**SWEET SAGE - SUGARCANE YIELD PREDICTION**

*Submitted in partial fulfillment of the Degree of*

**Bachelor of Technology**

**in**

**Computer Science Engineering**



Department of Computer Science

Jaypee Institute of Information Technology

Sector- 62, Noida

**Submitted by: Supervisor:**

**Nikita Bansal: 21103069 Dr. Kavita Pandey**

**Samyak Jain: 21103075**

**Ribha Nishal: 21103098**

**B.Tech 5th Semester**

**December 2023**

**Table of Contents**

**Chapter No. Topics Page No.**

**Chapter-1 Introduction 7-13**

1.1 General Introduction

1.2 Problem Statement

1.3 Novelty of the problem

1.4 Empirical Study

1.5 Brief Description of the Solution Approach

1.6 Comparison of existing approaches to the problem

**Chapter-2 Literature Survey 14-22**

2.1 Summary of papers studied

2.2 Integrated summary of the literature studied

**Chapter-3 Requirement Analysis and Solution Approach 23- 39**

3.1 Overall description of the project

3.2 Requirement Analysis

3.3 Solution Approach

**Chapter-4 Modeling and Implementation Details 40-49**

4.1 Design Diagrams

4.1.1 Use Case Diagrams

4.1.2 Class Diagrams/Control Flow Diagrams

4.1.3 Sequence Diagrams/Activity Diagrams

4.2 Implementation details and issues

4.3 Risk Analysis and Mitigation

**Chapter-5 Testing 50-52**

5.1 Testing Plan

5.2 Component decomposition and type of testing required

5.3 List all test cases in the prescribed format

5.4 Error and Exception Handling

5.5 Limitations of the solution

**Chapter-6 Findings, Conclusion, and Future Work 53-58**

6.1 Findings

6.2 Conclusion

6.3 Future Work

**References IEEE format 59-59**

**DECLARATION**

We hereby declare that reported in the B.Tech Minor Project-1 entitled **“Sweet Sage”** submitted at **Jaypee Institute of Information Technology, Noida, India** is our own work, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

**Place: Signature:**

**Department of Computer Science** Nikita Bansal (21103069)

**Jaypee Institute of Information Technology, Noida** Samyak Jain (21103075)

Ribha Nishal (21103098)

**Date:**

**CERTIFICATE**

This is to certify that the work titled **“Sweet Sage”** submitted by **“Nikita Bansal, Samyak Jain, and Ribha Nishal”** in partial fulfillment for the Bachelor of Technology in Computer Science of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature

**Dr. Kavita Pandey**

**ASSISTANT PROFESSOR**

Date :

**ACKNOWLEDGEMENT**

We want to express our heartfelt gratitude to those who played a vital role in successfully completing this research project. We sincerely appreciate **Dr. Kavita Panday ma’am** for her invaluable guidance and inspiration throughout the project. The dedication and hard work invested in this endeavor were significant, Her meticulous supervision made the implementation possible.

Special thanks also go to our Seniors whose contributions, big or small, greatly influenced the preparation of this project report. We are truly indebted to them for their support.

This study has not only deepened our understanding of the subject but has also opened up new avenues of knowledge. We are confident that the insights gained will continue to benefit us in our future endeavors.

Signature

Nikita Bansal (21103069)

Samyak Jain (21103075)

Ribha Nishal (21103098)

Date:

**SUMMARY**

In this project, we utilized NDVI data and weather data to forecast sugarcane yield in the sugarcane belt districts of Uttar Pradesh. Employing Multiple Linear Regression (MLR) and Random Forest (RF) models, we computed yield predictions based on Sentinel-2 satellite imagery from Google Earth Engine and weather data from the ERA-5 Land dataset by ECMWF (2015-2023). The investigation involved analyzing expected versus actual yields, comparing model performances, and visualizing NDVI time series data annually.

The preprocessing phase included merging NDVI and meteorological data, interpolating values, and converting time series data into a supervised format suitable for machine learning algorithms. The primary objective was to accurately predict sugarcane yields in Uttar Pradesh districts. Mean Squared Error (MSE) served as a comparison metric, revealing that the MLR model outperformed others, becoming the chosen model for yield estimation.

To extend the project's utility, we integrated a Flask web page for model deployment. The web page allows users to input meteorological data, obtaining accurate yield predictions. The Flask deployment, with its user-friendly interface, transforms complex models into an accessible tool. Policymakers and farmers benefit from precise production projections, supporting informed crop choices. This project underscores the significance of NDVI data, weather data, and MLR models in precision agriculture, empowering farmers to make accurate decisions about crop planning and productivity.

Nikita Bansal **Dr. Kavita Pandey**

Samyak Jain

Ribha Nishal

Date :

**CHAPTER-1**

**INTRODUCTION**

**1.1 INTRODUCTION**

Machine learning’s application in the field of agriculture is gaining immense popularity as it has shown to be a reliable and efficient method for predicting crop yield. Making accurate yield forecasts can enable farmers and policymakers to make informed decisions and sustainable choices. Regression, decision trees, and neural networks are the most popular algorithms in these fields.

Sugarcane is one of the most important cash crops in the world. Sugarcane cultivation is a major contributor to the Indian economy. In India too, it provides livelihood for a million people across the country. India is the second largest producer and exporter of sugarcane in the world. Sugarcane is used to make molasses, ethanol, and sugar. With the use of precise yield prediction made possible by machine learning models, farmers may increase the effectiveness of their crop management techniques, reduce losses brought on by under- or overproduction, and decide how best to market their produce.

The Indian state of Uttar Pradesh is well-known for its massive sugarcane farming; locals refer to this area as the "sugarcane belt." Muzaffarnagar, Saharanpur, Meerut, Baghpat, Bulandshahr, Ghaziabad, Amroha, Moradabad, Bijnor, Rampur, and Hapur are among the districts that make up this region. Sugarcane is a significant cash crop in these districts, which are situated in the western and central regions of the state. In this region, sugarcane is grown under optimal conditions thanks to the warm, humid weather and rich soil. The abundance of sugar mills and facilities for processing sugarcane in this area attests to the importance of sugarcane in this area.

The Uttar Pradesh sugarcane belt is essential to India's sugar industry, therefore any knowledge obtained by applying machine learning to forecast sugarcane yield might have a big impact on farmers, decision-makers, and the economy. Machine learning-based yield prediction for sugarcane has the potential to completely change how we calculate the yield of this valuable crop. By improving accuracy and precision in yield prediction, machine learning models can aid farmers and policymakers in making better decisions. In addition, the significance of sugarcane as a cash crop in India emphasizes the need for precise and trustworthy yield prediction to guarantee the viability of this sector.

**1.2 PROBLEM STATEMENT**

Traditional methods of sugarcane yield prediction are time-consuming and often lack accuracy. In the context of extensive research conducted in the field of sugarcane yield prediction, we aim to improve the precision and efficiency of this process. Our approach integrates different machine learning algorithms like MLR, LSTM and RF in combination with meteorological and NDVI data obtained from remote sensing technology. The primary objective is to develop a predictive model that can accurately estimate yields of Sugarcane fields of the respective Area of Interest. By doing so, we aim to provide sugarcane farmers with a valuable tool for optimizing resource management, enhancing crop productivity, and contributing to sustainable agriculture practices.

**1.3 SIGNIFICANCE OF THE PROBLEM**

The main novelty in the problem statement lies in the fact that it uses NDVI data which is a remote sensing technique that measures plant health and vitality based on the reflection of light from vegetation. Integrating NDVI with ML algorithms allows for precise monitoring of sugarcane fields, enabling farmers to implement targeted interventions for better sugarcane yield.

Traditionally, manual techniques including field observations, interviews, and surveys have been used to predict sugarcane yield. These procedures take a lot of time and effort, are subjective, and are prone to mistakes. It is additionally difficult to obtain a comprehensive picture of crop production due to the restricted spatial and temporal coverage provided by these technologies. As a result, existing methodologies' ability to predict sugarcane production are frequently criticized.

Furthermore, the growing demand for sugarcane as a cash crop in India may not be met by relying solely on conventional methods of yield prediction. The Uttar Pradesh sugarcane belt, which comprises the districts of Shamli, Muzaffarnagar, Saharanpur, and Meerut, is one of the major producers of sugarcane in the nation.

Accurate crop yield prediction can have a big impact on agriculture and related industries in numerous ways.

· Firstly, yield prediction can assist farmers in preparing for the growing season, such as when to plant, when to irrigate, and when to fertilize. To ensure that crops grow effectively and ideally, this information may be essential.

· Secondly, forecasting agricultural yields can help in making wise choices about when to harvest the crops. Farmers can use this knowledge to prevent harvesting too soon or too late, which would produce a poor yield.

· ThirdlyThirdly, yield prediction can help farmers choose which crops to sow in upcoming growing seasons by allowing them to evaluate the performance of various cultivars or hybrids.

For crops to be produced sustainably and effectively, especially in India where agriculture plays a substantial economic role, accurate yield prediction is crucial.

**1.4 EMPIRICAL STUDY**

A remote sensing technique called NDVI assesses the quantity and condition of plant cover in a given area. In the case of sugarcane, NDVI data can offer insightful information about the crop's prospective production. The difference in the amount of visible and near-infrared light reflected by the plants is used to calculate the NDVI values. These numbers can be used to calculate sugarcane crop density and growth, which in turn can be used to calculate yield.

Extraction of meteorological data from satellites using datasets like ERA-5 (European Reanalysis) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) offers several advantages and applications in various fields. ERA-5 provides global coverage, offering meteorological data on a consistent grid across the entire Earth.

Machine learning methods like LSTM, MLR, and RF models can further enhance the utilization of NDVI and meteorological data in sugarcane yield prediction. To forecast sugarcane yield, this data can be utilized as an input for machine learning algorithms. LSTM models are useful for predicting crop yields over a long period of time because they can analyze time-series data and work better when including temporal dependencies, such as the influence of weather conditions over time .

Likewise we have used Multivariate Linear Regression (MLR) which is another machine learning approach, which can be used for forecasting sugarcane yield, especially when considering multiple input features. Its simplicity lies in its ability to handle multiple input features simultaneously, providing a clear understanding of the relationships between various factors and the crop's output. Additionally, MLR demands fewer computational resources compared to more intricate models, making it an efficient and practical choice for forecasting sugarcane yields. This combination of simplicity, interpretability, and resource efficiency makes MLR a valuable tool in agricultural predictions, allowing for effective yield forecasting while maintaining a manageable computational load.

For precise yield projections, these models can examine the yield along with other pertinent factors including weather patterns and soil conditions. Sugarcane growers and policymakers can acquire trustworthy yield estimates that can support them in making decisions regarding crop management and marketing by combining NDVI data with machine learning techniques. Overall, the use of NDVI data along with meteorological data in sugarcane yield prediction is a promising strategy that could boost the sugarcane industry's productivity and profitability and hence becomes the focus of this study.

**1.5 BRIEF DESCRIPTION OF THE SOLUTION APPROACH**

The goal of the project was to create a machine learning solution to calculate the sugarcane production in kg-hectares for the districts of Uttar Pradesh that produce the most sugarcane, also known as the sugarcane belt.

The project used the GEE API to retrieve Sentinel-2 satellite imagery for the designated sugarcane belt, which was created using the coordinates in the form of longitudes and latitudes of the major districts in the sugarcane belt as vertices in order to form a polygon to accomplish this goal. The NDVI score for each image was then determined using the extracted satellite imagery. Also meteorological data gathered from GEE API, the ERA-5 Land dataset provided by ECMWF was also used.

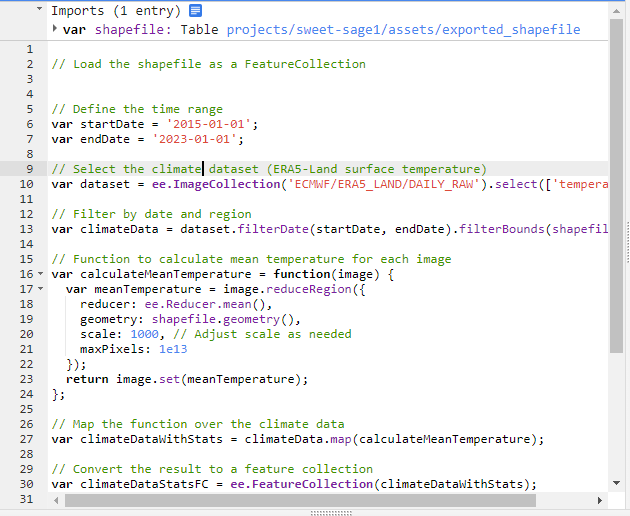
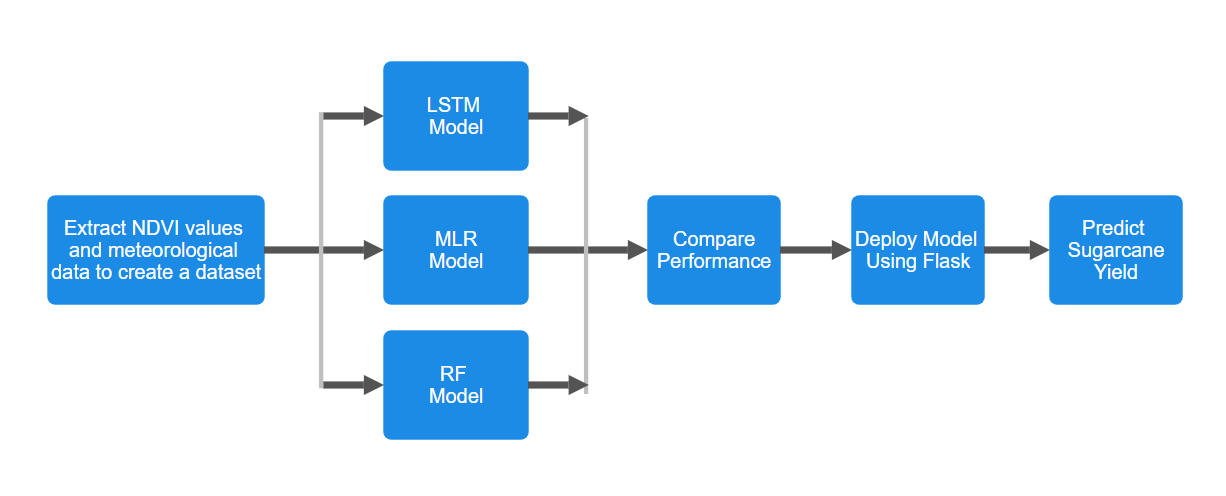


Figure 0: Extracting meteorological data

Both the weather and NDVI time series datasets, csv’s file were combined into a final csv file on the basis of interpolated NDVI and weather datas’ common parameter i.e. date of every month of years from 2015 to 2023.

The retrieved dataset was then used to train and test the LSTM, MLR, and RF models. The models were then used to accurately forecast future NDVI scores on the basis of which yield was predicted.

The yield of sugarcane in kg-hectares was forecast using expected NDVI scores, and the yield calculated using actual NDVI data was compared. The outcomes of the investigation showed that the yield forecasts using MLR algorithm were relatively accurate. ****

**Fig1:** gives an overview of the solution approach.

The MLR model was then saved and deployed using flask which provided a user-friendly interface as shown in Figure 6.10 and Figure 6.11, in which users can enter all the required meteorological data on a daily basis using which our model predicts sugarcane yield.

**1.6 COMPARISON OF THE EXISTING APPROACHES TO THE PROBLEM FOUND**

| **Traditional Approaches** | **Proposed Approach** |
| --- | --- |
| Traditional methods are manual and time consuming | Machine learning-based methods are faster and more automated |
| Not possible to collect large amounts of data over extended periods of time | Large amounts of data can be gathered and processed over extended periods of time. |
| Involved surveys, and statistics and could be biased | Make use of remote sensed data |
| Practically difficult to collect accurate data through field observations | Can provide real-time accurate predictions |

**Table 1.1:** Existing approaches Vs Proposed approach

**CHAPTER-2**

**LITERATURE SURVEY**

**2.1 SUMMARY OF PAPERS STUDIED**

**Paper- 1**

| Title | Ensemble Machine Learning Methods to Estimate the Sugarcane Yield Based on Remote Sensing Information |
| --- | --- |
| Author(s) | Sandeep Kumar Singla\* , Rahul Dev Garg, Om Prakash Dubey |
| Year | 2020 |
| Publisher | IIETA |
| Summary | The proposed model based on the random forest algorithm is best among all the scenarios and growth stages, whereas the model based on classification and regression tree performs worst in all the cases. |

**Paper- 2**

| Title | Sugarcane Yield Forecast in Ivory Coast (West Africa) Based on Weather and Vegetation Index Data |
| --- | --- |
| Author(s) | Edouard Pignède , Philippe Roudier, Arona Diedhiou, Vami Hermann N’Guessan Bi, Arsène T. Kobea, Daouda Konaté and Crépin Bi Péné |
| Year | 2021 |
| Publisher | MDPI |
| Summary | The resulting forecasting model allowed the correct categorization of 60% of the yields. This model can detect high or low yields in 50% and 69% of cases, respectively, while being wrong in only 10% and 2% of cases, respectively. |

**Paper- 3**

| Title | Evaluation of Remote Sensing and Meteorological parameters for Yield Prediction of Sugarcane (Saccharum officinarum L.) Crop |
| --- | --- |
| Author(s) | Preeti, S; et al. |
| Year | 2022 |
| Publisher | BABT |
| Summary | LTP\_MLR technique was used to build the empirical model for predicting the yield using weather variables and remote sensing derived VCI.  Accuracy of the proposed methodology achieved R^2 value in a range of 0.95 and 0.94 in most of the districts of Maharashtra state with a standard error of 1.04% followed by the Uttar Pradesh state with 5.36% |

**Paper- 4**

| Title | Detecting Sugarcane Crop Yield using Decision Tree Classifier in the District of Muzaffarnagar |
| --- | --- |
| Author(s) | Ankit Kumar and Anil Kumar Kapil |
| Year | 2021 |
| Publisher | ijemr |
| Summary | We are also analyzing the curse of dimension problem of ML algorithms by using dimension reduction techniques like Lasso Regression ML algorithm. |

**Paper- 5**

| Title | Sugarcane Yield Mapping Using High-Resolution Imagery Data and Machine Learning Technique |
| --- | --- |
| Author(s) | Tatiana Fernanda Canata, Marcelo Chan Fu Wei, Leonardo Felipe Maldaner and José Paulo Molin |
| Year | 2021 |
| Publisher | MDPI |
| Summary | Integrating multi-temporal imagery data from Sentinel-2 and the RF regression method enabled the development of predictive yield models for commercial sugarcane fields. In addition, the RF regression showed greater accuracy (lowest RMSE and higher R2) when compared with the MLR. |

**Paper- 6**

| Title | Remote Sensing-Based Yield Forecasting for Sugarcane (Saccharum officinarum L.) Crop in India |
| --- | --- |
| Author(s) | S. K. Dubey • A. S. Gavli • S. K. Yadav • Seema Sehgal • S. S. Ray |
| Year | 2020 |
| Publisher | ISRS |
| Summary | The various combinations of attributes using correlation matrix and feature selection techniques using the different NDVI data (Remote sensing data). |

**Paper- 7**

| Title | Sugarcane Yield Estimation through Remote Sensing Time Series and Phenology Metrics |
| --- | --- |
| Author(s) | Dimo Dimov , Johannes H. Uhl , Fabian Löw, Gezahagn Negash Seboka |
| Year | 2021 |
| Publisher | Springer |
| Summary | This study was carried out to explore the suitability of satellite remote sensing-based index for operational district-level sugarcane yield forecasting. The Vegetation Condition Index (VCI) derived from long-term low-resolution satellite data was found to explain the sugarcane yield variability up to 86%, in some cases. |

**Paper- 8**

| Title | Sugarcane Crop Yield Forecasting Model Using Supervised Machine Learning |
| --- | --- |
| Author(s) | Vijay Rajpurohit, Ramesh Medar |
| Year | 2019 |
| Publisher | Research Gate |
| Summary | In this paper, a novel approach to sugarcane yield forecasting in the Karnataka(India) region using Long Term-Time-Series (LTTS), Weather-and-soil attributes, Normalized Vegetation Index(NDVI) and Supervised machine learning(SML) algorithms have been proposed. Sugarcane Cultivation Life Cycle (SCLC) in Karnataka(India) region is about 12 months, with plantations beginning at three different seasons. Our approach divides yield forecasting into three stages. |

**2.2 INTEGRATED SUMMARY OF THE LITERATURE STUDIED**

Table 2 lists the advantages as well drawbacks of all the algorithms studied in the literature review.

| **Serial** **No.** | **Algorithms** | **Advantages** | **Drawbacks** |
| --- | --- | --- | --- |
| **1** | MLR | MLR allows for the analysis of more complex relationships between the dependent variable and multiple independent variables. This is particularly useful when the relationship is not adequately captured by a simple linear model. | While MLR can model complex relationships, too much complexity without adequate data can lead to overfitting or difficulties in interpreting the model. |
| **2** | Random Forest | It can easily handle noisy data.Highest accuracy and efficiency in predicting crop yields. | Random Forests can be prone to overfitting, especially if the model is not properly tuned. Overfitting occurs when the model learns the training data too well, capturing noise and outliers rather than the underlying patterns. |
| **3** | LSTM | It provides high accuracy for the task at hand. Highly efficient for dealing with time series data. | Data might not always be available in a favorable format. |
| **4** | SVM | SVMs are less prone to overfitting, especially in high-dimensional spaces.  This makes them suitable for scenarios with complex data. | SVMs may not scale well to very large datasets. The training time and memory requirements can become prohibitive as the dataset size increases. |
| **5** | CNN | CNNs naturally learn hierarchical representations of features, from low-level features like edges to high-level features like complex patterns. | CNNs are designed for spatial information and might not inherently capture temporal dynamics. |

**Table 2.1:** Comparison of Algorithms

Table 3 lists the advantages as well drawbacks of various solution approaches that were observed in the literature survey.

| **Serial** **No.** | **Approach** | **Advantages** | **Drawbacks** |
| --- | --- | --- | --- |
| **1** | Random forest, SVR, CART, KNN, NDVI, MDA, MDG | Sugarcane yield estimation model based on the temporal profile of spectral information of Landsat-8 has been explored. | Random forest algorithm was best among all the scenarios and growth stages, whereas model based on classification and regression tree performed worst in all the cases. |
| **2** | Random forest (RF) and linear regression (LR) | Takes into account cultural practices in the model & improves the score and enables forecasting 3 months before harvest in 50% and 69% of cases whether yields will be high or low, respectively, with errors of only 10% and 2%, respectively. | At the plot level, the noise due to cultivation practices hides the effects of climate on yields and leads to poor forecasting performance. |
| **3** | MLR, NDVI, VCI | Impact of remote sensing-based derived products with Climate data on the accuracy of a prediction model for the sugarcane yield. | In some districts, the coefficient of determination was found to be low. |
| **4** | Decision Tree Classifier | NDVI images from the SPOT sensor used in the sensor's determination with the ECMWF model. | The classifier accurately predicted 70% of the time, with error rate 40%. |
| **5** | RF (Random forest) and MLR (Multiple Linear Regression) | Developed predictive sugarcane yield models integrating time-series orbital imaging and a machine learning technique. | Linear correlation between VIs and sugarcane yield did not present values greater than 0.50. |

**Table 2.2:** Comparison of Solution approaches.

**CHAPTER-3**

**REQUIREMENT ANALYSIS AND SOLUTION APPROACH**

**3.1 OVERALL DESCRIPTION OF THE PROJECT**

Sugarcane is one of the most important cash crops in India and provides livelihood to a major part of the population. India is the second largest producer and exporter of sugarcane in the world and its contribution in the economy cannot be denied.

Even in the 21st century the methods used by farmers in India are traditional and the use of technology is almost non-existent. If the majority of the agriculture sector can be penetrated with state-of-the-art technological advancements it can prove to be extremely beneficial for the farmers, economy as well as the environment because it will promote more efficient and sustainable agriculture while at the same time increasing profits for the farmers.

Uttar Pradesh state is the largest sugarcane producing state of the country with majority of the yield coming from the Muzaffarnagar, Saharanpur, Meerut, Baghpat, Bulandshahr, Ghaziabad, Amroha, Moradabad, Bijnor, Rampur, and Hapur districts. These districts comprise what is known as the Sugarcane Belt of Uttar Pradesh.

In this project we used GEE API in order to access Sentinel-2 satellite imagery of these districts. Using longitude and latitude coordinates of these districts a polygon was formed. Satellite Imagery of this polygon over 8 years starting from 1.1.2015 to 1.1.2023 comprised the core dataset for this project. The collection of images was then used to compute NDVI values with timestamps. Also meteorological data gathered from the ERA-5 Land dataset provided by ECMWF is also used by integrating it with the NDVI dataset to get more accurate results.

This data was then used to train and test our LSTM, MLR and RF models which were able to accurately predict NDVI and yield values of the future. These predicted NDVI values were then compared to the actual NDVI values, and the results were found to be reasonably accurate.

Further, various regression models were used to calculate the average yearly yield of the Sugarcane belt districts. MLR regression model was able to calculate the yield with reasonable accuracy of 98% when compared to the actual data.

Then the MLR model was saved and deployed using flask to develop a user interface with input parameters like soil temperature, evaporation, precipitation, air temperature etc to predict sugarcane yield on daily basis.

**3.2 REQUIREMENT ANALYSIS**

This Section lays down the specific software, hardware requirements as well as any other tools which were necessary in the development and completion of this project and act as dependencies.

**3.2.1 HARDWARE REQUIREMENTS**

This project is not computationally intensive and does not require extraordinary hardware components. However, during certain steps of the project having more computational power can prove to be useful.

During the extraction phase of Sentinel-2 Imagery and ERA-5 Land dataset as well as computing NDVI from these imagery a good bit of computer memory is used as the dataset is very large.

Also during the training and testing phase of the LSTM model decent computational power is required.

| **Particulars** | **Specifications** |
| --- | --- |
| CPU | 500 Mhz (Minimum) |
| Computer Processor | Intel i5 or i3 |
| Computer Memory | 4Gb |
| Graphics Hardware | Not Required |
| Network | Internet Connection Required |

**Table 3.1:** Hardware Requirements

**3.2.2 SOFTWARE REQUIREMENTS**

The Entire project was developed using Python3 as the coding language with Jupyter Notebook as the IDE of choice. Python was chosen due its easy interoperability and integrability with the Google Earth Engine API.

GEE API was essential as it allowed us to access Sentinel-2 Imagery and ERA-5 Land dataset provided by ECMWF as well as specify our AOI i.e. the Sugarcane Belt of Uttar Pradesh. GEE API played an important role in the availability of the data. This section lays down the softwares and tools utilized in the development of this project and are summarized in the following table. Along with the basic software requirements, numerous python libraries were used in the completion of this project. A list of all the libraries have been summarized in table 4. These python libraries enabled the implementation of the project.

| **Particulars** | **Specifications** |
| --- | --- |
| Operating System | Linux, Windows or Mac OS |
| Coding Language | Python |
| Coding IDE | Jupyter Notebook / Google Colab |
| APIs Required | Google’s Earth Engine API |

**Table 3.2:** Software Requirements

| **Name of the library** | **Usage** |
| --- | --- |
| Ee | For working with the Google Earth Engine Platform, in order to extract satellite data. Also provides necessary tools for analysis. |
| Pandas | To analyze and manipulate data and create a dataset from the extracted imagery |
| Folium | To visualize geo-spatial data i.e., the data extracted from Sentinel-2. Also to create interactive maps and display NDVI. |
| IPython.display | To display interactive images, videos, etc. In jupyter notebook |
| Matplotlib | To visualize the data in the form of graphs, plots, etc. |
| Datetime | To work with date, timestamps and other date formats. |
| Pylab | For data visualization and performing calculations. |
| Seaborn | For creating line charts and displaying the final results. |
| Keras | For building and training LSTM model |
| Sklearn | For data pre-processing, classification and implementation of regression model |
| Shapely | For manipulating and analyzing planar objects like polygons which are visualized using longitude and latitude coordinates. |
| Pyproj | Used for conversion of coordinates in different referencing systems, |
| Numpy | For working with arrays and other mathematical functions. |

**Table 3.3:** Python Libraries Used

**3.3 SOLUTION APPROACH**

The project can be divided into 4 steps:

1. Visualising Sentinel-2 Imagery

2. Visualizing ERA-5 Land dataset provided by ECMWF

3. Creation of Dataset combining both the data

4. Defining and Implementing LSTM, MLR and RF Model

5. Save the MLR model

6. Deploy using flask

5. Estimate Crop Yield.

These steps will be discussed in detail in the following subsections.

**3.3.1 VISUALISING SENTINEL-2 IMAGERY**

In order to access and visualize Sentinel-2 Imagery using the GEE API, we first need to define our area of interest. For this project, our area of interest is the Sugarcane belt of the Uttar Pradesh state. Uttar Pradesh is the largest Sugarcane producing state in India, 95% of this sugarcane is cultivated in the Sugarcane Belt of Uttar Pradesh which makes this the perfect choice for our area of interest for this project. This belt consists of many districts, including Muzaffarnagar, Saharanpur, Meerut, Baghpat, Bulandshahr, Ghaziabad, Amroha, Moradabad, Bijnor, Rampur, etc. The entire list of districts that lay in the Sugarcane Belt has been listed in Table 5.

| Bijnor | Bulandshahr | Etah | Saharanpur |
| --- | --- | --- | --- |
| Muzaffarnagar | Meerut | Moradabad | Rampur |
| Amroha | Bareilly | Pilibhit | Shahjahanpur |
| Ghaziabad | Gautum Budh Nagar | Baghpat | Hapur |
| Lucknow | Barabanki | Hardoi | Lakhimpur Kheri |
| Sitapur | Gorakhpur | Kushinagar | Deoria |
| Basti | Sant Kabir Singh Nagar | Siddharthnagar | Azamgarh |

**Table 3.4:** Sugarcane Belt District

However, these districts are sparse apart and all cannot be included in our AOI. So the central and highest producing part of this Sugarcane Belt is incorporated in our AOI.

In order to visualize this Sugarcane belt, we constructed a polygon using Google Earth Engine which covered the majority of the area. We selected vertices of this polygon in the form of Longitude and Latitude coordinates. The polygon consisted of 17 vertices and covered an area of roughly 206066.10 hectares.

The coordinates of these vertices in the form of Longitude and latitude points have been listed Table 6 below.

| **Latitude** | **Longitude** | **Name** |
| --- | --- | --- |
| 28.1124 | 80.1728 | Point 1 |
| 28.0459 | 80.3233 | Point 2 |
| 27.9413 | 80.3648 | Point 3 |
| 27.8406 | 80.3476 | Point 4 |
| 27.8007 | 80.1756 | Point 5 |
| 27.8132 | 80.0897 | Point 6 |
| 27.8557 | 80.0149 | Point 7 |
| 27.8978 | 79.9669 | Point 8 |
| 27.9468 | 79.9190 | Point 9 |
| 27.9993 | 79.8889 | Point 10 |
| 28.0641 | 79.8704 | Point 11 |
| 28.1435 | 79.8015 | Point 12 |
| 28.2381 | 79.7522 | Point 13 |
| 28.2827 | 79.7847 | Point 14 |
| 28.3554 | 79.7744 | Point 15 |
| 28.3964 | 79.8495 | Point 16 |
| 28.4119 | 79.8778 | Point 17 |

**Table 3.5:** Polygon Vertices

These vertices were constructed using our polygon which was visualized on a map as an overlay.

It can be observed that it covers the most sugarcane producing areas as seen in figure 1.



**Figure 3.1:** AOI Polygon map.

With our AOI clearly defined we can now proceed to the next step, which is to visualize Sentinel-2 Imagery for this specific AOI.

This is made possible via the Copernicus Program. Copernicus is a European Union program for monitoring and analyzing Earth and its Environment. Under this program numerous satellites, ground stations and other infrastructure has been established. This program makes available its observations and the vast amount of data collected freely and openly through a dedicated Copernicus Open Access Hub which also includes the data collected through the Sentinel missions.

The GEE platform has integrated Sentinel-2 data from the Copernicus program and made it accessible to researchers and developers through its API. The API allows free access to all the Sentinel-2 Imagery and data.

Using the GEE API, we have accessed the Sentinel-2 Imagery over our specified AOI and visualized it in the form of an interactive widget. Snapshot of the interactive widget can be seen in Figure 2.

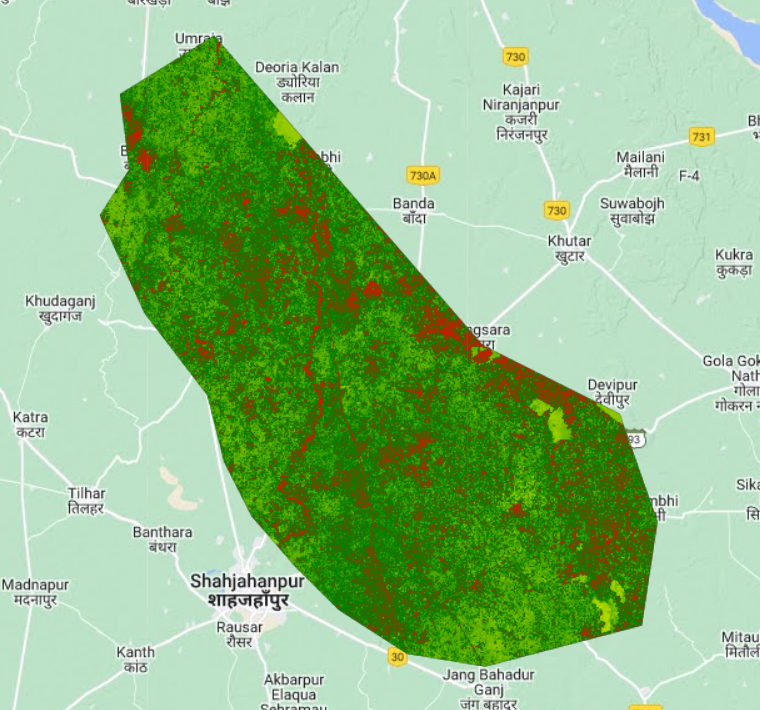


**Figure 3.2:** AOI Sentinel-2 Image.

The next step is to gather and visualize NDVI. It is a remote sensing index which is used to estimate the health and amount of vegetation in a particular area. The values for NDVI range from –1 to +1. Higher NDVI generally means denser and healthier vegetation. Water and soil have negative NDVI values while vegetation generally has positive NDVI values.

NDVI can be calculated using spectral bands captured by the Sentinel-2 satellite.

A snapshot of the visualized NDVI can be seen in Figure 3.



**Figure 3.3:** AOI NDVI Image

**3.3.2 VISUALISING ERA-5 DATASET**

With our AOI being clearly defined, we visualized the ERA-5 dataset provided by European Centre for Medium-Range Weather Forecasts (ECMWF) by first converting all the latitudes and longitudes of our polygon area in a shapefile. Shapefiles are effective for visualizing spatial data. Geographic Information System(GIS) software can display shapefiles on maps, allowing users to easily interpret the geographic information. This is particularly useful when dealing with complex geometries such as polygons. After importing the shapefile we selected our required ERA-5 Land Surface Daily Temperature dataset and visualized different weather parameters like soil temperature, precipitation, evaporation etc and finally exported to a csv file. Code for the same was done in google earth engine code editor as shown in Figure 0.

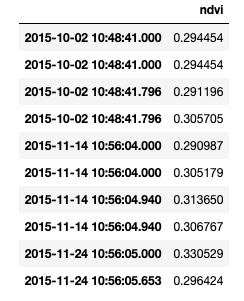
**3.3.3 CREATION OF DATASET**

To create our dataset, we collected NDVI and meteorological data of our AOI over a period of 8 years starting from 1.1.2015 to 1.1.2023 and integrated both the csv files on the basis of time series parameters.

We extracted NDVI for these dates and created a collection of images. However, these images cannot be directly used to train our LSTM model and thus data pre-processing was required.

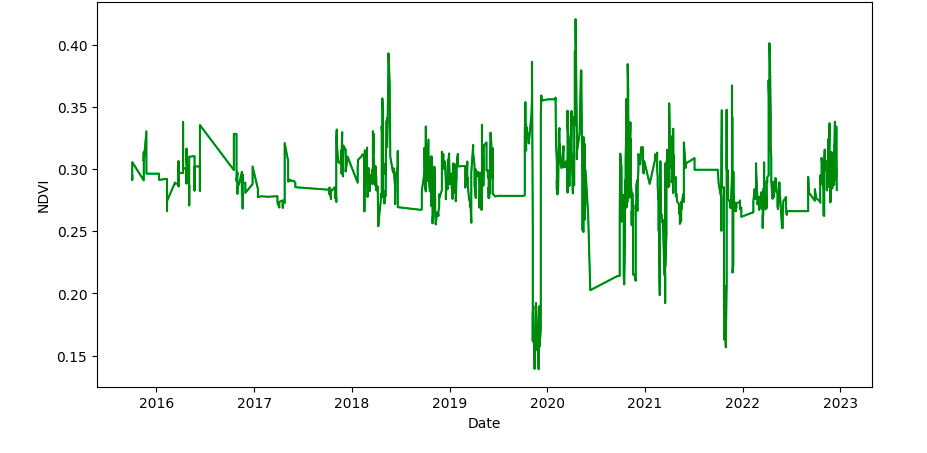
The image collection was transformed in the form of a python dictionary with two keys- namely NDVI and timestamp.

Each image in the collection was reduced to represent a mean NDVI and was stored in a dictionary along with its timestamp. As represented in Figure 4.



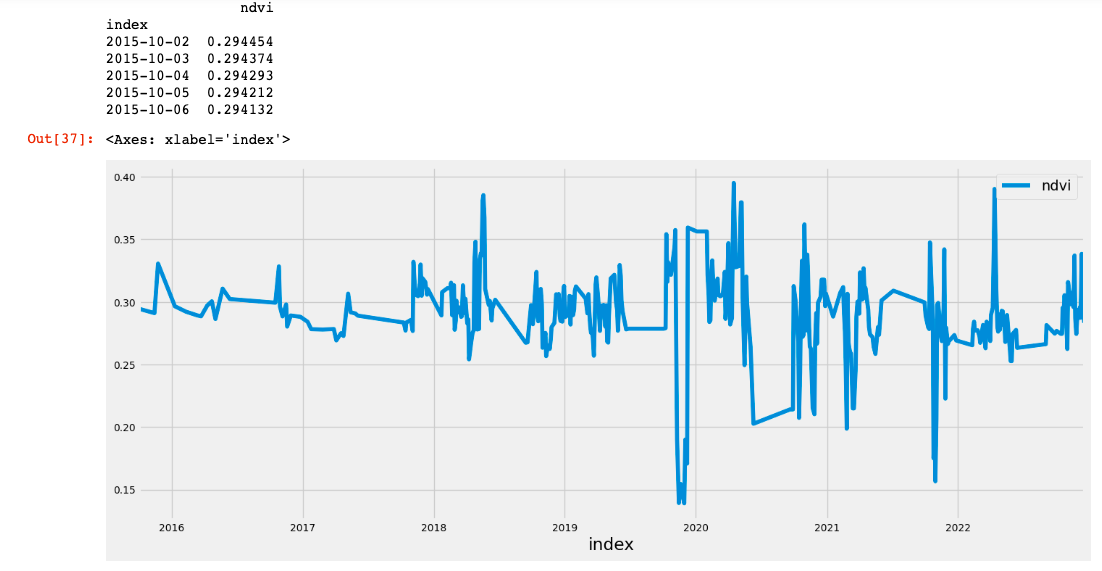
**Figure 3.4:** NDVI with Timestamps

This data was then plotted in the form of a graph.



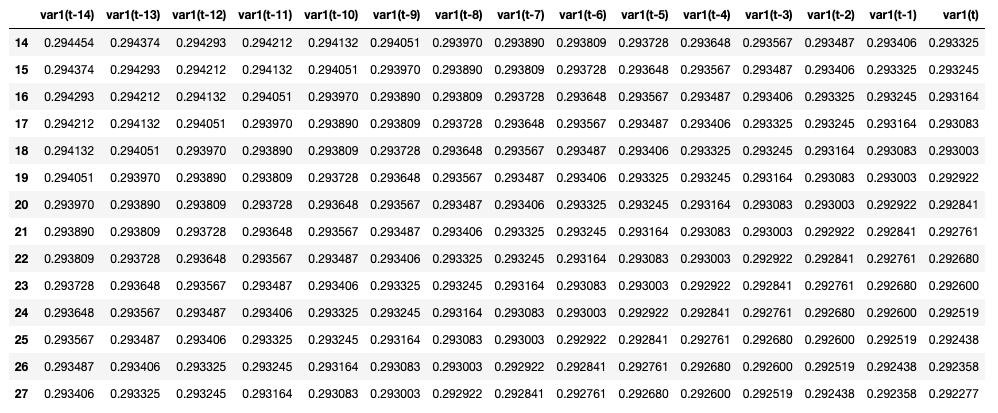
**Figure 3.5:** NDVI vs Date Graph

However, this data was raw and required pre-processing. Duplicates were removed from the data and only one reading per day was kept in the dataset. Finally, interpolation was done to ensure that readings for all days were available.



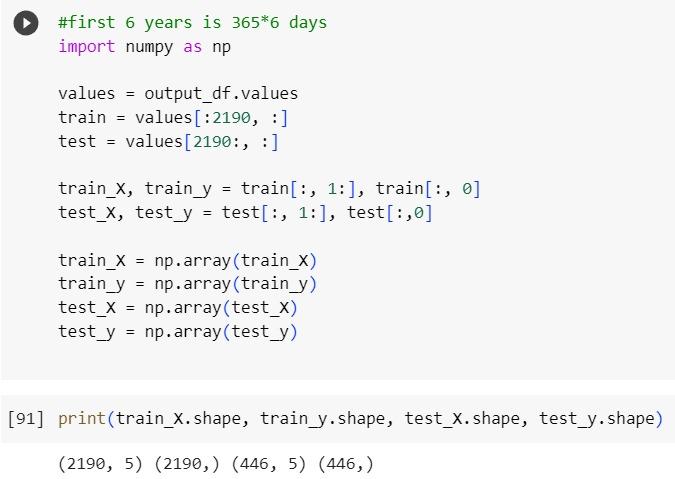
**Figure 3.6:** Interpolated NDVI

After the preprocessing the data was then converted to a time series data and then supervised data which is a favorable format for training and testing the LSTM model. As can be observed in figure 7.



**Figure 3.7:** Supervised data using series\_to\_supervised function

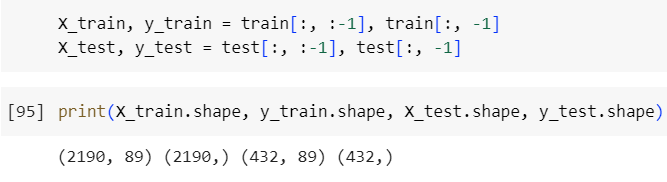
**Splitting of dataset into train and test**

****

**Figure 3.8:** Splitting time series data for RF and MLR model

**For LSTM model**

Converted time series data to supervised data

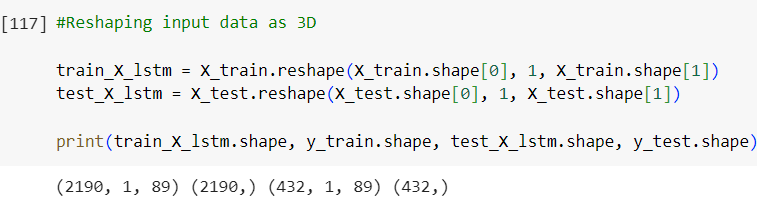


**Figure 3.9:** Using supervised data for LSTM model

**3.3.4 DEFINING AND IMPLEMENTING MACHINE LEARNING MODELS**

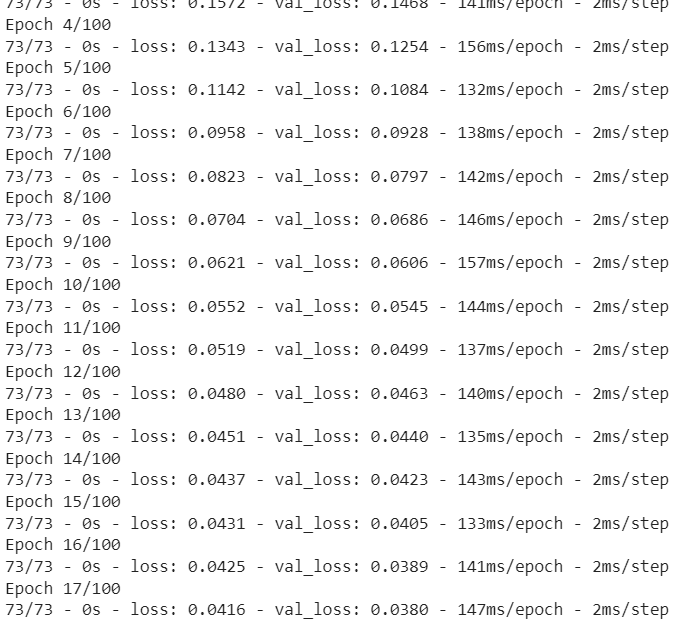
In order to implement the LSTM, MLR and RF models the dataset was split into training and testing sets. Where the first 6 years of data was used for training and the rest was used for testing as well as predicting the models’ performances.

We divided the dataset into the training and testing phase, we used the first 6 years of data i.e., starting from 1.1.2015. This corresponded to 2190 days of data or training and 432 days of data for testing. The final shape of the dataset in the favorable format can be seen in Figure 10.



**Figure 3.10:** Favorable data format for LSTM

After reshaping the dataset, the models are defined, and the data is used to train the models. The model attempts to minimize the difference between the predicted values and the actual values as it learns to predict using the training data. Up to the time where the model's performance on the validation data begins to degrade, the training procedure is repeated for a predetermined number of epochs. At that point, the training procedure is terminated.



**Figure 3.11:** Sample EPOCHs

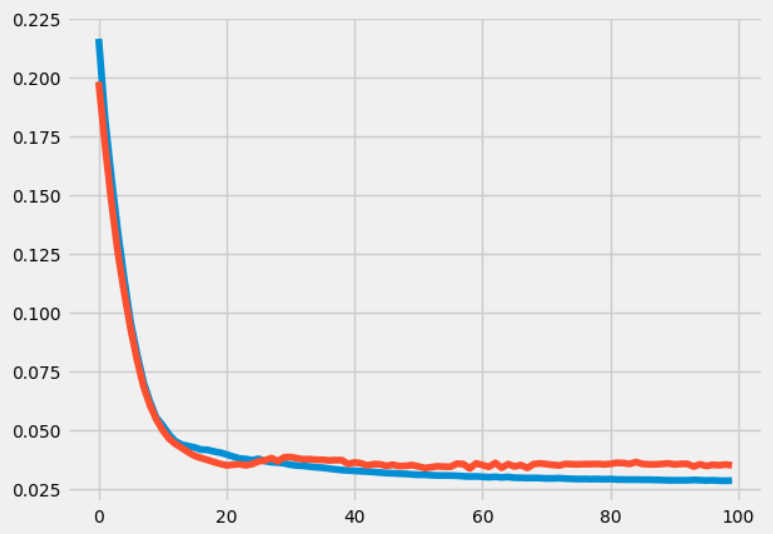
Two frequent issues with machine learning algorithms are overfitting and underfitting.

When a model tries to match the training data too closely and is too complicated, overfitting occurs. Due to this, the model has poor generalization to new information and shows poor results on test data. In other words, the model is unable to apply to new data since it has absorbed the training data too thoroughly.

When a model is too straightforward and fails to recognise the deeper patterns in the data, under fitting occurs. Additionally, this leads to subpar performance on both the test and training sets of data.

To visualize underfitting and overfitting, we can use a plot of loss values over each EPOCH.

As can be observed in figure 10, overfitting and underfitting issues were mitigated to the best extent possible, and the model was trained and tested reasonably well.



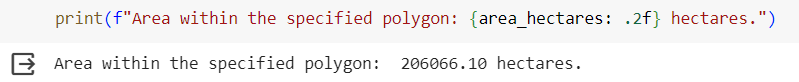
**Figure 3.12:** Loss in EPOCHs

**3.3.5 ESTIMATING CROP YIELD**

The range of the NDVI is from -1 to 1, with values nearer 1 indicating more vegetation. More chlorophyll in the plants, which implies healthy and dense vegetation, translates into a higher NDVI rating. A lower NDVI score, on the other hand, denotes less vegetation, which may be brought on by elements like drought or pests.

We used our predicted NDVI values to estimate the crop yield and compared it to the crop yield calculated using the actual NDVI data.

To calculate the crop yield, we first had to calculate the area in hectares of our AOI. The area was calculated to be 206066.10 hectares.



**Figure 3.13:** Computed area of AOI

To calculate the crop yield in kg-hectares, we used a regression model. Based on the correlation between NDVI and crop yield, the regression model takes the NDVI data as input and produces a predicted crop yield value. The model can forecast crop production from new NDVI data by examining the link between NDVI and crop yield.

Regression models are commonly used in crop yield predictions as they allow for the analysis of the relationship between variables such as NDVI values, weather parameters, and soil characteristics, and the prediction of crop yield. In this context, regression models can be used to build a mathematical equation that can predict crop yield based on the input variables.

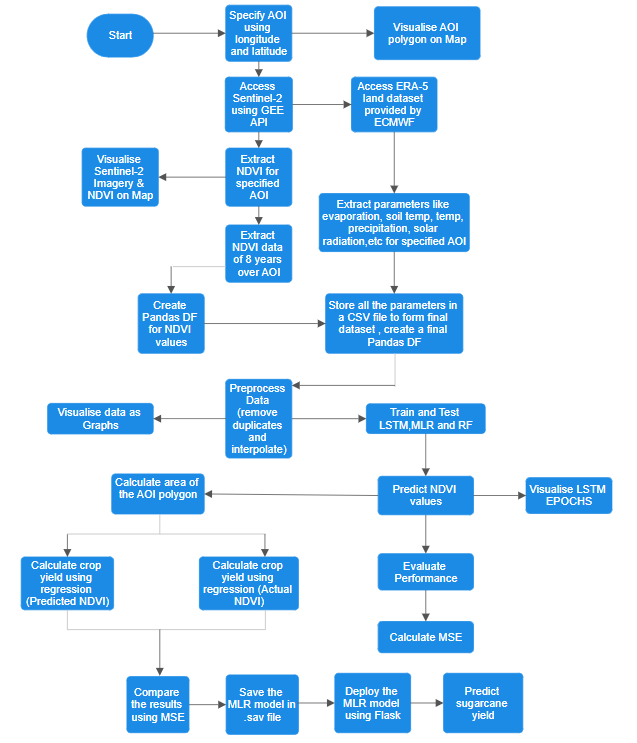
Finally using this model, the yield estimates in kg-hectares were calculated and compared to actual yield.

**CHAPTER-4**

**MODELING AND IMPLEMENTATION DETAILS**

**4.1 DESIGN DIAGRAMS**

**4.1.1 CONTROL FLOW DIAGRAM**

****

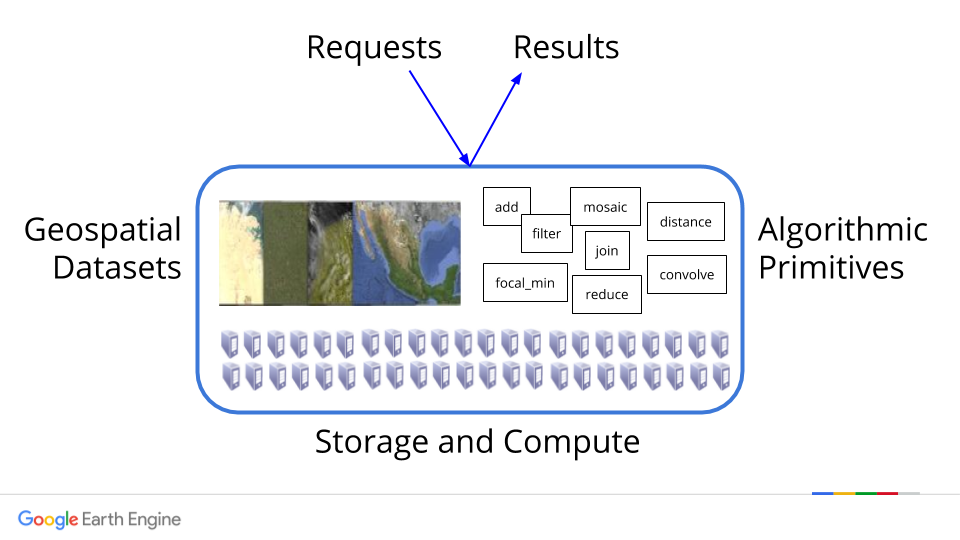
**Figure 4.1:** Control Flow Diagram

**4.2 MODELS AND FRAMEWORKS USED**

* **Google’s Earth Engine**

A cloud-based technology called Google Earth Engine enables users to access and examine geographical data. It offers petabytes of remote sensing data and permits massive data processing utilizing Google's infrastructure.

Google Earth Engine may be used to obtain and analyze satellite imaging data from sensors like Landsat and Sentinel-2 for agricultural yield prediction projects. Input characteristics for crop yield prediction models include the NDVI, temperature, humidity, precipitation, etc., which the platform can extract.



**Figure 4.2:** Google Earth Engine

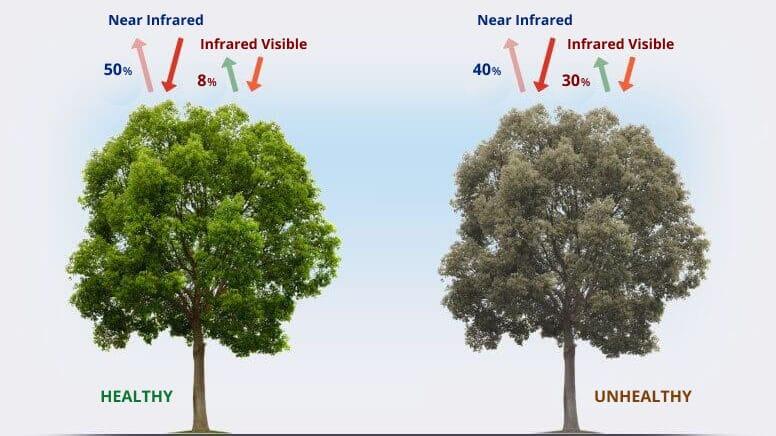
Researchers may quickly obtain long-term, multi-spectral satellite data using Earth Engine, which offers regular, reliable observations that allow investigation of agricultural conditions throughout the growing season. With the use of this information, patterns, trends, and crop production predictions may be made. To assess crop stress, water availability, and other variables that affect crop development, the platform also provides a variety of analysis methods, such as time series analysis.

Utilizing Google Earth Engine for crop yield prediction projects has several advantages, one of which is that it makes it possible to analyze huge, complex data sets without the need for sophisticated computing infrastructure. The platform's user interface makes it simple for researchers to access and analyze the data, saving time and money on pre-treatment and analysis of the data.

* **NDVI (Normalised Difference Vegetation Index)**

A popular statistic for determining vegetation growth and health is the Normalised Difference Vegetation Index (NDVI). It makes advantage of the red and near-infrared light spectrums that plants absorb and reflect, respectively. The near-infrared and red reflectance values are divided by their total to determine the NDVI, which produces a number between -1 and 1. Negative NDVI denotes the absence of any vegetation, 0 barren soil, and 1, thick green vegetation.

We may utilize bands 4 (red) and 8 (near-infrared) of the satellite pictures to calculate NDVI using Sentinel-2 data. The satellite records these bands, which are then processed to produce reflectance values for each pixel. The NDVI values for each pixel are then determined using these reflectance values.



**Figure 4.3:** NDVI Working

By monitoring the development and overall well-being of crops throughout time, NDVI may be used to forecast agricultural yields. We may determine which areas of an area of interest have good vegetation and which regions have low or no vegetation by examining the NDVI values over that area. Decisions about crop management techniques such as fertilizer application, irrigation, and other can be made using this information.

We have to first collect NDVI data over time in order to use it to estimate crop production. This may be achieved through the collection of frequent satellite images, such as every two weeks. Once we have a time series of NDVI values and different meteorological parameters, we can predict future NDVI values using machine learning algorithms like the LSTM, RF, MLR model. The crop yield may then be calculated based on these projections.

By analyzing NDVI and weather metrics over time and using machine learning algorithms to forecast future values, crop production may be predicted.

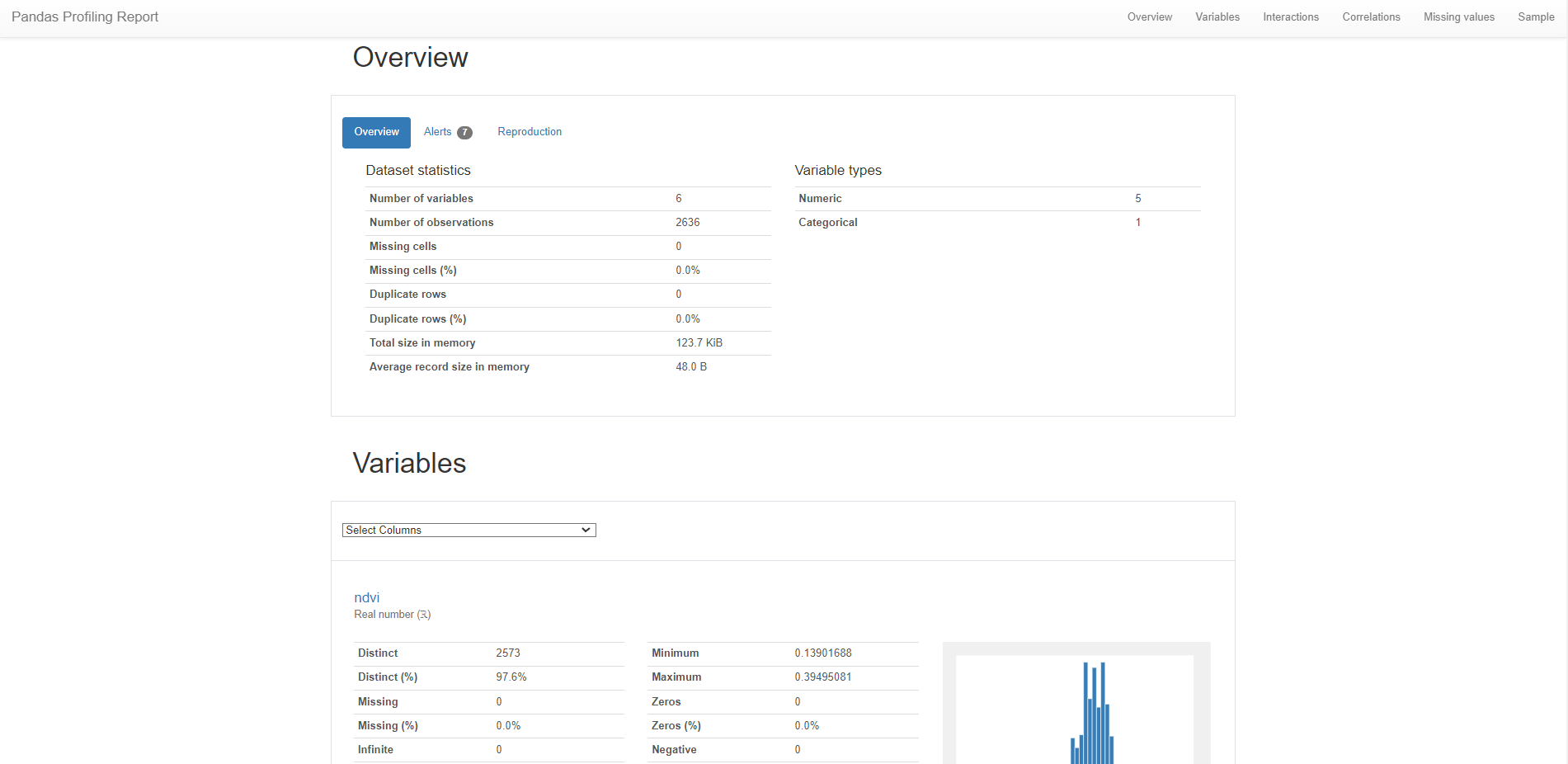
* **ERA-5 Land dataset**

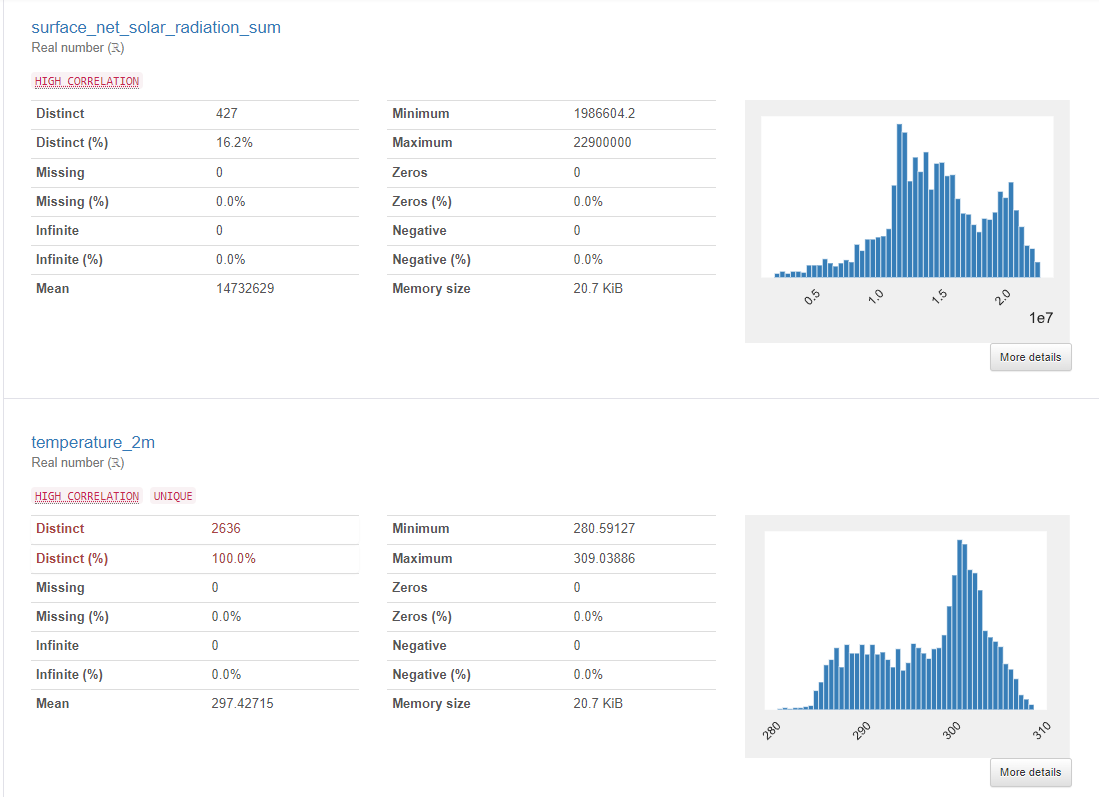
The ERA-5 Land dataset, sourced from the European Centre for Medium-Range Weather Forecasts (ECMWF), is a valuable resource for obtaining high-resolution reanalysis data encompassing various meteorological parameters. Access to this dataset should be secured with the appropriate permissions to ensure compliance with usage regulations. The dataset covers global land areas on a regular grid, making it suitable for a wide range of applications. In order to utilize this dataset for generating meteorological data specific to soil temperature, air temperature, solar radiation, precipitation, and evaporation, several key steps must be followed.

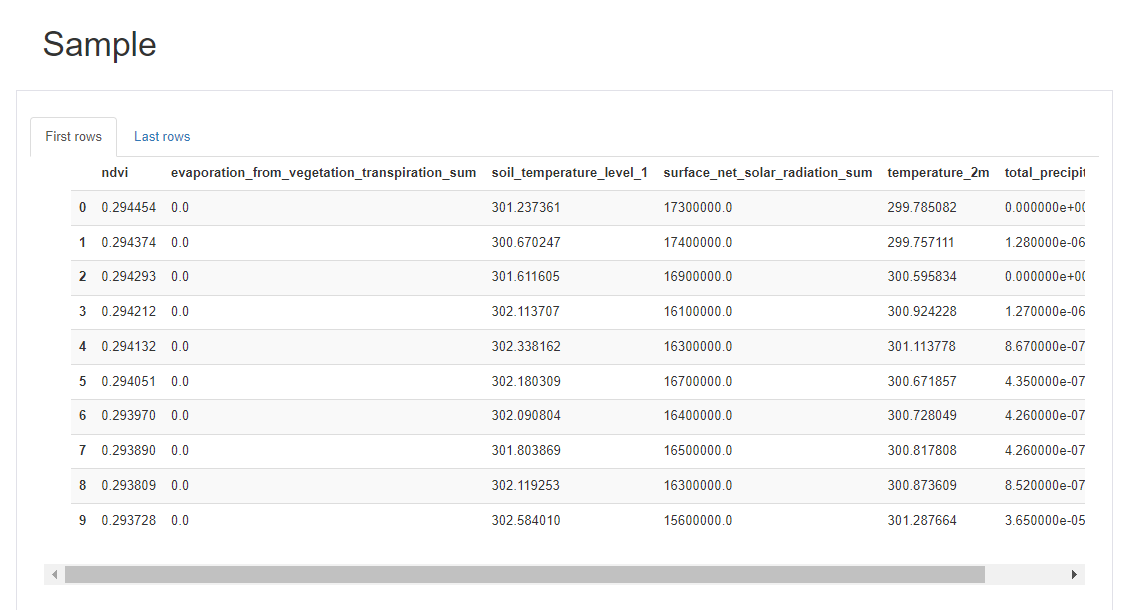
The processed meteorological data can be applied to sugarcane yield prediction models or analyses. This application may involve the use of machine learning models, statistical techniques, or crop simulation models. Validation steps, including the comparison of model outputs with independent datasets or known observations, and sensitivity analyses to understand the impact of different meteorological variables, contribute to the robustness and reliability of the sugarcane yield predictions or analyses. Throughout these steps, it is essential to refer to ECMWF documentation for dataset details and adhere to the specified terms of use to ensure responsible and compliant data usage.

* **Visualization of the final dataset**

Done using Pandas profiling Report.



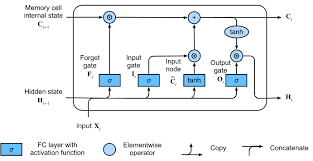




* **LSTM Model**

The LSTM model is a type of machine learning model that can be used to predict future values in a time series dataset. It stands for "Long Short-Term Memory," which refers to its ability to remember important patterns in data over long periods of time.

The LSTM model works by taking in a sequence of data points, such as daily NDVI readings, and trying to learn patterns in the data. It then uses these patterns to predict future values in the sequence.



**Figure 4.4:** LSTM Working

Two important parameters in the LSTM model are "epochs" and "batch size." Epochs refer to the number of times the entire dataset is passed through the model during training. Batch size refers to the number of data points that are used in each training iteration.

In crop yield prediction, the LSTM model can be used to analyze historical NDVI data and predict future values. By training the model on past data, it can learn to recognize patterns that may indicate changes in crop growth and yield.

The model's predictions can then be compared to actual values to evaluate its accuracy. One advantage of using the LSTM model in crop yield prediction is its ability to handle noisy and incomplete datasets.

* **MLR Model**

A MLR is a type of artificial intelligence algorithm that's used for analyzing and recognizing images. It's designed to work similarly to the way our brains process visual information. A MLR has different layers that process the image in various ways, such as by highlighting patterns like edges and corners. The final layer produces a set of scores or probabilities for different possible objects in the image. The MLR is a useful tool for analyzing images in many fields, including self-driving cars and medicine.

Overall, the Regression model is an important tool in crop yield prediction that allows farmers to make informed decisions about crop management. Its accuracy and efficiency can be improved by using appropriate techniques and metrics and by ensuring that high-quality data is used.

* **RF Model**

The Random Forest model is a machine learning algorithm that can be used for classification and regression tasks. It consists of multiple decision trees that work together to make predictions. The trees are built on random subsets of the data and features, which helps to reduce overfitting and increase accuracy. The final prediction is based on the majority vote of all the individual trees. Random Forests are easy to interpret and have many practical applications in different fields, such as finance and healthcare.

**4.3 RISK ANALYSIS AND MITIGATION**

Table 7 below gives an overview of the risks associated with the project.

| Risk\_ID | Classification | Description of Risk | Risk Area | Probability | Impact |
| --- | --- | --- | --- | --- | --- |
| Risk\_1 | Design | The possibility of low accuracy as we are not incorporating many important factors which can have a significant impact on sugarcane crop yield. | Completeness | High (H) | Moderate (M) |
| Risk\_2 | Engineering Specialties | The project scope demands maximum possible reliability on the predicted outcomes. | Reliability | Moderate (M) | Moderate (M) |
| Risk\_3 | Requirements | The NDVI and meteorological data collected may not be complete, and data interpolation was done. This can impact the overall quality of predictions that the model made. | Performance | Moderate (M) | Low (L) |
| Risk\_4 | Design | The collection of images which was extracted was then modified such that each image in the collection was reduced to represent the mean value of NDVI that was represented in the image. | Performance | Moderate (M) | Moderate (M) |

**Table 4.1:** Risk Analysis

**CHAPTER-5**

**TESTING**

Projects must include testing since it helps to assure the excellence and dependability of software. It entails the procedure of confirming and validating that a system or application complies with the given specifications and operates as anticipated. The purpose of testing is to find flaws, problems, and mistakes in the programme and make sure it runs properly.

Prior to deployment to production, it aids in finding bugs and defects in the software, lowering the possibility of later, expensive and time-consuming fixes. Second, it guarantees that the programme executes as expected, complies with the requirements, and operates appropriately. This promotes user assurance and faith in the system. Thirdly, testing helps to raise the caliber of the programme by pointing out areas that may be enhanced and optimized.

**5.1 TESTING PLAN**

In order to evaluate the efficacy of the model and approach used in the project, the performance of both the regression model used to predict the actual sugarcane yield as well as the quality and accuracy of the predictions testing was done.

In order to test the accuracy of the LSTM,MLR and RF models MSE error was calculated and was compared to aid the selection of the most optimal model.

In order to test the estimation of actual sugarcane tonnage, the yield was calculated using the actual NDVI values as well as Meteorological data (Weather) as well.

The calculated yields in both the cases were almost identical indicating that the yield prediction model performed decently.

The yields were compared to the actual data released by the government where slight error was found however, it was within the acceptable range.

**5.2 COMPONENT DECOMPOSITION AND TYPE OF TESTING REQUIRED**

The objectives behind the testing of our developed model are:

● Evaluation of Parameters of the developed system

● Calculating accuracy

● Speed of the model

● Evaluation of the predicted numbers

● Comparing the predicted values with the real world yield.

**5.3 TYPE OF TEST EXPLANATION SOFTWARE COMPONENT**

| Type of Testing | Description | Executed On |
| --- | --- | --- |
| Requirement Testing | Validation checks were made to ensure that hardware and software specifications meet the minimum requirements. | Self Check |
| Performance Testing | Performance testing is the process of determining the speed, accuracy, and consistency of the proposed model. This was achieved by checking the accuracy of the predicted price as well as the predicted trend with the actual figures. | Jupyter Notebook / Google Collab |
| Experimental Testing | Our model was checked against various trials and error tests to fine-tune the hyperparameters in order to ensure the best results. | Jupyter Notebook / Google Collab |
| Unit Testing | The purpose is to validate that each unit of the software performs as expected. The output of the steps within data preprocessing and the result of data segmentation was randomly tested in order to ensure valid and consistent results. | Jupyter Notebook / Google Collab |

**Table 5.1:** Testing

**5.4 LIMITATIONS OF THE SOLUTIONS**

Predicting sugarcane yield or any crop yield is a very difficult task as a majority of technical as well as non-technical factors influence it. Incorporating all the factors is beyond the scope of this research. Some such factors are listed below.

· The impact of meteorological conditions.

· The accessibility and quality of data are further restrictions.

· Non technical factors, such as economy, policies etc.

**CHAPTER-6**

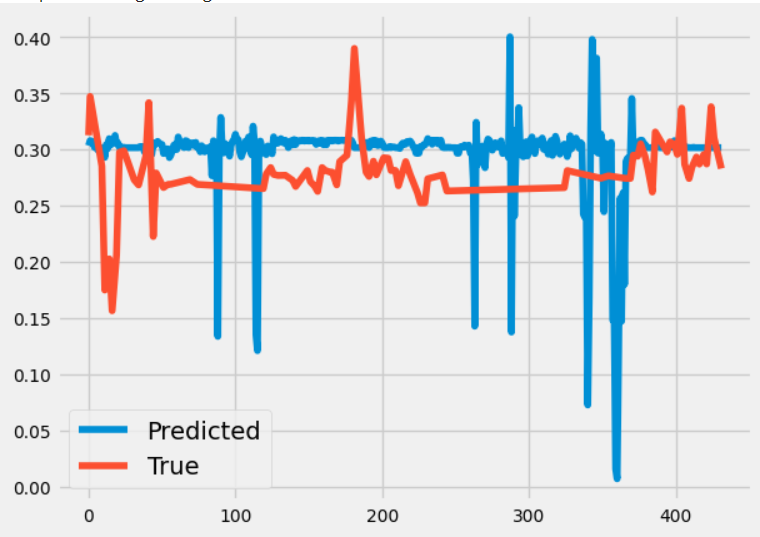
**RESULTS & CONCLUSION**

**6.1 FINDINGS**

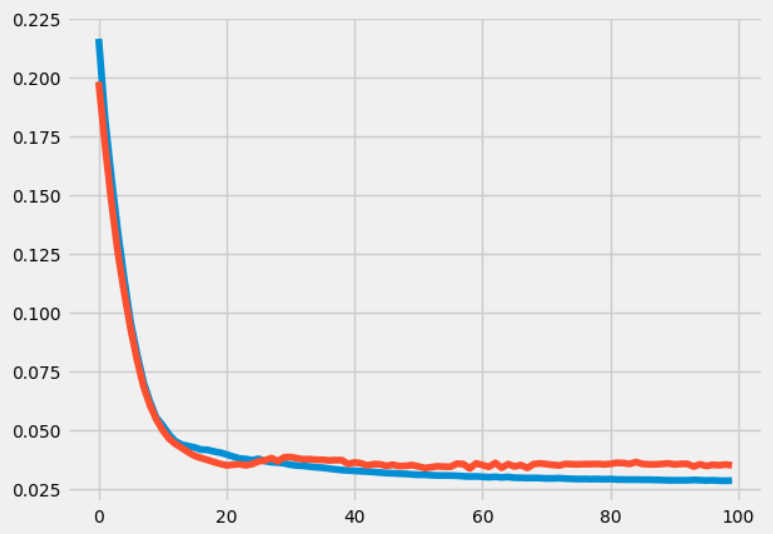
After analyzing the data and training, and testing the LSTM, MLR and RF models, we were able to accurately predict the NDVI values and subsequently calculate the average yearly crop yield in kg-hectares.

The first 6 years of data i.e., 2190 days of the data were used in the training phase and the 432 days of data were used for testing.

Figure 6.1 shows the correlation between the actual and predicted NDVI values in the form of a graph using the LSTM model. Figure 6.2 shows the validation loss in the LSTM model.

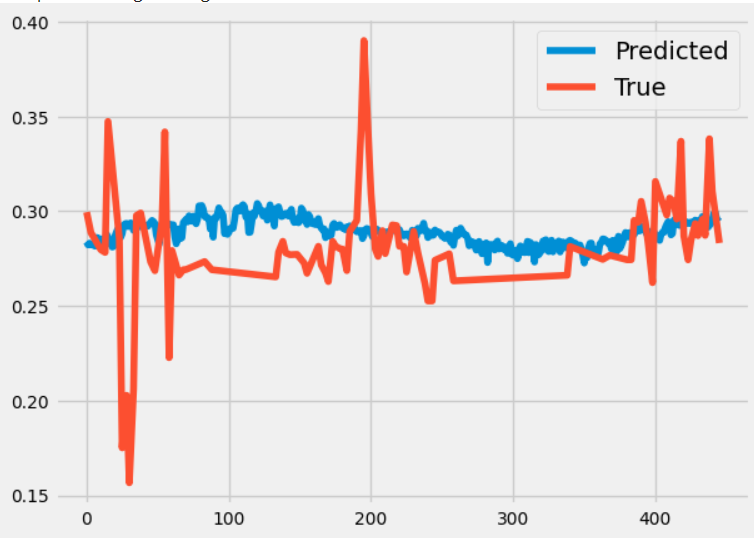


**Figure 6.1**: Predicted vs Actual NDVI using LSTM



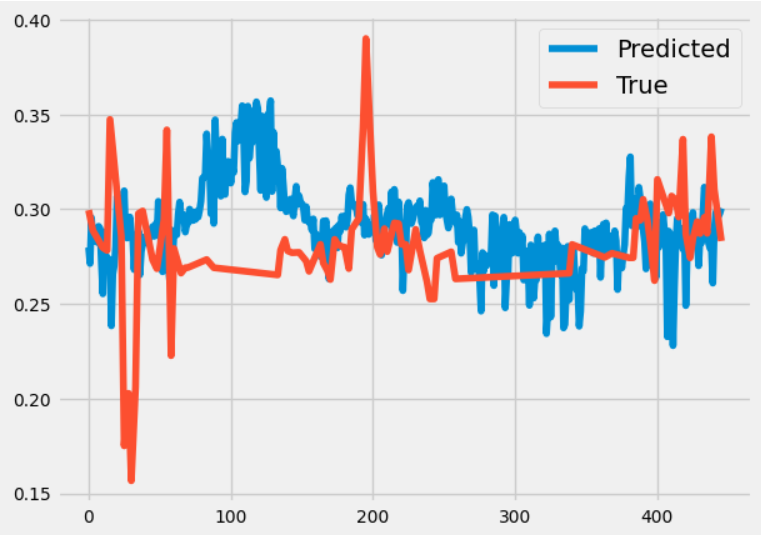
**Figure 6.2**: Validation Loss in LSTM

Figure 6.3 shows the correlation between the actual and predicted NDVI values in the form of a graph using the MLR model.



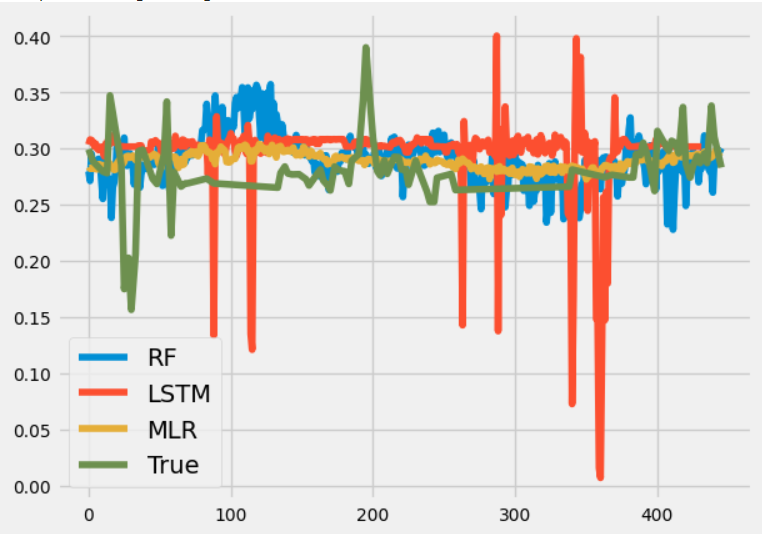
**Figure 6.3**: Predicted Vs Actual NDVI using MLR

Figure 6.5 shows the correlation between the actual and predicted NDVI values in the form of a graph using the RF model.



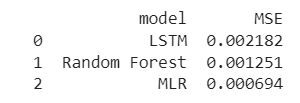
**Figure 6.5**: Predicted Vs Actual NDVI using RF

Finally Figure 6.6 shows the consolidated graph indicating the performance of the algorithms.



**Figure 6.6**: Consolidated Results

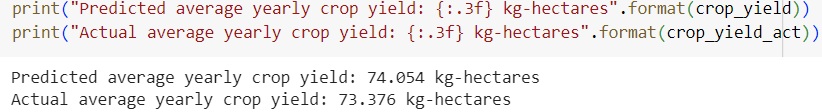
To compare the accuracy of the three models, and select the most accurate one, their MSE scores were calculated and are presented in figure 6.7.



**Figure 6.7**: MSE

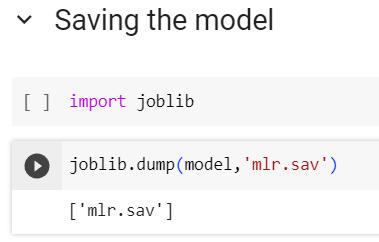
As we can see, the MLR algorithm outperforms the Random Forest as well as the LSTM model. So the Multiple Linear Regression (MLR) algorithm was chosen to predict the sugarcane yield.

Finally, the average yearly of sugarcane produced was calculated from the estimated as well as the actual NDVI values. The results of which can be seen in Figure 6.8.



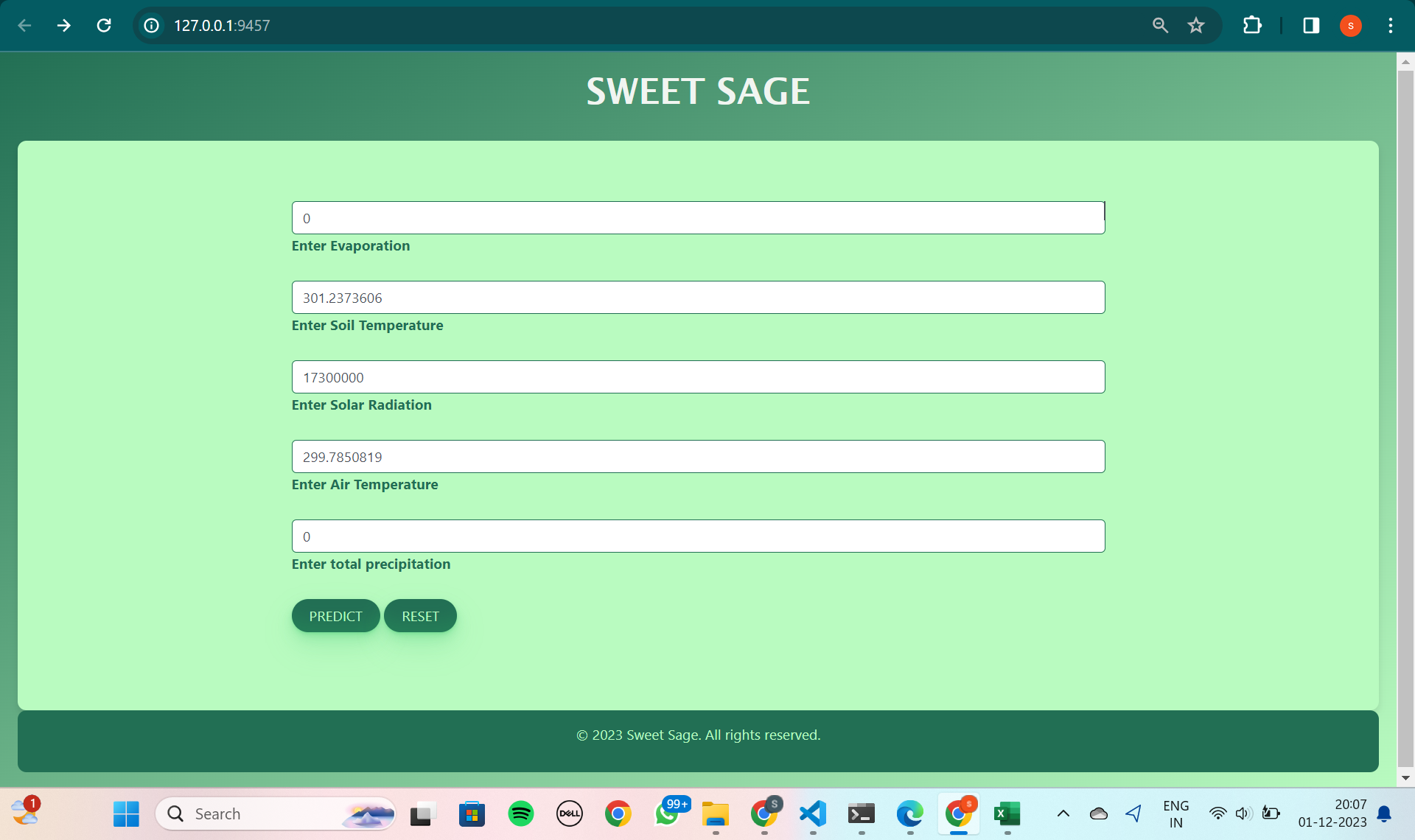
**Figure 6.8**: Predicted yields

After predicting the yield, we saved the best performer algorithm(MLR) for deployment.

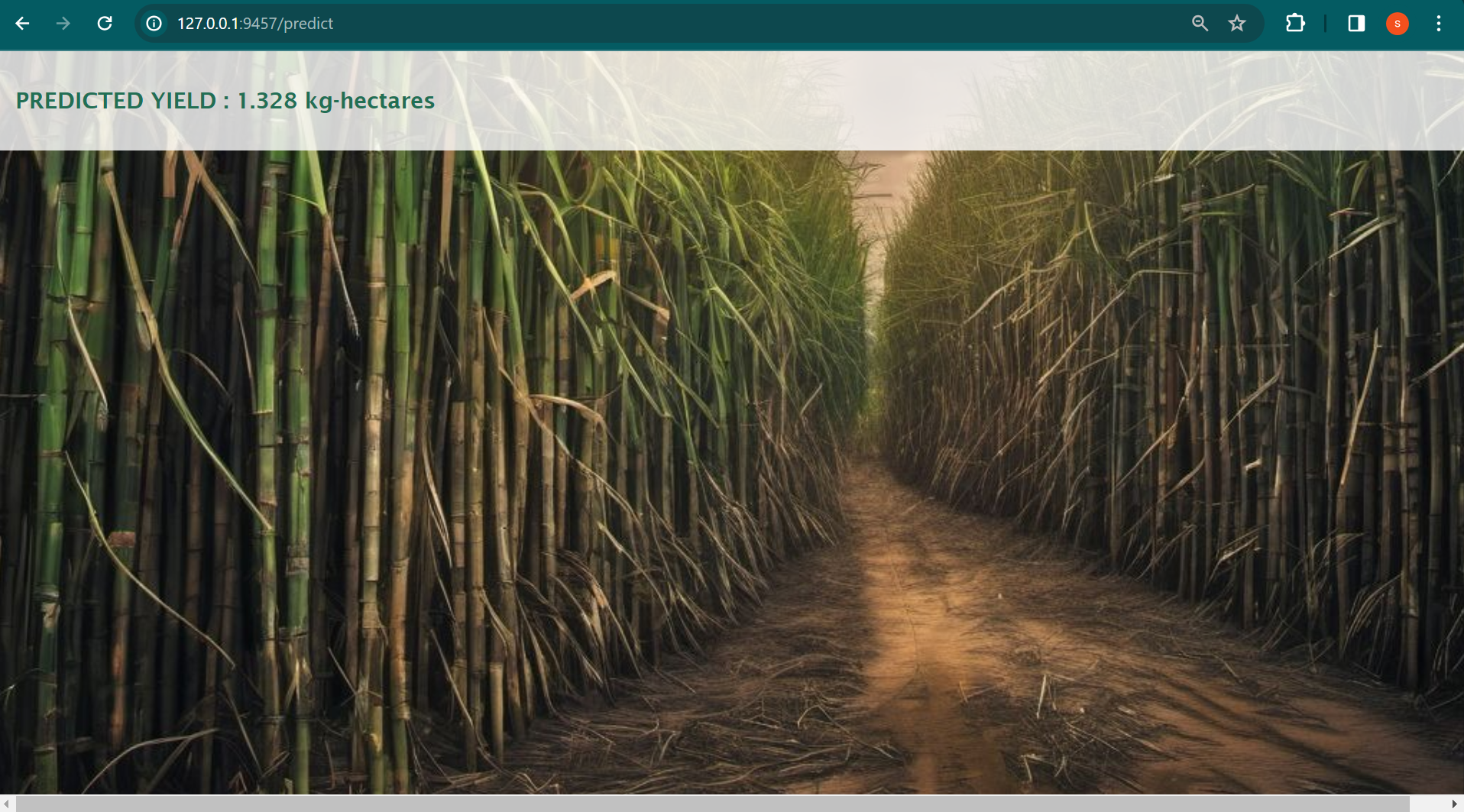


**Figure 6.9**: Saved MLR model using joblib library

After saving the model, we deployed the model using flask making a user-friendly interface, namely Sweet Sage shown in Figure 6.9.



**Figure 6.10**: Web Page



**Figure 6.11:**Predicted yield obtained after entering the meteorological data.

However, when comparing these yields to the actual sugarcane production data released by the government it was found that our predictions are slightly exaggerated.

**6.2 CONCLUSION**

Sugarcane is one of the most important cash crops in India. Its cultivation is a major contributor in the Indian economy and provides livelihood to a major part of the population, especially in the northern parts of India. Thus, making the task of sugarcane yield prediction a very important one as it can impact the lives of millions of people across the country.

Our study concluded that the MLR model outperforms RF as well as LSTM. Moreover, our solution approach of using NDVI and weather data values along with machine learning algorithms proved accurate and efficient.

To put it all together, the detailed analysis and experimentation carried out by us raise the conclusion that the model we have implemented is a highly accurate one in predicting the sugarcane yield of The Sugarcane Belt in Uttar Pradesh and takes us one step closer to a holistic and affordable solution.

**6.3 FUTURE WORK**

· Customize the solution approach so that predictions for any area of interest can be made.

· Improving the MLR model to improve its efficiency and accuracy.

**REFERENCES**

[1] Preeti Saini, Bharti Nagpal, Puneet Garg, Sachin Kumar, "Evaluation of Remote Sensing and Meteorological parameters for Yield Prediction of Sugarcane (Saccharum officinarum L.) Crop," vol. 66, e23220781, 2023.

[2] Edouard Pignède, Philippe Roudier, Arona Diedhiou, Vami Hermann N’Guessan Bi, Arsène T. Kobea, Daouda Konaté, and Crépin Bi Péné, "Sugarcane Yield Forecast in Ivory Coast (West Africa) Based on Weather and Vegetation Index Data," vol. 12, p. 1459, 2021.

[3] Sandeep Kumar Singla, Rahul Dev Garg, Om Prakash Dubey, "Ensemble Machine Learning Methods to Estimate the Sugarcane Yield Based on Remote Sensing Information," vol. 34, no. 6, pp. 731-743, 2020.

[4] Stephen Leo, Massimiliano De Antoni Migliorati, Peter R. Grace, "Predicting within-field cotton yields using publicly available datasets and machine learning," vol. 113, pp. 1150–1163, 2021.

[5] Ramesh A. Medar, Vijay S. Rajpurohit, Anand M. Ambekar, "Sugarcane Crop Yield Forecasting Model Using Supervised Machine Learning," International Journal of Intelligent Systems and Applications (IJISA), vol. 11, no. 8, pp. 11-20, 2019. DOI: 10.5815/ijisa.2019.08.02.

[6] Sunil Kumar Jha, Virupakshagouda C. Patil, Rekha B.U, Shyamal S. Virnodkar Sergey A. Bartalev, Dmitry Plotnikov, Evgeniya Elkina, Nilanchal Patel, "Sugarcane Yield Prediction Using Vegetation Indices in Northern Karnataka, India," Universal Journal of Agricultural Research, vol. 10, no. 6, pp. 699-721, 2022. DOI: 10.13189/ujar.2022.100611.

[7] M. Rahimi Jamnani, A. Liaghat and N. Sadeghi Loyeh, "Sugarcane yield prediction at farm scale using remote sensing and artificial neural network," Water Utility Journal, vol. 24, pp. 1-10, 2020.

[8] HAYAM R. SEIREG, YASSER M. K. OMAR, FATHI E. ABD EL-SAMIE, ADEL S. EL-FISHAWY, AHMED ELMAHALAWY, "Ensemble Machine Learning Techniques Using Computer Simulation Data for Wild Blueberry Yield Prediction," vol. 10, 2022. DOI: 10.1109/ACCESS.2022.3181970.