**JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY**



**Minor Summary Report**

**TOPIC: Sweet Sage - Sugarcane yield prediction using ML algorithms and NDVI data**

### **Submitted To: Supervision of:**

### Mr. Vivek Kumar Singh Dr. Kavita Pandey

Ms. Suma Dawn Ms. Bhawna Saxena **Submitted by:**

Ms. Sonal Nikita Bansal (21103069) B3

Mr. Shivendra Kumar Singh Samyak Jain (21103075) B3

Mr. Shailesh Ribha Nishal (21103098) B4

**1. Motivation behind the project:**

India is an agricultural country and the major population of our country is directly or indirectly dependent on agriculture. It also has a huge role in the GDP of our country. However, a notable challenge persists as a substantial number of Indian farmers continue to operate with traditional methods, and they need to be made aware of modern technologies but they are still using old techniques for farming and prediction as well. This causes them loss in farming and sometimes increases the debt as well.

We took motivation from this and thought to make some changes for the country and made a prediction model for sugarcane crops. The model can predict them with better accuracy about sugarcane crops. With accurate yield predictions, farmers can use resources more efficiently. It will make Economic Planning better as accurate yield predictions contribute to better economic planning for both farmers and stakeholders in the sugarcane industry. This information is valuable for market forecasting, price stability, and overall economic sustainability. Understanding and optimizing agricultural practices through NDVI data can have positive environmental implications.

The motivation for this project comes from our desire to transform sugarcane farming practices through advanced technologies. By using Machine Learning (ML), utilizing Normalized Difference Vegetation Index (NDVI) data and meteorological data, the project aims to address key challenges in agriculture. Our primary objective is to promote precision agriculture, optimizing resource allocation, managing risks associated with crop production, and contributing to the development of economically viable farming practices for the farmers.

**2. Type of project:**

The project is categorized as a **research cum development** project. While it involves the development of practical applications, its foundation lies in thorough research to identify gaps in existing methodologies and integrate innovative approaches for sugarcane yield prediction.

**3. Critical Analysis of research paper read and gaps in work**

During the literature review, gaps in existing methodologies for sugarcane yield prediction were identified as some need for more accuracy in the model. After going through numerous research papers it was found that ML algorithms work well with up to 90% accuracy in this field. So in our project, we aim to improve **the model by 95% accuracy using the Multiple Linear Regression(MLR) ML Algorithm**, coupled with the integration of NDVI data. The project employs tools such as sci-kit-learn in Python for ML implementation and utilizes remote sensing software for processing NDVI data.

**Details of new technologies, tools, and software**

To begin with, our project is in python written on Jupyter Notebook using numerous machine learning algorithms.

Remote sensing and NDVI Data were extracted through Sentinel-2 Imagery, a dataset provided by the COPERNICUS program. These satellites provide high-resolution multispectral imagery, including NDVI data, which were used for monitoring sugarcane health and predicting yields. The ERA5-Land dataset provided by ECMWF was used to get high-quality global meteorological and land surface data, including variables such as temperature, precipitation, humidity, and other essential climate parameters on a daily basis.

**Technologies Used:**

**Jupyter Notebook:** Jupyter Notebook serves as the primary platform for developing and presenting the project. It allows for interactive coding and integration of documentation.

**Python Programming Language:** Python is the programming language used for implementing the project. Its versatility and extensive libraries make it suitable for data analysis and machine learning tasks.

**Tools:**

**Normalized Difference Vegetation Index (NDVI):** NDVI data is utilized to measure and analyze plant health and vegetation cover. It plays a crucial role in assessing the conditions of sugarcane crops.

**Meteorological Data:** Meteorological data is incorporated to understand weather patterns and their impact on sugarcane growth. Parameters such as temperature, precipitation, and humidity contribute to the analysis.

**Machine Learning Models:**

**LSTM (Long Short-Term Memory):** LSTM is a type of recurrent neural network (RNN) suitable for sequence prediction tasks. In this project, LSTM is applied for time-series analysis of NDVI and meteorological data.

**Random Forest (RF):** RF is an ensemble learning method used for classification and regression. In the project, RF is employed for predictive modeling, offering robust and accurate results.

**Multiple Linear Regression (MLR):** MLR is a statistical method used for modeling the relationship between multiple independent variables and a dependent variable. In this project, MLR aids in understanding the correlation between various factors affecting sugarcane farming.

**Software:**

**Scikit-Learn and TensorFlow Libraries:** Scikit-Learn is used for implementing machine learning algorithms, while TensorFlow facilitates the development and training of deep learning models, including LSTM.

**Pandas and NumPy Libraries:** Pandas is employed for data manipulation and analysis, and NumPy is used for numerical operations, enhancing the efficiency of data handling.

**Matplotlib and Seaborn Libraries:** Matplotlib and Seaborn are utilized for data visualization, providing insightful graphs and plots to interpret the results effectively.

**4. Overall design of the project:**

The project design is comprehensive, incorporating various visual representations. A flowchart illustrates the sequential processes of data collection, preprocessing, model training, and yield prediction. Additionally, a workflow model outlines both sequential and parallel processes within the project.

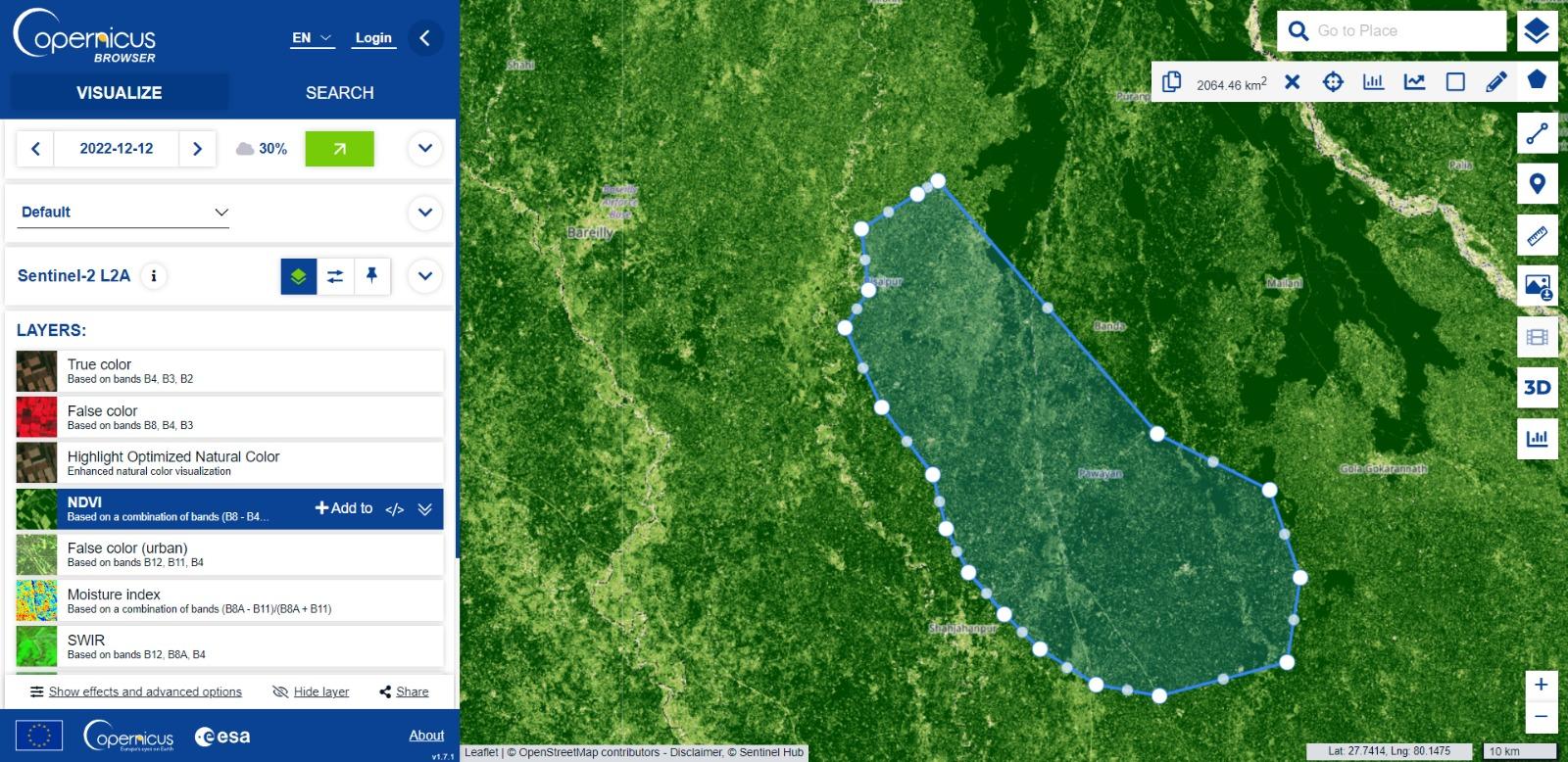
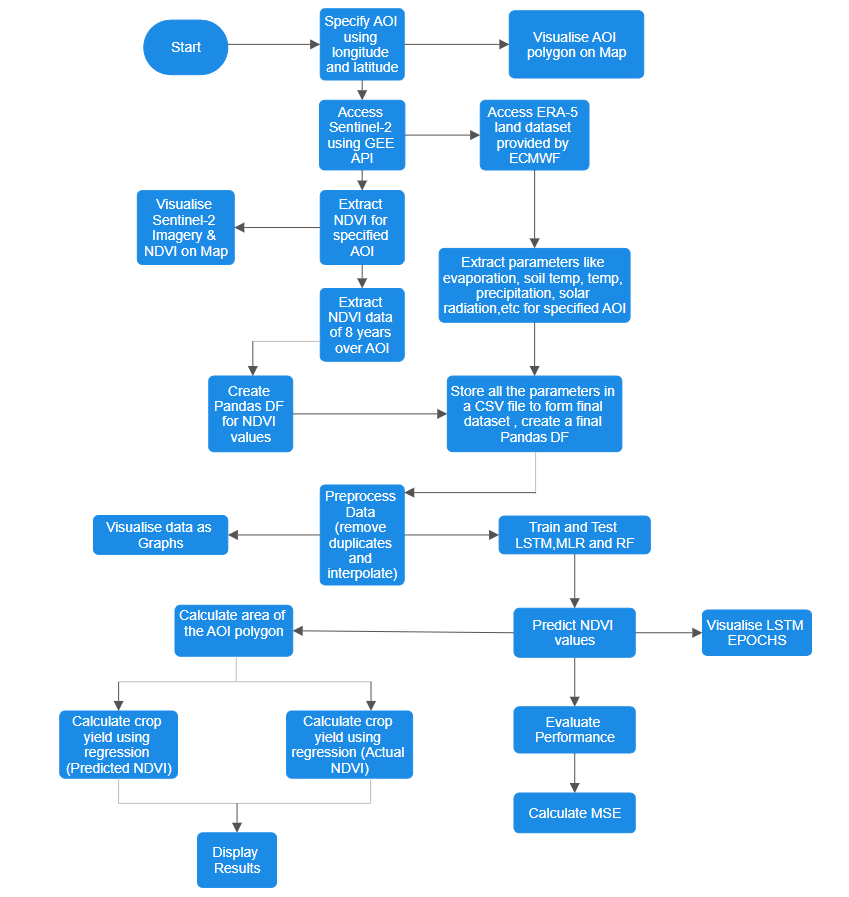


Fig1 : AOI

****

**Fig2: Flow Chart**

**5. Features built, the Programming language used:**

The project encompasses features such as data preprocessing, ML model training, and sugarcane yield prediction. The primary programming language is Python, chosen for its versatility and extensive libraries. Libraries such as sci-kit-learn, pandas, and numpy are employed for ML implementation.  
a) data preprocessing

b) ML libraries - numpy, pandas, folium, sci-kit-learn, google earth engine, etc.

c) Meteorological data - temperature, humidity, soil temperature, etc.

d) Remote sensing data - NDVI data

**6. Proposed Methodology:**

The proposed methodology involves a systematic approach. ML models, like Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM), are then trained using the integrated dataset. The trained models are subsequently applied for accurate sugarcane yield prediction. The procedural steps include -

* **Data Collection**: Gather NDVI and meteorological data from satellites and available datasets.
* **Area Selection**: The Sugarcane Belt of Uttar Pradesh consists of many districts, including Muzaffarnagar, Saharanpur, Meerut, Baghpat, Bulandshahr, Ghaziabad, Amroha, Moradabad, Bijnor, Rampur, etc so a polygon is constructed.
* **Feature Selection**: Use Sentinel-2A satellite to access satellite images to get NDVI data and ERA5-Land dataset provided by ECMWF to get meteorological data like temperature, humidity, soil temperature, precipitation, etc**.**
* **Data Preprocessing**: Clean the duplicate data and interpolate for the missing values of NDVI and weather parameters to ensure a representative dataset distribution, and handle outliers.
* **Model Learning**: Develop a predictive model using the above features and train the model using different ML algorithms like Multiple Linear Regression (MLR), Random Forest (RF), Long Short-Term Memory (LSTM).
* **Model Testing and Evaluation**: Assessed model performance using mean squared error (MSE) metric.
* **Yield Prediction**: Calculated the area of polygon of our area of interest using python’s GEE API and then average actual and predicted yearly sugarcane yield is predicted.

**7. Algorithms of the Work:**

Key algorithms include Random Forest, LSTM, and MLR for ML model training. Data preprocessing includes cleaning and interpolating NDVI and weather data. Model training utilizes cross-validation techniques to validate the accuracy and effectiveness of the developed models.

**8. Division of the work among students:**

To ensure a collaborative and efficient workflow, the project is divided into distinct phases. We divided work at all stages ranging from feature selection, and data extraction and preprocessing to ML model implementation and result analysis. We all decided and worked together to collect data and decided the area of interest for our project. We divided our project as -

Samyak Jain - Implementation of LSTM model.

Ribha Nishal - Implementation of RF model.

Nikita Bansal - Implementation of MLR model.

Finally after testing our model, we calculated the average yield of sugarcane.

**9. Results:**

The project yields from the MLR model gave promising results, demonstrating a substantial improvement in sugarcane yield prediction accuracy (95%) compared to traditional methods. Then after saving the model, deployment is done using flask creating a user interface. Our project provides valuable insights for farmers so that he can improve his yield per hectare.

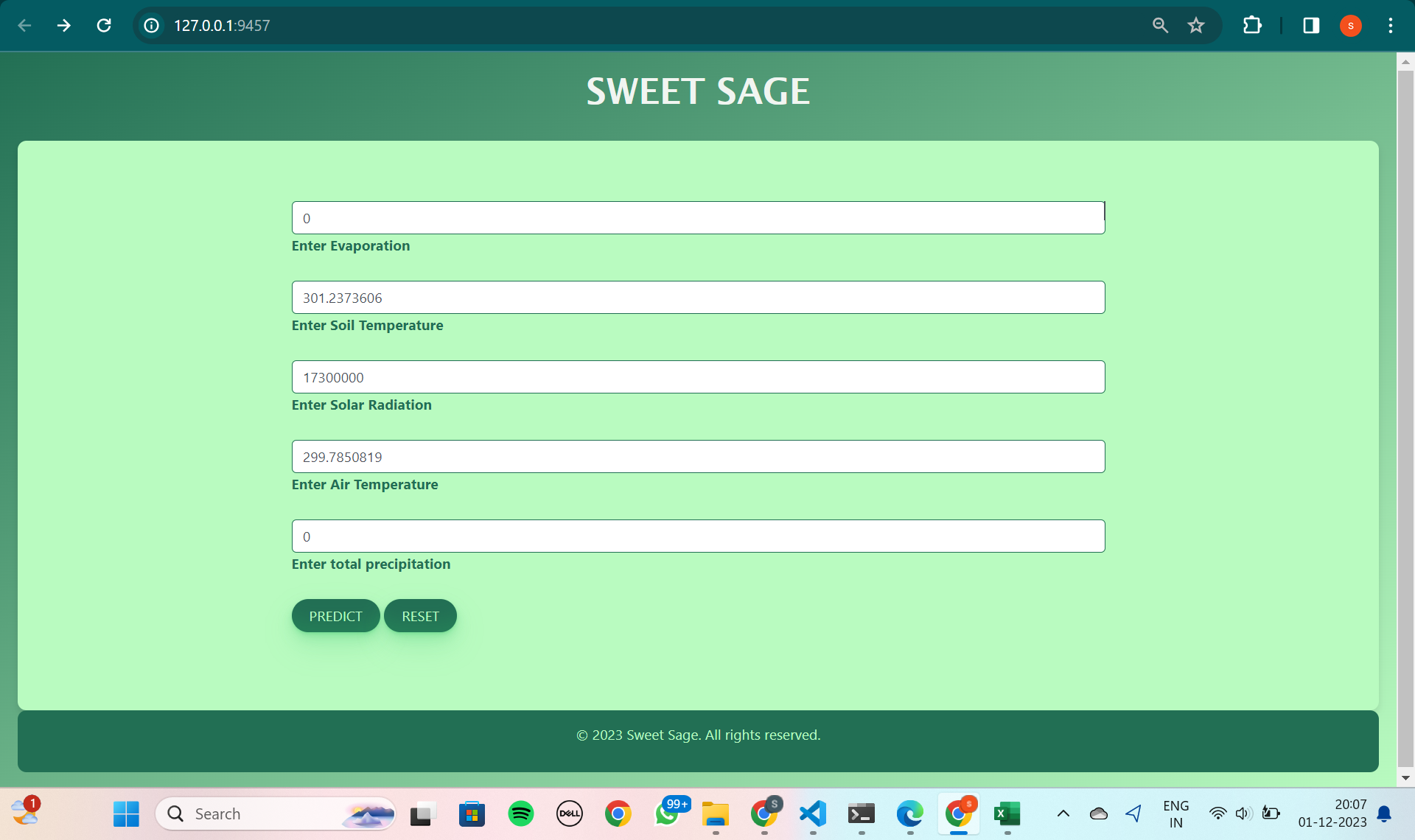


Fig 3. The user interface

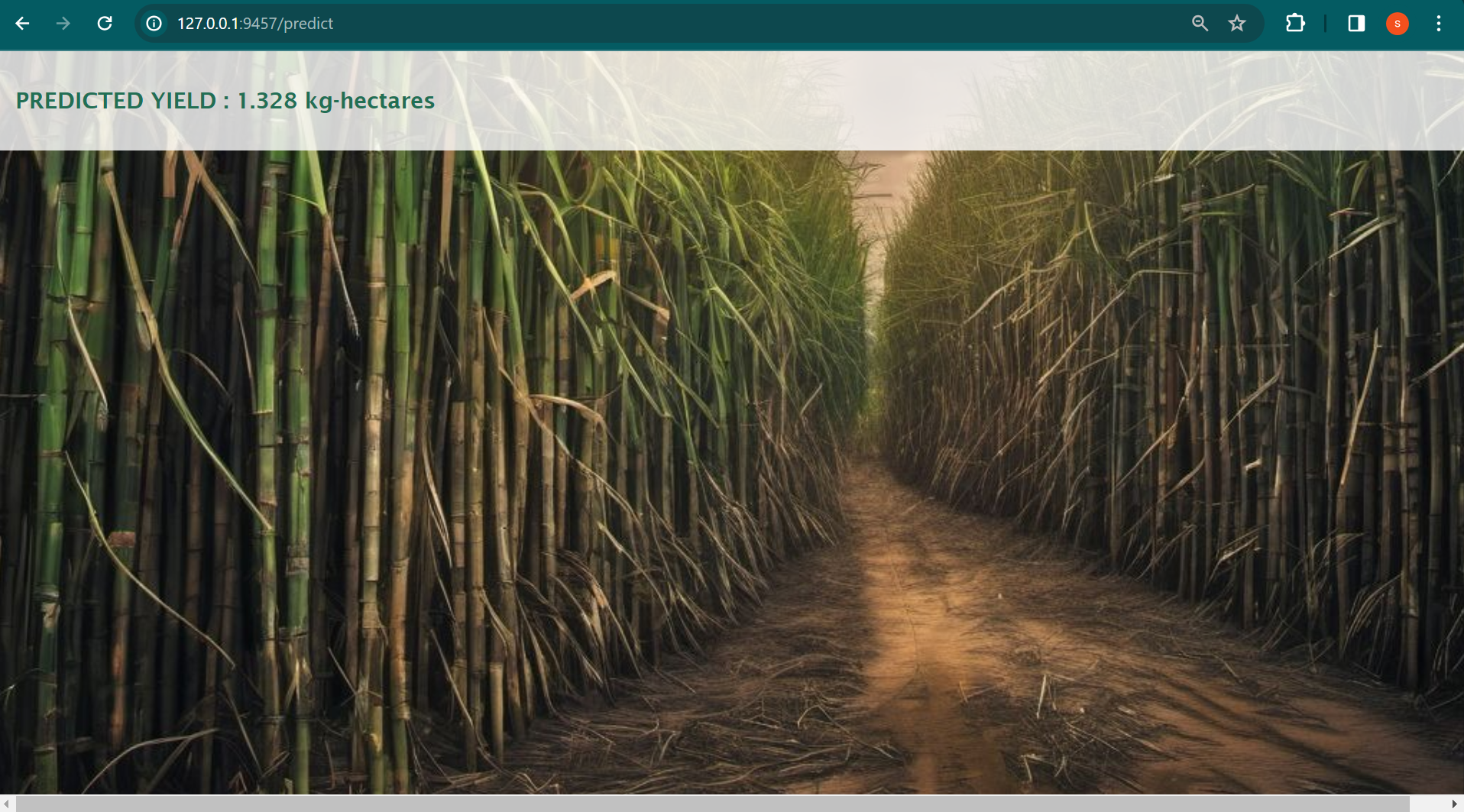


Fig 4. Predicted yield

**10. Conclusion:**

In conclusion, the project effectively utilizes advanced Machine Learning (ML) techniques in conjunction with Normalized Difference Vegetation Index (NDVI) and meteorological data to achieve precise predictions of sugarcane yield. By integrating ML models with NDVI information, the project not only facilitates accurate yield forecasts but also demonstrates the broader potential for transforming agricultural practices.

The significance of this project extends beyond accurate predictions. It underscores the role of precision agriculture in optimizing resource utilization, refining farming methodologies, and ultimately contributing to the establishment of sustainable and economically viable agricultural practices. Through the synergy of ML techniques and NDVI data, the project showcases a pathway towards a more efficient and informed approach to farming.

As the project concludes, it advocates for the continuation of research efforts in this domain and encourages real-world implementations. By doing so, the aim is to further amplify the positive impact of these methodologies on the agricultural landscape. The recommendations for ongoing research and practical applications stem from the belief that the integration of ML and NDVI holds substantial potential in revolutionizing how we approach and manage agriculture, offering benefits not only for farmers but also for the broader sustainability of our agricultural ecosystems.