# A black and red text Description automatically generated

**PROJECT AND TEAM INFORMATION**

## Project Title

|  |
| --- |
| Machine Learning Based Prediction of Crop Yield using Climate and Soil Parameters |

## Student/Team Information

|  |  |
| --- | --- |
| Team Name:  UG\_ML\_AIDS\_30 | AgriVision AI |
| Team member 1 (Team Lead) | Maahi Yadav - 231522650  Maahigeeta12@gmail.com |
| Team member 2 | Shubham Kumar - 231601105  shrshu648866@gmail.com |
| Team member 3 | Sachin Kumar - 23151298  [Skofficial116@gmail.com](mailto:Skofficial116@gmail.com) |

**PROJECT PROGRESS DESCRIPTION**

## Project Abstract

|  |
| --- |
| This project applies **machine learning techniques** to predict crop yield using historical agricultural and climate data. The dataset spans multiple states and districts of India, containing features such as crop type, area under cultivation, soil nutrient requirements, temperature, humidity, rainfall, and pH values. We aim to model the relationship between these variables and crop yield (kg/ha) using multiple ML algorithms.  The project demonstrates the full ML workflow: data cleaning, preprocessing, feature engineering, training regression models (Linear Regression, Random Forest, XGBoost), evaluating their performance, and visualizing predictions. By comparing models, we identify the best approach for accurate yield prediction. This has real-world implications in agricultural planning and policy decision-making. |

## Updated Project Approach and Architecture

|  |
| --- |
| **Dataset:** *Custom\_Crops\_yield\_Historical\_Dataset.csv* (1966–present, district/state/crop level).  **Preprocessing:** Categorical encoding (state, district, crop), normalization, missing value handling.  **Models:** Linear Regression (baseline), Random Forest Regressor, XGBoost Regressor.  **Evaluation Metrics:** R² Score, Mean Squared Error (MSE), Mean Absolute Error (MAE).  **Architecture:**   1. Data ingestion (CSV → Pandas DataFrame). 2. Preprocessing pipeline (encoding + scaling). 3. Model training + hyperparameter tuning. 4. Evaluation + visualization (actual vs predicted, feature importance). |

Tasks Completed

|  |  |
| --- | --- |
| Task Completed | Team Member |
| Project ideation and finalization of scope  Dataset collection and loading  Dataset exploration and descriptive analysis  Handling missing values  Encoding categorical variables  Normalization of numeric features  Baseline Linear Regression model training | All members  Maahi Yadav  Maahi Yadav  Shubham Kumar  Shubham Kumar  Sachin Kumar  Sachin Kumar |

## Challenges/Roadblocks

|  |
| --- |
| One initial challenge was ensuring that all team members had a consistent development environment with the correct versions of the Python libraries. We resolved this by creating a requirements.txt file to standardize the installation process. A second challenge is that the provided dataset, while having non-null values, lacks metadata about the units of production (e.g., metric tons, bushels). For our analysis, we are proceeding with the assumption that the units are consistent across all years. We plan to explicitly state this assumption in our final report.  1. Large dataset (~50k records) slowed down initial processing. 2. Handling categorical variables (many districts/states). 3. Choosing which features to keep for prediction (Year, nutrient columns may not be user-friendly). 4. Early models (Linear Regression) showed limited accuracy. |

## Tasks Pending

|  |  |
| --- | --- |
| Task Pending | Team Member (to complete the task) |
| Implementation of Random Forest model  Implementation of XGBoost model  Model comparison and evaluation  Hyperparameter tuning  Streamlit app development | Shubham Kumar  Sachin Kumar  Maahi Yadav  Sachin Kumar  Shubham Kumar |

## 

## Project Outcome/Deliverables

|  |
| --- |
| At the **current stage (25% completion)**, the project has produced the following **outcomes and intermediate deliverables**:   * **Dataset Cleaning & Preprocessing**   + Removed redundant columns (Year, nutrient totals, IDs).   + Handled missing values using median imputation.   + Encoded categorical features (State Name, District Name, Crop) using LabelEncoder.   + Normalized numeric features (Rainfall, Temperature, Humidity, pH, Wind Speed, Solar Radiation). * **Baseline Machine Learning Model**   + Implemented **Linear Regression** as the first predictive model.   + Conducted an initial performance check using R², MSE, and MAE metrics.   + Established a **benchmark** for future comparison with advanced models. * **In Progress**   + Preparation of model comparison framework (model\_comparison.py) to integrate Random Forest and XGBoost.   + Early visualization setup (scatter plots, performance graphs). * **Pending Deliverables**   + Training and evaluation of **Random Forest and XGBoost** models.   + **Model comparison report** with accuracy metrics and feature importance plots.   + Saving and exporting the **best-performing model** with joblib.   + Development of **Streamlit web application** for user-friendly crop yield prediction.   + Preparation of **final project documentation and presentation slides**. |

# Progress Overview

|  |
| --- |
| So far, our team has completed the **initial 25% of the project work**, which primarily covers **dataset understanding and preparation** along with a **baseline machine learning model**:   * **Data Exploration:** We successfully explored the dataset (~50,000 records), checked distributions, and identified important variables such as rainfall, soil pH, temperature, and crop type. * **Data Preprocessing:** We dropped irrelevant columns (Year, nutrient totals, IDs), handled missing values using median imputation, encoded categorical features (State, District, Crop), and normalized numeric features using StandardScaler. * **Baseline Model:** A **Linear Regression model** was implemented to test the feasibility of predicting crop yields. While the accuracy was limited, it provided a useful benchmark for further improvements.   At this stage, the project has built a **solid foundation** for the machine learning pipeline, ensuring the dataset is ready for advanced modeling. |

# Codebase Information

|  |
| --- |
| * **Repository:** [GitHub/Drive Link] * **Branch:** main * **Key Files (So Far):**   model\_comparison.py → includes baseline Linear Regression training and initial evaluation. |

## 

## Testing and Validation Status

|  |  |  |
| --- | --- | --- |
| Test Type | Status(Pass/Fail) | Notes |
| Missing value handling  Encoding of categorical features  Train/test split validation  Model accuracy evaluation  Streamlit app testing | Pass  Pass  In Progress  In Progress  Pending | Filled with median values.  LabelEncoder applied.  80/20 split used for baseline model.  Linear Regression baseline tested.  Random Forest, XGBoost not done yet. |

# Deliverables Progress

|  |
| --- |
| The key outcomes and deliverables of the project are as follows:   * Dataset Cleaning (✔ Completed) * Baseline Model (✔ Completed) * Model Comparison (❌ Pending) * Streamlit App Development (❌ Pending) * Final Report & Presentation (❌ Pending) |