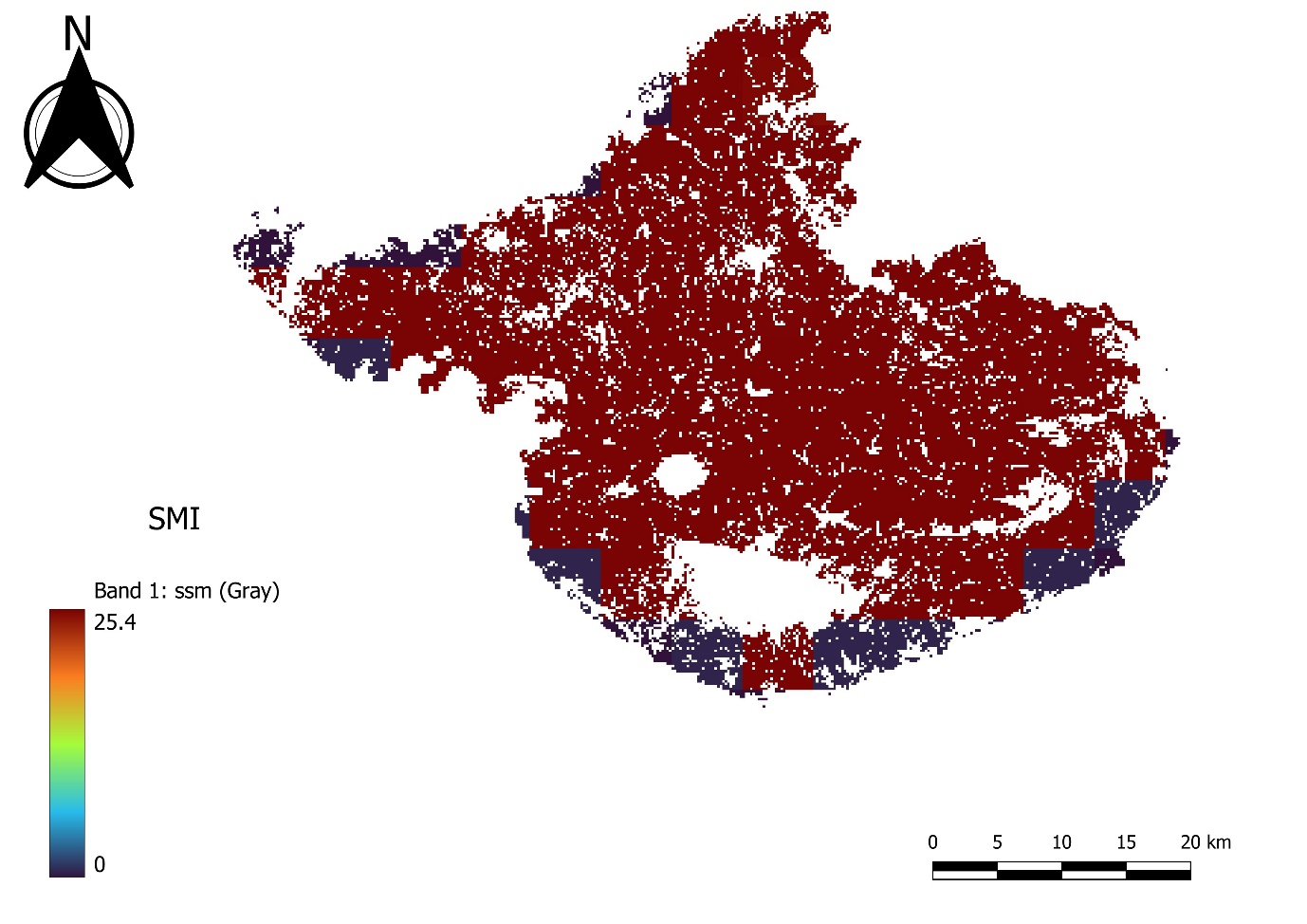
An essential indicator for estimating crop output and evaluating agricultural productivity is the Soil Moisture Index (SMI). It measures the amount of moisture in the soil, which has a direct impact on crop health overall, nutrient uptake, and plant development. The ratio of actual soil moisture to field capacity, represented as SMI, is used in the formula.

**SMI = (𝐿𝑆𝑇 𝑚𝑎𝑥 − 𝐿𝑆𝑇)/(𝐿𝑆𝑇 𝑚𝑎𝑥 − 𝐿𝑆𝑇 min)** [29]

In this case, field capacity denotes the highest quantity of water the soil can hold against gravity, whereas actual soil moisture content refers to the water content of the soil as of right now. For several physiological functions in plants, such as photosynthesis, nutrient transport, and cell division, there must be sufficient moisture in the soil. Reduced growth of plants, water stress, and eventually lower crop yields can result from inadequate rainfall. On the other hand, too much moisture in the soil can lead to nutrient leaching, root rot, and waterlogging, all of which have a detrimental effect on crop health and yield (Figure 10).

Farmers and agronomists may make well-informed judgments about crop selection, irrigation timing, and soil management techniques by keeping an eye on SMI. Enhancing water use efficiency, reducing crop losses from drought or waterlogging, and advancing sustainable farming practices can all be achieved by optimizing soil moisture levels based on SMI. SMI is therefore a useful tool for increasing crop yields while protecting soil health and water supplies.

No table of figures entries found.



**Figure 10: Spatial Variation of SMI within the crop Masked Area**

# **Methodology**

Machine learning provides computers with the ability to learn new things on their own and grow as a result of their experiences—even if they are not specifically designed to do so. The number of applications in the field of machine learning (ML) has increased due to advancements in computing.

Machine learning is becoming used more and more often globally due to its efficacy in several fields, including forecasting, defect identification, pattern recognition, and so on. Predicting yield is a significant challenge in agriculture.

Using ML, farmers will be able to forecast their crop's yield before planting it in the field, enabling them to make well-informed decisions. It helps farmers for better decision making with optimum resourse utilization and avoid over use of resourses like fertilizers, pesticides, irrigation and many more.

**STEP 1**

**STEP 3**

**Data Collection**

**(GEE)**

**Data Preparation**

**Understand The Problem Statement**

**STEP 2**

**STEP 6**

**STEP 4**

**Training The Model**

**Hyperparameters Tunning**

**Evaluation of Model**

**STEP 5**

**STEP 7**

**STEP 8**

**Model Deploy**

**(Stream lit)**

**Dashboard Creation**

**(looker studio)**

## **5.1. Algorithms (Models)**

### **5.1.1. Training the Model**

The machine learning model is trained on the training dataset once the data has been prepared by processing and splitting it into training and testing sets. Because the prepared data is used to educate the model how to identify patterns and make predictions, this is an important phase in the machine learning process. By doing this, the model is able to learn from the data and do its designated task. Over time, as the model is taught, its ability to predict results improves.

Regression analysis is employed in this procedure to utilize the model to predict a continuous dependent variable based on a collection of independent inputs. Nine distinct regression models, were utilized to train the data. These included the following: SVR, Ridge Regressor, Lasso Regressor, AdaBoost, Gradient Boosting Regression, Random Forest Regression, XGBoost Regression, and Linear Regression. Based on the independent factors included in the dataset, each of these models employs a different methodology to provide predictions.

The machine learning model learns from the patterns in the data and improves its ability to make precise predictions by being trained on the data. This is a crucial stage in the machine learning process since it enables us to use data to forecast outcomes and obtain understanding of a range of real-world issues.

#### **Linear Regression**

The relationship between a numerical output and one or more explanatory variables—also referred to as dependent and independent variables—is described in statistics using a technique called linear regression[30]. When there are one or more inputs, an iterative approach can be used to reduce the model's error on the training set of data, hence improving the coefficient values. Known as Gradient Descent, the process starts with random values for every coefficient. The coefficients are updated in the direction of reducing the error using a learning rate as a scaling factor. The procedure is repeated until the total squared error is as low as possible or there is no more room for improvement.

#### **Decision Tree Regression**

A machine learning technique called decision tree regression represents decisions and their potential effects using a structural framework. It is a supervised learning algorithm suitable for categorical and continuous output variables[31]. When it comes to regression, continuous-valued outputs are predicted using a regression tree as opposed to discrete outputs.   
  
One possible issue with decision tree regression is overfitting. This may occur when an excessively complex model matches too closely to the training set of data, resulting in subpar performance when applied to newly created, untested data. To avoid overfitting, it is essential to carefully adjust the model's parameters and use techniques like pruning to decrease the tree structure.

#### **Gradient Boosting Regression**

Gradient boosting is a potent machine learning technique that lowers mistakes and raises prediction accuracy. It functions by merging several poor predictors or models, which are subsequently trained one after the other to fix the errors of the earlier models.   
Gradient boosting has the advantage of being able to predict both continuous and categorical variables[32]. This implies that models that forecast variables like as home prices or the likelihood that a client will purchase a particular item can be constructed using it.

Gradient boosting operates through the optimization of a loss function. The accuracy with which the model predicts the target variable is gauged by this function. By changing the parameters of the weak predictors, the model attempts to minimize the loss function, which eventually raises the overall model's accuracy.

#### **Random Forest**

A machine learning approach called random forest regression makes predictions about data by utilizing several decision trees. The algorithm trains these decision trees using a method known as Bootstrap and Aggregation or bagging. Because of their high levels of variability, the decision trees in the random forest could not always produce precise predictions. On the other hand, the algorithm can generate more dependable and accurate predictions by aggregating the output of multiple decision trees[33].

Subsets of data are chosen at random and utilized to train each decision tree in the random forest regression bagging technique. In doing so, each tree is trained on a marginally distinct set of data, assisting in the reduction of overfitting and raising the general model's accuracy.   
The average of all the outputs from each individual decision tree is the ultimate result for a regression problem once the random forest has been trained. This aids in further minimizing any possible biases or mistakes in the forecasts.

#### **XGBoost Regressor**

XGBoost is a machine learning method that predicts data by using ensemble learning. In ensemble learning, the output of several models is combined to provide a final prediction that is more reliable and accurate. The individual models in XGBoost are referred to as "base learners." The final prediction is derived from combining the predictions of these base learners, which are trained on various subsets of the data[34].

The ability of XGBoost to accommodate base learners that routinely perform poorly at outcome prediction is one of its distinctive qualities. To do this, the contributions of each base learner are weighted according to their performance, giving the better performers a bigger influence on the final forecast. XGBoost culminates in a solitary model that surpasses the accuracy and dependability of every single foundation learner. This makes it a well-liked option for many applications, including as analyze consumer behaviour, predicting market prices, and diagnosing illnesses.

#### **Support Vector Regression (SVR)**

A supervised learning technique called Support Vector Regression (SVR) is utilized for regression problems and works especially well with high-dimensional and non-linear data. SVR is tailored for regression and functions similarly to Support Vector Machines (SVM) for classification. In order to minimize prediction errors, it seeks to identify a function that most closely approximates the mapping of input variables to the target variable[35].

SVR adds a margin of tolerance, permitting a given amount of error or deviation from the target variable, in contrast to classic regression models, which seek to reduce errors between predicted and actual values. SVR finds a subset of data points that are essential for establishing the regression function. The model's performance is mostly determined by these support vectors, which are closest to the regression hyperplane. To avoid overfitting, SVR incorporates a regularization parameter that strikes a compromise between maximizing the margin of tolerance and lowering prediction errors.

#### **Ridge Regressor**

Modeling the association between a dependent variable and one or more independent variables is done using the linear regression approach known as "ridge regression." When there is multicollinearity among the predictor variables, it works especially well[36].

The ordinary least squares (OLS) objective function gains a regularization element thanks to Ridge Regression. Large coefficients are penalized by this term, which lessens the effect of overfitting and multicollinearity. Ridge Regression "shrinks" the parameter estimates towards zero, hence lowering their variance, by penalizing big coefficients. This enhances the model's stability and capacity for generalization.

A tuning parameter (λ, also known as alpha) governs the degree of regularization in Ridge Regression. Stronger regularization and more shrinkage are the outcomes of higher values of λ. Variance Bias Ridge Regression optimizes the model's overall performance by balancing the reduction of model complexity (bias) and the minimization of prediction mistakes (variance).

Ridge Regression offers a closed-form solution, which facilitates easy implementation and computational efficiency.

#### **Lasso Regressor**

A regularization term is added to the ordinary least squares (OLS) method in the Lasso regression, also known as the Least Absolute Shrinkage and Selection Operator. In addition to penalizing the absolute size of the regression coefficients, it seeks to minimize the sum of squared residuals (as in OLS)[37].

Lasso regression efficiently carries out feature selection by automatically choosing a subset of the most pertinent predictors. This helps to improve model interpretability and decrease overfitting, and it is especially helpful in high-dimensional datasets with numerous predictors.

Lasso regression is a potent method for creating parsimonious models. It works especially well with datasets that require interpretable models with the ability to pick variables.

#### **5.1.1.9. AdaBoost Regressor**

An example of an ensemble learning algorithm for regression tasks is AdaBoost Regressor, which combines several weak learners to produce a powerful predictive model. This is how it operates:

AdaBoost Regressor begins by fitting the training set to a weak learner, which is usually a shallow-depth decision tree. Simple models that do marginally better than chance are these poor learners. Following the training of the first weak learner, AdaBoost gives occurrences that the prior weak learner mis predicted a higher weight. This makes it possible for later weak learners to concentrate more on the difficult data points. AdaBoost Regressor trains a series of weak learners iteratively in sequential learning. Every iteration prioritizes the cases that were incorrectly categorized in the preceding iteration[38].

Following the training of all weak learners, AdaBoost Regressor aggregates their predictions using a weighted total. The performance of each weak learner during training determines the weight of its prediction. The weighted total of each weak learner prediction is the AdaBoost Regressor model's final prediction, which yields a reliable and precise regression model.

AdaBoost Regressor is renowned for being easy to use, adaptable, and efficient when solving challenging regression issues. When compared to individual weak learners, it is less likely to overfit and frequently performs well with only minor tuning.

# **Result and Discussion**

## **Evaluation of the Model**

To determine a machine learning model's capacity for generalization, its performance on untested data must be evaluated after training. Because the model is too familiar with the training set, using the same data for testing could lead to an unduly optimistic outcome. Separate testing data is therefore essential for an objective assessment. R2 Score, which quantifies the percentage of the target variable's variance that the model can account for, is one of three often used metrics for this evaluation. It offers a standardized way to quantify model performance, making it possible to compare results from various datasets and models and guaranteeing accurate evaluation of prediction accuracy and efficiency.

## **Model Assessment**

In a regression model, the R2 score measures the percentage of the dependent variable's variance that can be attributed to the independent variable or variables. The formula for calculating it is the ratio of the total squared differences between the actual and projected values to the total squared differences between the actual and mean values. A higher R2 score, which ranges from 0 to 1, denotes a better fit between the model and the data, with 1 denoting a perfect fit.

**Equation: R2 score = 1 - (sum of squared differences between predicted and actual values / sum of squared differences between actual values and their mean)**

While a perfect score of 100% indicates a perfect fit, there is no set standard for an ideal R2 score, and models with low scores can still be useful.

Nine regression models are used in this work to provide exact yield predictions: Linear Regression, Random Forest Regression, Decision Tree Regression, XGBoost, Gradient Boosting Regression, Lasso Regressor, Ridge Regressor and SVR. Out of all of them, XGBoost routinely produces forecasts with the best R2 Score and accuracy (Figure 11). Strong ensemble learning algorithms like XGBoost are excellent at handling intricate relationships and identifying minute patterns in the data. For precise yield predictions, its scalability, resilience, and capacity for performance optimization make it the go-to option. XGBoost outperforms other models, indicating that it is capable of reliably predicting agricultural yield in the relevant areas (Table 2).

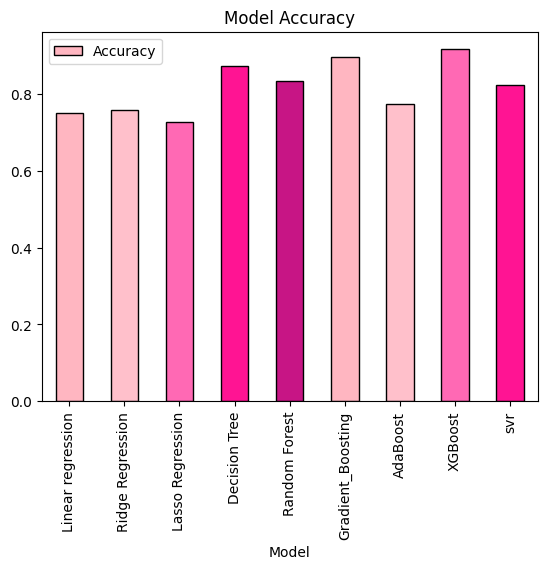
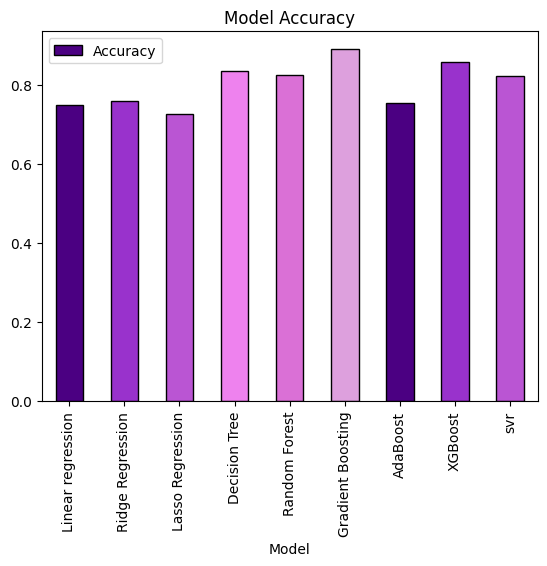
## **Hyperparameter Tuning**

Optimizing model performance, enhancing generalization, cutting down on resource waste, and guaranteeing adaptability across various datasets and domains all depend on hyperparameter adjustment. It's an important phase in the machine learning process that can result in notable gains in the efficacy and quality of the model. XGBoost model performs best at learning rate (0.1), Maximum depth (3) and n-estimator (300) (Figure 12).

**The outcomes of these ML models are as follows**

**Table 2: Accuracy Score**

|  |  |
| --- | --- |
| ML MODEL NAME | R2 SCORE |
| Linear Regression | **0.749976** |
| Random Forest Regression | **0.832916** |
| Decision Tree Regression | **0.872531** |
| XG Boost Regressor | **0.915705** |
| Gradient Boosting Regression | **0.896246** |
| Lasso Regressor | **0.727212** |
| Ridge Regressor | **0.758049** |
| Support Vector Regressor (SVR) | **0.823192** |

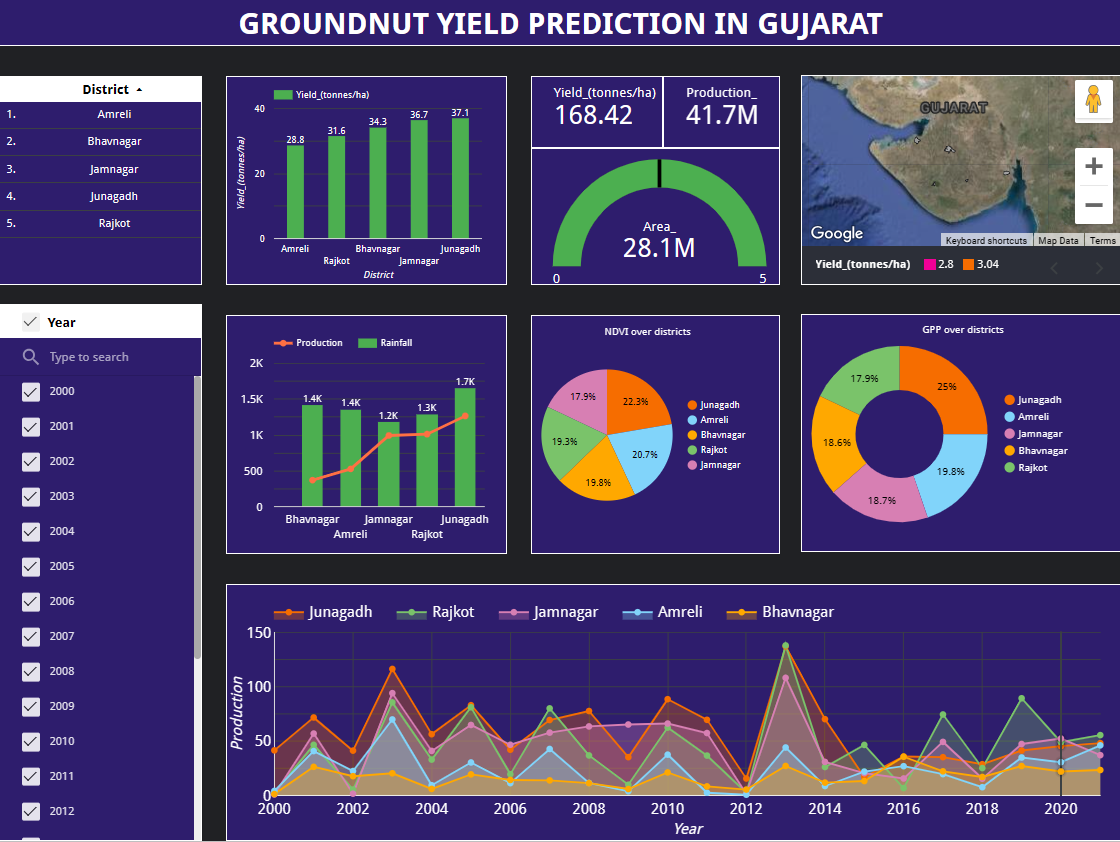


**Figure 12: Accuracy Assessment: After Hyperparameter Tunning**

**Figure 11: Accuracy Assessment: Before Hyperparameter Tunning**

## **Dashboard**

A robust platform for building, modifying, and sharing interactive dashboards that facilitate data-driven decision-making within and between enterprises is offered by Looker Studio. It is a well-liked option for companies wishing to efficiently visualize and analyze their data because of its user-friendly interface, strong integration possibilities, and sophisticated analytics tools (Figure 13).



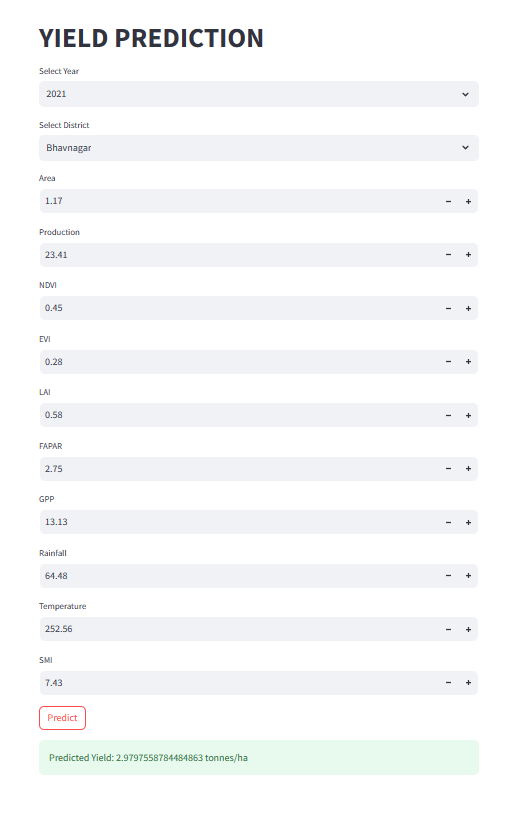
**Figure 13: Dashboard**

<https://lookerstudio.google.com/reporting/ec76b4e8-2a47-4e96-95e1-a543a5ad3852>

## **Model Deploy**

Using the pickle module in Python, one can deploy a model by saving the trained model to a file and loading it into memory when making predictions.

Machine learning models can be easily and quickly deployed to production by using a pickle file. But, pickle files should be loaded with caution due to security issues; if not handled appropriately, they can run arbitrary code and present a security risk (Figure 14).



**Figure 14: Model Deploy**

# **Conclusion**

In the districts of Junagadh, Jamnagar, Amreli, Bhavnagar, and Rajkot in Gujarat, groundnut is an important oilseed crop. Accurate crop forecast using various algorithms is beneficial. Precise projections play a crucial role in enhancing agricultural practices and guaranteeing effective output, so augmenting the agricultural productivity and economic stability of the area.

In order to anticipate groundnut yield, we combine five vegetation indicators (NDVI, EVI, GPP, FPAR, and LAI) with meteorological variables including rainfall, land surface temperature (LST), and soil moisture index (SMI).

To predict yield, nine regression models were used; XGBoost consistently produced the highest accuracy and R2 Score. It is skilled at identifying intricate data patterns thanks to its robust ensemble learning capability. When it comes to accurate agricultural yield forecasting, XGBoost stands out due to its scalability and speed optimization.

The efficacy of the yield estimating model for groundnut crops is greatly impacted by the features that are chosen, such as vegetation indices and soil moisture. These elements have a direct impact on the growth, development, and general health of plants, which impacts yield. Predictions become more accurate as pertinent features are added, as this improves the model's capacity to represent the complexity of groundnut agriculture. On the other hand, leaving out important details could cause an overestimation or underestimating of yield, which would reduce the model's usefulness for making agricultural decisions.

Accurate assessment of crop production is crucial for various stakeholders, including farmers, policymakers, and Gujarati agricultural extension agencies. Farmers are adept at-risk management, financial planning, and resource allocation. Yield projections are used by policymakers to prepare for food security, formulate policies, and adjust to climate change. Based on precise production projections, agricultural extension services improve capacity building, encourage technology uptake, and offer customized advice services. In the end, accurate yield estimation guarantees knowledgeable decision-making, raises productivity, and promotes all-around sustainable agriculture.

# **References**

[1] “Groundnut: Multifarious utilities of the ‘King of Oilseeds’”.

[2] “Value addition, Price spread and Profit margins along the Groundnut value chain- A study in Anantpur district of Andhra Pradesh”.

[3] “Study on Yield Sustainability and Water Productivity of Groundnut on Farmers’ Fields through Improved Technology under Hyper Arid Partially Irrigated Zone of Rajasthan”.

[4] “Growth and Instability of Major Oilseed Crops in Gujarat”.

[5] “Forecasting of Crop Yield using Remote Sensing Data, Agrarian Factors and Machine Learning Approaches”.

[6] “Crop Yield Prediction using Machine Learning and Deep Learning Techniques”.

[7] “Land Cover classification of Punjab state using Sentinel-2 data and Machine Learning within the Google Earth Engine Cloud Platform”.

[8] “Five Necessary Policy Changes to Help Achieve Improved Nutrition and Sustainable Agriculture through Smallholder Vegetable Horticulture.”.

[9] B. Panigrahi, K. C. R. Kathala, and M. Sujatha, “A Machine Learning-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning With Regression Models,” *Procedia Computer Science*, vol. 218, pp. 2684–2693, 2023, doi: 10.1016/j.procs.2023.01.241.

[10] “Groundnut Crop Yield Prediction Using Machine Learning Techniques”.

[11] N. R. Prasad, N. R. Patel, and A. Danodia, “Crop yield prediction in cotton for regional level using random forest approach,” *Spat. Inf. Res.*, vol. 29, no. 2, pp. 195–206, Apr. 2021, doi: 10.1007/s41324-020-00346-6.

[12] P. Arumugam, A. Chemura, B. Schauberger, and C. Gornott, “Remote Sensing Based Yield Estimation of Rice (Oryza Sativa L.) Using Gradient Boosted Regression in India,” *Remote Sensing*, vol. 13, no. 12, p. 2379, Jun. 2021, doi: 10.3390/rs13122379.

[13] A. Nihar, N. R. Patel, and A. Danodia, “Machine-Learning-Based Regional Yield Forecasting for Sugarcane Crop in Uttar Pradesh, India,” *J Indian Soc Remote Sens*, vol. 50, no. 8, pp. 1519–1530, Aug. 2022, doi: 10.1007/s12524-022-01549-0.

[14] “Machine learning methods for crop yield prediction and climate change impact assessment in agriculture”.

[15] “District-wise Area, Production and Yield of Important Food & Non-food Crops in Gujarat State: Directorate of Agriculture, Gujarat”.

[16] “Train-Test Split for Evaluating Machine Learning Algorithms”.

[17] “Terra (Satellite) - an overview”.

[18] “CHIRPS: Rainfall Estimates from Rain Gauge and Satellite Observations”.

[19] “Effect of using crop specific masks on earth observation based crop yield forecasting across Canada”.

[20] “NDVI and its Practical Applications in Agriculture”.

[21] “Index: Normalized Difference NIR/Red Normalized Difference Vegetation Index, Calibrated NDVI - CDVI”.

[22] “Development of a two-band enhanced vegetation index without a blue band”.

[23] “Comparative analysis of different vegetation indices with respect to atmospheric particulate pollution using sentinel data”.

[24] “An indirect method of estimating leaf area index in a tropical deciduous forest of India”.

[25] “Ground‐based measurements of leaf area index: a review of methods, instruments and current controversies”.

[26] “Fraction of photosynthetically active radiation (FPAR) absorbed by corn canopies at different growth stages.”.

[27] “Allocation of gross primary production in forest ecosystems”.

[28] “The relationship between GPP and PARa at each of three grassland sites”.

[29] A. Saha, M. Patil, V. C. Goyal, and D. S. Rathore, “Assessment and Impact of Soil Moisture Index in Agricultural Drought Estimation Using Remote Sensing and GIS Techniques,” in *ECWS-3*, MDPI, Nov. 2018, p. 2. doi: 10.3390/ECWS-3-05802.

[30] “What is Linear Regression?”.

[31] “Decision Tree – A Step-by-Step Guide”.

[32] “Gradient Boosting Algorithm: A Complete Guide for Beginners”.

[33] “Random Forest - an overview”.

[34] “XGBoost: A Scalable Tree Boosting System”.

[35] “Support vector regression based models to predict fracture characteristics of high strength and ultra high strength concrete beams”.

[36] “Ridge Regression and Multicollinearity: An In-Depth Review”.

[37] “Lasso & Ridge Regression | A Comprehensive Guide in Python & R”.

[38] “AdaBoost Algorithm: Understand, Implement and Master AdaBoost”.

­

# **Appendix**

