Crop Yield Prediction

Group-2

Aradhya A. Rathnakar, Mahamaya Panda, ReddySaketh R. Chappidi, and ShashiKumar K. Mallikarjuna

Department of Applied Data Science, San Jose State University

DATA 270: Data Analytics Process

Dr. Eduardo Chan

December 1, 2023

Agenda

- Project Background
- Project Deliverables
- Project Requirements
- Project Workflow
- Technology Survey
- Literature Survey
- Project Planning and Management
- Task Partition
- Project Resource Requirements
- Data Source
- Data Preparation Workflow
- Data Cleaning
- Data Transformation
- Model Selection and Updates
- Model Comparison and Justification
- Conclusion
- References

Project Background

Motivation

- Tackling the global challenge of securing a consistent and sustainable food supply amidst a growing world population.
- Recognizing the impact of supply-and-demand dynamics on agricultural prices and the need for proactive planning through advanced crop yield prediction.

Target Problem

- Crop yield prediction is currently inadequate, limiting the ability of stakeholders to effectively plan and make decisions towards global food security goals.
- Factors like changing climate and growing population present challenges.

Needs

- Collect data on crop yield, location, soil, weather to formulate solution.
- Ensure data quality and develop machine learning models (LSTM, XGBoost Regressor, ARIMA, LightGBM).

Innovation and Impact

- Evaluate models using metrics RMSE, MSE, MAE, and R2 scores to select the best performing ones for accurate prediction.
- Enabling real-time agricultural decisions by farmers to adapt practices based on soil, weather changes; stabilizing global food supply.

Project Deliverables

Deliverables	Description	Delivery Dates
Project proposal	A document outlining the project scope, problem statement, and proposed problem solution along with the SWOT analysis and literature survey related to crop yield prediction.	9/22/2023
JIRA setup	Set up the JIRA environment for project planning and management.	9/29/2023
Work breakdown structure	Project planning framework used to plan the crop yield prediction project using CRISP-DM methodology by breaking down into high-level functional units.	10/1/2023
Gantt chart	Graphical depiction of timeline used to plan schedules for different epics, stories, and sub-tasks based on dependencies for the project and assign them to team members.	10/8/2023
PERT chart	A tool used to plan the Crop Yield Prediction project by graphically mapping out the milestones with their dependencies and estimating the time needed for each milestone based on which we can find the critical path of the project.	10/13/2023
Project Introduction	A document discussing the project background, introducing the problem statement, and proposing a solution. It also contains details about project requirements, project deliverables, technology surveys, and literature reviews.	10/15/2023
Data and project management plan	A document outlining the collection and storage plan for the crop yield, weather, and location data for crop yield prediction. It also holds the planned details about the development methodology, resource allocation, and project schedule.	10/27/2023
Data engineering plan	Document containing the planned details of the data collection, cleaning, transformation, and analysis.	11/3/2023
Machine Learning (ML) model development and evaluation	Develop LSTM, ARIMA, LightGBM, and XGBoost Regressor models, and train them on the transformed data from the Data Engineering processes. Test the models using various evaluation metrics like R2, Mean Squared Error (MSE), etc. to understand the model performance and accuracy.	11/16/2023
ML model deployment to production	Deploy the final version of the models to the production environment for the end-users to access the model and use it for prediction.	11/22/2023
Project presentation	The Project presentation is to showcase the problem statement and a step-by-step breakdown of the implementation of the problem solution.	12/1/2023
Model development reports	Documents that contain the proposed Machine Learning model to solve the Crop Yield Prediction issue along with the model survey, development plan, and model evaluation metrics to choose the best model to solve the prediction problem in the agriculture sector.	12/6/2023
Final project report	The final document summarizes the entire project including the problem statement, proposed problem solution, methodology, source data analysis, model development and evaluation, and solution deployment.	12/8/2023

Project Requirements

Functional Requirements

- Users can input historical crop, soil, weather and location data.
- Machine learning models process the user data to predict crop yield.
- Users can obtain and customize forecasts for their fields that are exportable to other software.

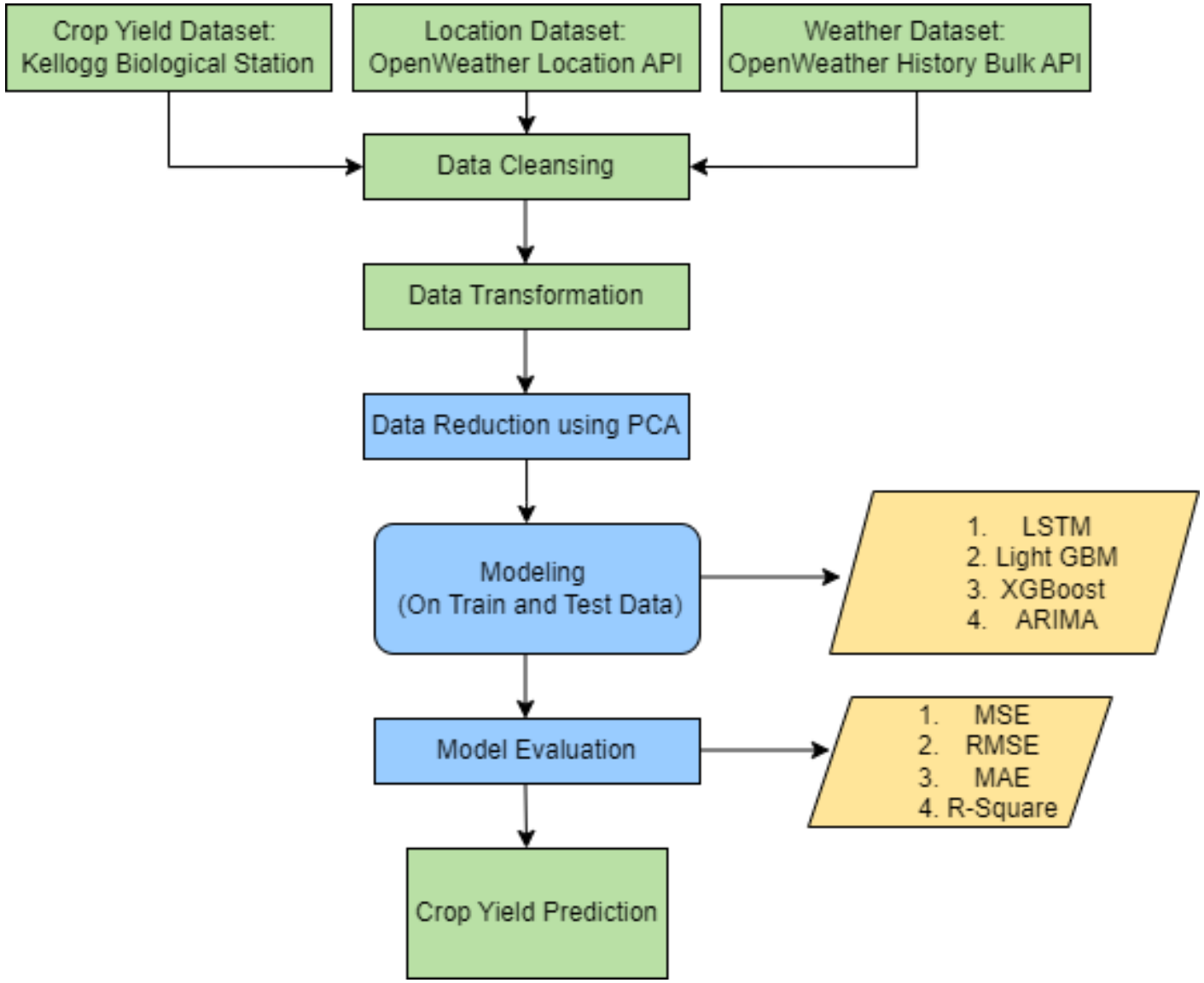
AI Requirements

- Models must be efficient to train on edge devices with constrained hardware using algorithms like LightGBM and XGBoost.
- Models should have acceptable accuracy (<5% error) and should be able to handle acute outliers.
- Models must be able to incorporate emerging factors and adapt to changes over time using flexible algorithms like deep LSTMs.

Data Requirements

- Historical crop yield, soil, weather and location data is needed from sources like Kellogg Biological Station, OpenWeather Weather API, and OpenWeather Geocoding API.
- Data preprocessing is required to verify quality, handle missing values, and structure for modeling.
- A storage strategy is needed to guarantee availability and advanced time-series based analysis.

Project Workflow



TECHNOLOGY SURVEY

Authors	Dataset	Models	Result	Conclusion
Geetha et al. (2022)	Rice crop yield data from 1966 to 1980 collected from ICRISAT website	Stacked LSTM model	Best performance achieved with batch size of 64, Adam and RMSProp optimizers	Including additional features like weather and soil data could improve prediction of crop yields
Saini and Nagpal (2022)	Yearly wheat yield data from 1950 to 2019 from Directorate of Economics and Statistics, India	Deep LSTM model	Deep LSTM model achieved least RMSE of 0.2 compared to other models	Hybrid methods and extensive dataset can improve predictions for wheat yields
Garg et al. (2021)	PlantVillage dataset containing 54,000 crop images across 38 classes and six crops	LightGBM, CNN models like VGG16, ResNet, DenseNet	LightGBM achieved 96.5% accuracy, VGG16 achieved 98.2% accuracy for disease detection	IoT and machine learning can aid better crop growth through damage/disease prediction
Begum et al. (2023)	Dataset from Kaggle on soil nutrients, weather parameters	LightGBM algorithm	Achieved 90% accuracy, 0.82 precision, 0.90 recall, 0.85 F1 score	LightGBM system can effectively match predicted and actual optimal crops
Islam et al. (2023)	Major crop data from Bangladesh Bureau of Statistics and weather data from Bangladesh Meteorological Department from 1976 to 2018	EnsemblMachine Learning Approach (EMLA) using Catboost Regressor and XGBoost Regressor	EMLA achieved 88.084%-91.776% accuracy for different crop predictions based on R-squared scores	Ensemble machine learning improves predictions for overcoming food difficulties and making efficient farming decisions in Bangladesh by leveraging strengths of individual models.
Mariadass et al. (2022)	Historical crop yield data with attributes related to climate fluctuations and pesticides from Malaysia	XGBoost machine learning algorithm	XGBoost model achieved 0.98 R-squared value for variable evaluation using SHAP	Machine learning frameworks like XGBoost can accurately predict crop yield. A universal regression model can help predict crop yield across countries.
Haque et al. (2022)	Weather prediction data including temperature, humidity, wind speed for Bangladesh wheat	ARIMA, ARIMA-WNN hybrid models	ARIMA(2,1,1) model achieved 88% accuracy for wheat production forecasting	Accurate early wheat production predictions aid better farming decisions in Bangladesh
Behera et al. (2023)	Rainfall and groundwater data from Indian government websites	Gradient boosting algorithm, ARIMA model	ARIMA achieved 87.5% accuracy, gradient boosting achieved 91.2% accuracy	The models can accurately suggest optimal crops based on rainfall and underground water analysis for increased yields and profits in India

LITERATURE SURVEY (Contd.)

AUTHORS	DATASET	MODELS	RESULT	CONCLUSION
Sneha and Bhavana (2023)	Crop, soil and weather data from different states in India including parameters like area, temperature, rainfall etc.	Decision Tree, Multiple Linear Regression, Random Forest for yield prediction. Autoregressive Integrated Moving Average (ARIMA) model for price prediction.	Decision Tree performed best for yield prediction with 97.24% accuracy. ARIMA performed best for price prediction.	Combining machine learning algorithms and time series analysis helps farmers avoid losses and make informed decisions for sugarcane farming by predicting yield and price accurately.
Bramantoro et al. (2022)	Weather data from BDMD and a reservoir station in Brunei including temperature, rainfall, humidity, wind speed and direction.	ARIMA, Linear Regression, Artificial Neural Network, Decision Tree	Artificial Neural Network performed best for wind speed prediction. ARIMA performed best for temperature prediction with lowest MSE score.	Data analytics techniques like machine learning can effectively forecast weather parameters and their impact on paddy yield. This helps achieve precision agriculture and improve farming decisions in Brunei.
Mehrmolaei and Keyvanpour (2016)	Monthly New York City births dataset from January 1946 to December 1959 containing 168 data points.	ARIMA model, proposed improved ARIMA model using estimation error mean.	The proposed approach achieved lower MSE, RMSE and MAE compared to basic ARIMA model on test data, demonstrating improved forecasting accuracy.	The proposed approach was shown to outperform the basic ARIMA model for time series forecasting based on experimental results on the NYC births dataset.
Kandan et al. (2021)	Crop production data along with parameters like temperature, humidity, rainfall from districts in India.	Random Forest (RF) and Decision Tree (DT) classifiers.	Random Forest achieved an accuracy of 90% whereas Decision Tree only had 20% accuracy.	Random Forest performs better than Decision Tree in addressing the impact of climate factors on crop production. It provides a better predictive model for crop planning and management compared to Decision Tree, based on the dataset from India.
Krishna et al. (2022)	Crop production data from 1991-2020 for India including variables like crop yield, rainfall, temperature.	K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF) classifiers.	KNN performed best with an accuracy of 97%, followed by LR and RF.	Machine learning algorithms can effectively predict crop yield trends by considering the inherent uncertainties associated with India's climate and crop production systems. KNN worked best for this dataset in accounting for these uncertainties.
Rai et al. (2022)	Crop yield data including parameters like rainfall, temperature for major crops in India.	Logistic Regression (LR), Decision Tree (DT), Linear Regression (LR), Random Forest (RF) classifiers.	Random Forest achieved the highest accuracy of 94% for crop yield prediction, followed by Decision Tree.	Random Forest is shown to be an effective machine learning approach for crop yield prediction, which can support India's economic growth through informed agricultural planning and policymaking. Accurate yield prediction is important considering the sector's contribution to India's GDP.
Dwivedi et al. (2021)	Nifty 500 indices historical data from January 2000 to March 2020.	Seasonal Auto-Regressive Integrated Moving Average (SARIMA), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM).	After cross-validation, the testing MSE of SARIMA was highest at 0.003842 compared to CNN and LSTM which were similar with MSE of 0.003369 and 0.003300 respectively.	Deep learning models like CNN and LSTM outperform regular machine learning SARIMA model for trend prediction on time series data like stock price indices.
Ranjani et al. (2021)	Climate, crop and location datasets from Indian government websites like data.gov.in and imd.gov.in.	Random Forest algorithm for crop yield prediction.	The study uses Random Forest algorithm for crop yield prediction which aids farmers in selecting the best crop to cultivate based on economic and environmental changes.	Crop yield prediction using machine learning can help farmers confront challenges due to unpredictable variations in yield. It also mentions the possible usage of web pages by users to get crop yield prediction by providing crop and climate details as inputs.

LITERATURE SURVEY

Authors	Dataset	Models	Result	Conclusion
Mateo-Sanchis et al. (2023)	Crop yield data from 2015-2018 obtained from the US Department of Agriculture website, along with remote sensing/satellite data and Meteorological data obtained from the zenodo website.	Remote Sensing Data LSTM	When comparing single-variable and multi-variable input to the LSTM model, the LSTM model with multi-variable input provided a better R2 score which was greater than 0.56.	LSTM is a useful model for crop yield prediction. Using IG and SHAP attributions help improve the interpretability of the LSTM model and the crop yield data.
Babbar et al. (2023)	Soil data including details like soil temperature, humidity, fertilizers, phosphorus, nitrogen content. Weather data including temperature, rainfall.	LSTM neural network was used. Two versions of the LSTM model were trained - one with soil data and one with weather data.	The version of the LSTM model trained with weather-related data had better accuracy of 86% and lower mean absolute error (MAE) of 0.389 compared to the version trained with soil data which had accuracy less than 35% and MAE of 0.1695.	Weather data is better than soil data alone for predicting wheat yield when using an LSTM model.
Venkatesh. et al. (2022)	Includes variables such as production, temperature, humidity, rainfall from a crop production dataset.	Simple Linear Regression (SLR) and Polynomial Regression (PR) were used. Group 1 data was used for SLR and Group 2 for PR.	SLR achieved a higher accuracy of 95.40% compared to PR which had an accuracy of 75%.	SLR performs better than PR for the task of crop selection based on the dataset. Considering other parameters like soil type, crop type, fertilizer could help improve the accuracy of machine learning models.
Aruvansh et al. (2019)	Includes crop yield data, weather data like temperature, rainfall collected over 17 years, and 100 years of temperature and rainfall data.	Simple RNN, LSTM, Random Forest (RF), XGBoost were used.	RF provided the best accuracy of 68% for crop name prediction. Simple RNN had the lowest mean absolute error for temperature and rainfall prediction.	Random Forest performs best for crop yield prediction. The study addresses the technology gap in Indian agriculture by offering an applicable solution using machine learning to aid farmers. Simple RNN works best for weather prediction per the results.
Seireg et al. (2022)	Meteorological data from a weather station and simulation data from the Agricultural Production Systems Simulator model.	LightGBM, Gradient Boosting Regression (GBR), XGBoost, Ridge models.	Stacking combination of LightGBM, GBR, XGBoost and Ridge models provided highest accuracy with R2 value of 0.89.	Ensemble machine learning approach of stacking multiple models is effective for predicting wild blueberry yield and has potential for predicting yields of other crops when more data is available.
Choudhary et al. (2020)	Includes crop yield data, meteorological parameters, vegetation indices collected from an agriculture site in Rajasthan, India.	LightGBM, Random Forest, AdaBoost Regressor models were used.	LightGBM model performed best for both Kharif and Rabi crop datasets based on R2 and RMSE scores.	The YieldPredict tool incorporating machine learning models can assist farmers in making decisions. However, external factors like pests and diseases need to be considered to improve predictive performance.
Indira et al. (2022)	Soil nutrients, rainfall, temperature	MobileNet for disease detection from images, XGBoost to predict suitable crops, Random Forest to propose fertilizers.	Combined model recommends crop for soil with 99% accuracy.	The machine learning recommendation system helps farmers increase crop yield and make informed decisions by identifying diseases, forecasting suitable crops, and recommending fertilizers.
Shuvessa et al. (2022)	Jute production data from 2007-2021 collected from government organizations, climatic data like rainfall, temperature, humidity.	Random Forest, Decision Tree, Gradient Boosting, XGBoost	Decision Tree and Random Forest algorithms achieved highest accuracy of 99% for jute yield prediction compared to other models.	Machine learning algorithms can effectively predict jute yield by considering climate impact. There is scope to develop a recommendations system for jute production and distribution to farmers.

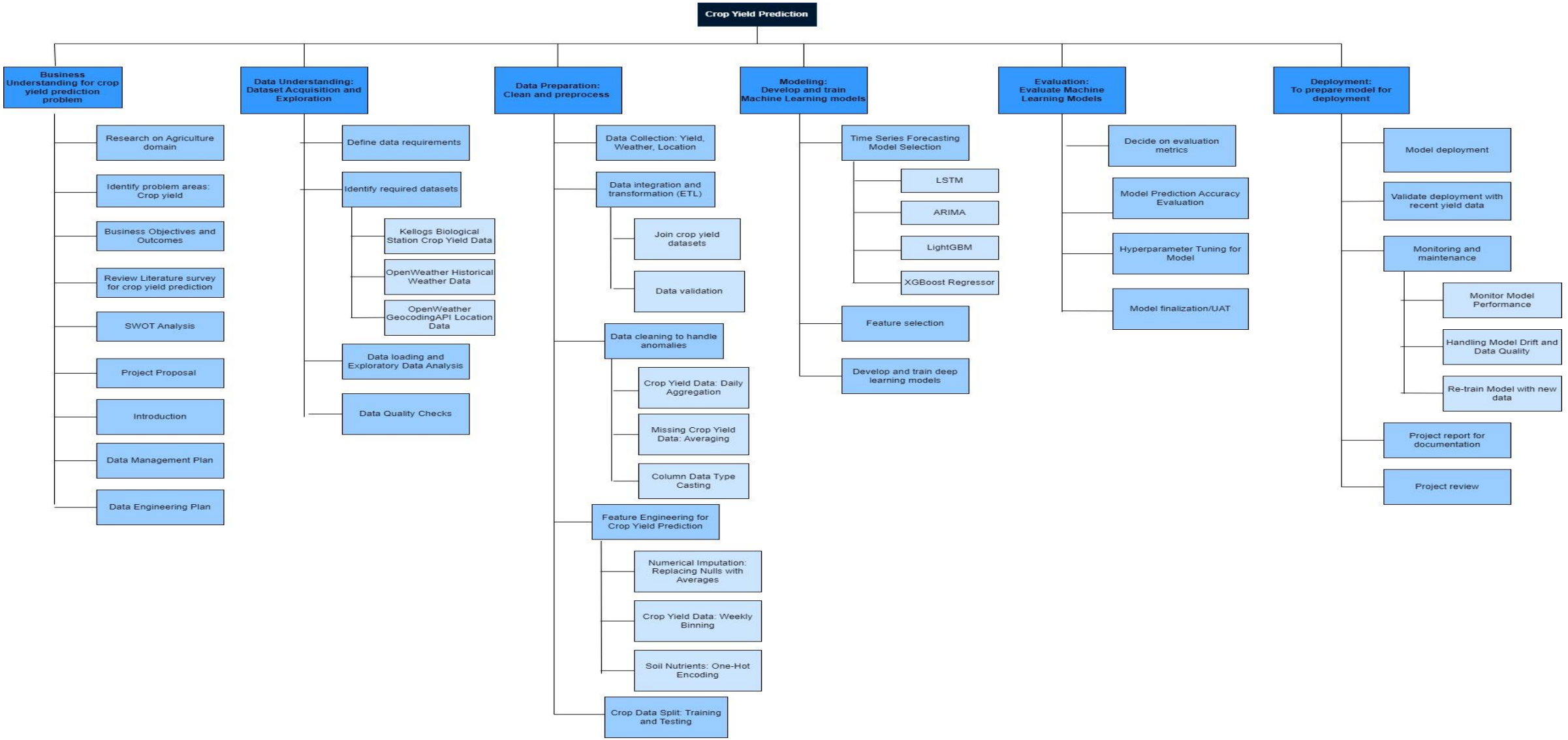
Project Planning and Management

CRISP-DM METHODOLOGY: BUSINESS UNDERSTANDING,
DATA UNDERSTANDING,
DATA PREPARATION,
MODELING,
EVALUATION,
DEPLOYMENT.

EACH PHASE CONTAINS SEVERAL WORK PACKAGES THAT ARE FURTHER BROKEN DOWN INTO TASKS.

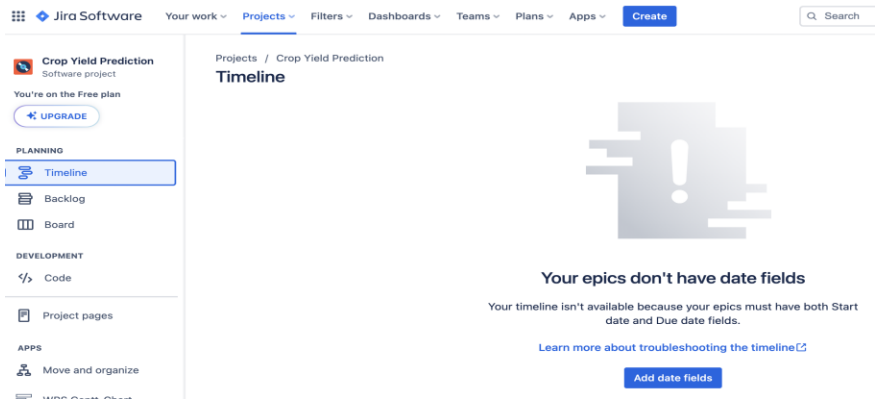
TOOLS USED FOR PROJECT PLANNING AND MANAGEMENT:
JIRA
WORK BREAKDOWN STRUCTURE (WBS)
GANTT CHART
PERT CHART

Work Breakdown Structure

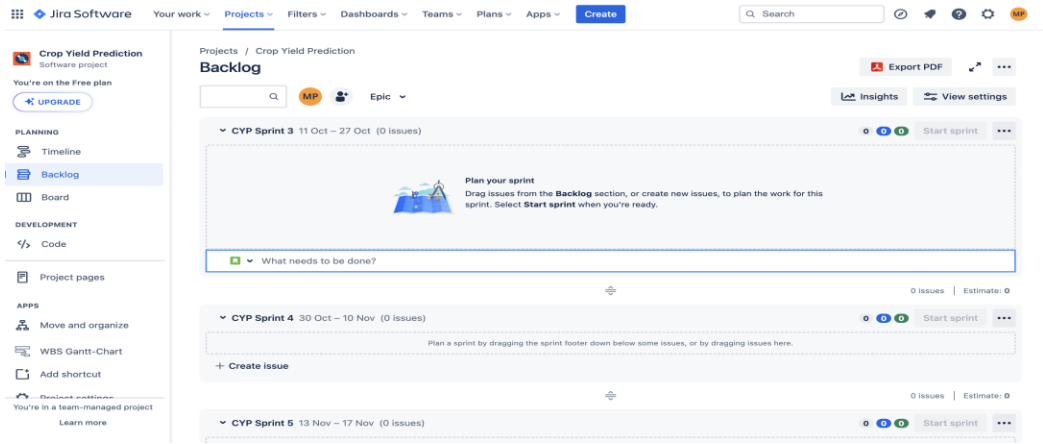


JIRA

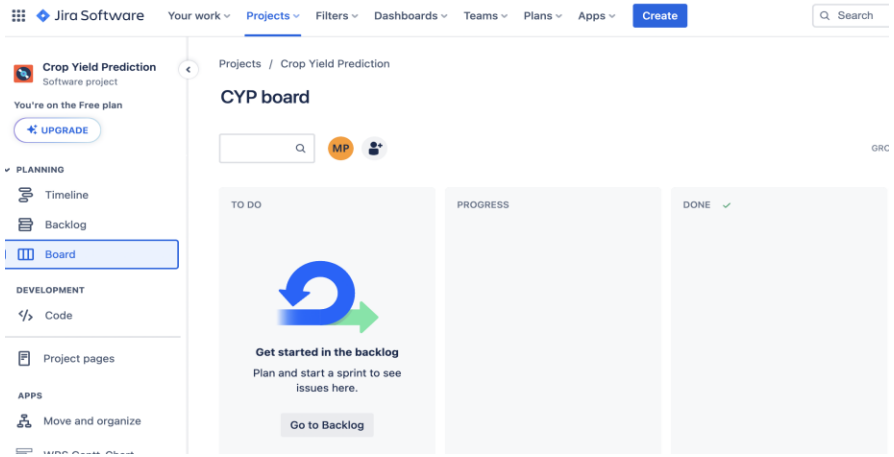
Timeline



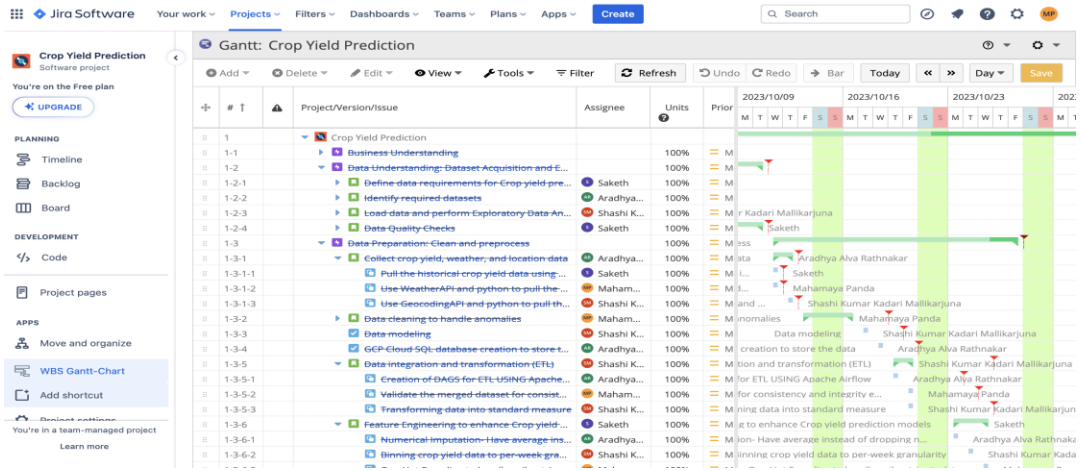
Backlog



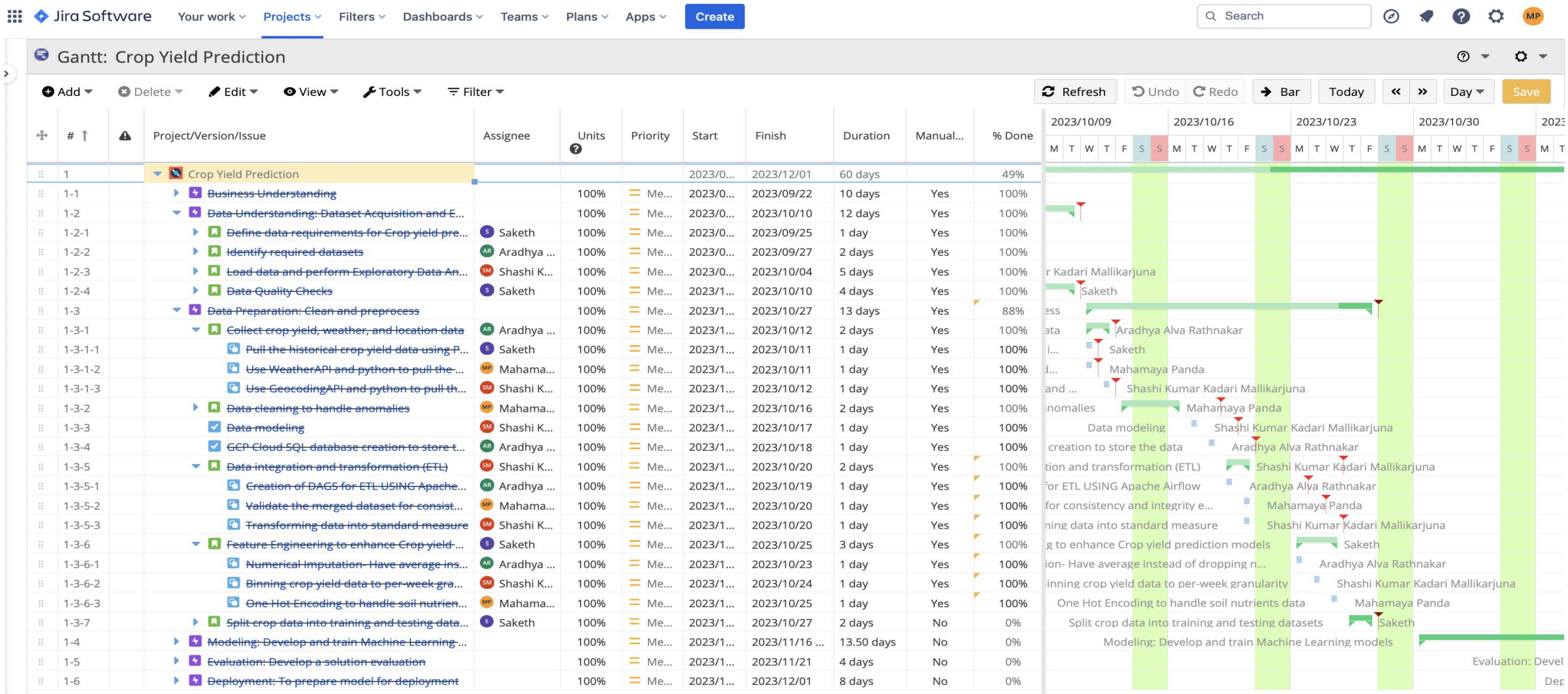
Board

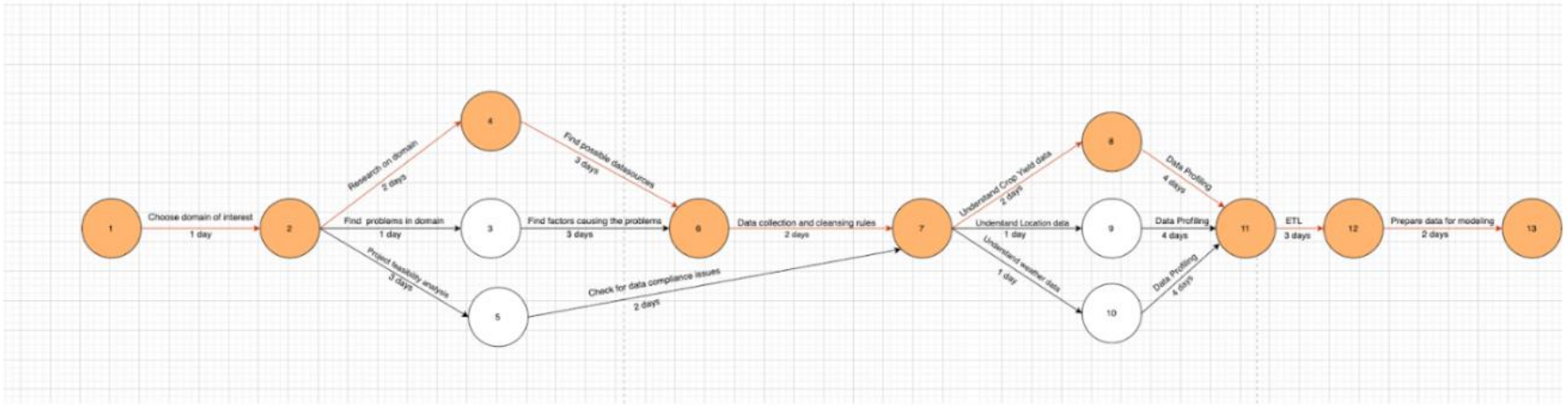


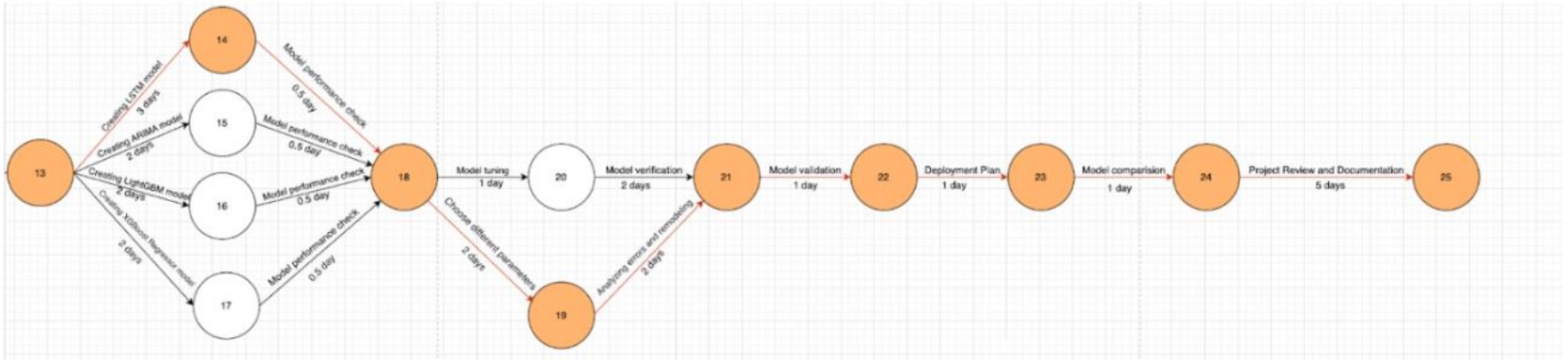
Gantt Chart



GANTT CHART


Jira Software
Your work
Projects
Filters
Dashboards
Teams
Plans
Apps
Create
Search
Gantt: Crop Yield Prediction
Add
Delete
Edit
View
Tools
Filter
Refresh
Undo
Redo
Bar
Today
Day
Save
Project/Version/Issue Assignee Units Priority Start Finish Duration Manual... % Done
1 Crop Yield Prediction 2023/0... 2023/12/01 60 days 49%
1-1 Business Understanding 100% Me... 2023/0... 2023/09/22 10 days Yes 100%
1-2 Data Understanding: Dataset Acquisition and E... 100% Me... 2023/0... 2023/10/10 12 days Yes 100%
1-2-1 Define data requirements for Crop yield pre... Saketh 100% Me... 2023/0... 2023/09/25 1 day Yes 100%
1-2-2 Identify required datasets Aradhya ... 100% Me... 2023/0... 2023/09/27 2 days Yes 100%
1-2-3 Load data and perform Exploratory Data An... Shashi K... 100% Me... 2023/0... 2023/10/04 5 days Yes 100%
1-2-4 Data Quality Checks Saketh 100% Me... 2023/1... 2023/10/10 4 days Yes 100%
1-3 Data Preparation: Clean and preprocess 100% Me... 2023/1... 2023/10/27 13 days Yes 88%
1-3-1 Collect crop yield, weather, and location data Aradhya ... 100% Me... 2023/1... 2023/10/12 2 days Yes 100%
1-3-1-1 Pull the historical crop yield data using P... Saketh 100% Me... 2023/1... 2023/10/11 1 day Yes 100%
1-3-1-2 Use WeatherAPI and python to pull the ... Mahama... 100% Me... 2023/1... 2023/10/11 1 day Yes 100%
1-3-1-3 Use GeocodingAPI and python to pull th... Shashi K... 100% Me... 2023/1... 2023/10/12 1 day Yes 100%
1-3-2 Data cleaning to handle anomalies Mahama... 100% Me... 2023/1... 2023/10/16 2 days Yes 100%
1-3-3 Data modeling Shashi K... 100% Me... 2023/1... 2023/10/17 1 day Yes 100%
1-3-4 GCP Cloud SQL database creation to store t... Aradhya ... 100% Me... 2023/1... 2023/10/18 1 day Yes 100%
1-3-5 Data integration and transformation (ETL) Shashi K... 100% Me... 2023/1... 2023/10/20 2 days Yes 100%
1-3-5-1 Creation of DAGS for ETL USING Apache... Aradhya ... 100% Me... 2023/1... 2023/10/19 1 day Yes 100%
1-3-5-2 Validate the merged dataset for consist... Mahama... 100% Me... 2023/1... 2023/10/20 1 day Yes 100%
1-3-5-3 Transforming data into standard measure Shashi K... 100% Me... 2023/1... 2023/10/20 1 day Yes 100%
1-3-6 Feature Engineering to enhance Crop yield ... Saketh 100% Me... 2023/1... 2023/10/25 3 days Yes 100%
1-3-6-1 Numerical Imputation- Have average ins... Aradhya ... 100% Me... 2023/1... 2023/10/23 1 day Yes 100%
1-3-6-2 Binning crop yield data to per-week gra... Shashi K... 100% Me... 2023/1... 2023/10/24 1 day Yes 100%
1-3-6-3 One Hot Encoding to handle soil nutrien... Mahama... 100% Me... 2023/1... 2023/10/25 1 day Yes 100%
1-3-7 Split crop data into training and testing data... Saketh 100% Me... 2023/1... 2023/10/27 2 days No 0%
1-4 Modeling: Develop and train Machine Learning ... 100% Me... 2023/1... 2023/11/16 ... 13.50 days No 0%
1-5 Evaluation: Develop a solution evaluation 100% Me... 2023/1... 2023/11/21 4 days No 0%
1-6 Deployment: To prepare model for deployment 100% Me... 2023/1... 2023/12/01 8 days No 0%
2023/10/09
2023/10/16
2023/10/23
2023/10/30
Aradhya Alva Rathnakar
Mahamaya Panda
Shashi Kumar Kadari Mallikarjuna
Saketh
Data modeling
Split crop data into training and testing datasets
Modeling: Develop and train Machine Learning models





The critical path for our project is:
1 -> 2 -> 4 -> 6 -> 7 -> 8 -> 11 -> 12 -> 13 -> 14 -> 18 -> 19 -> 21 -> 22 -> 23 -> 24 -> 25 = **34.5 days**

PERT CHART

Task Partition

The project tasks are divided among four members based on their skills and expertise.

The roles assigned are Data Collection, Data Processing, Data Management, Data Exploration, Modeling and Deployment.

Project Resource Requirements

Hardware Specifications

Hardware	Memory	Configuration	Purpose
Chip - 8-core CPU with 4 performance cores and 4 efficiency cores	8GB unified memory	250GB SSD	Computation engine for running ML Model in Jupyter Notebooks
8-core GPU	(Configurable to 16GB)	(Configurable to 500GB SSD, 1TB, or 2TB SSD)	ML Model
16-core Neural Engine	(Configurable to 16GB)		ML Model

Software Specifications

Libraries/Packages	Purpose	Version
Python	Data Cleaning Data Preparation Data Analysis	3.8
Pandas	Creating Dataframe	1.3.2
Numpy	Dealing with Dataframe arrays	3.2
Matplotlib	Data Interpretation using plots	3.2
Seaborn	Advance Interactive plots	0.12.0
sci-kit-learn	Machine Learning Model Building	1.1.2
Tensorflow	Machine Learning Model Building	2.14.0
Keras	Machine Learning Model Building	2.14.0

Project Resource Requirements (Contd.)

Tools and Licenses

Tools	Purpose	License
OpenWeather Weather API	Weather Dataset	$10.00 for getting historical weather data for one location
OpenWeather Geocoding Location API	Location Dataset	Free subscription
GitHub	Source and version control	Free
Draw.io	WBS, PertChart, Figures	Free
Jira	Assigning Tasks, Gantt Chart	Student
DMPTool.org	For designing our Data management plan	Student
Zoom	Project meetings	Student
Google Docs	Documentation	Free
Anaconda	For Python Code	Free

Project Cost and Justification

Resource	Justification	Duration	Cost
OpenWeather Weather API	To collect weather data per day	Once	$10.00 for getting historical weather data for one location

Data Sources

- Historical crop yield data is collected from Kellogg Biological Station in Hickory Corners, MI, spanning the period from October 1996 to July 2013.
- Weather data collected from OpenWeather Weather API.
- Location data collected from OpenWeather Geocoding API.

Crop Yield Dataset

- The Robertson 2016 dataset contains detailed crop yield data from 1996-2013 recorded by IoT sensors at the Kellogg Biological Station in Michigan.
- The raw crop yield data is in a comma-separated text file format with values for variables like crop yield, date, location, weather characteristics.
- File size 100.1 MB
- Record count 767,826

Sample Raw **Crop Yield Dataset**

```
longitude,latitude,crop_flow_lb_s,datetime,duration ,distance_in,swth_wdth_in,moisture,status,pass_num,serial_number,field,dataset,product,elevation_ft
-85.372824,42.408276,3.29,2013-07-24 19:20:04+00,1,56,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372838,42.408288,2.89,2013-07-24 19:20:03+00,1,47,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372847,42.408296,2.89,2013-07-24 19:20:02+00,1,38,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372855,42.408301,3.69,2013-07-24 19:20:01+00,1,18,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372855,42.408304,3.39,2013-07-24 19:20:00+00,1,39,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372847,42.408293,3.99,2013-07-24 19:19:59+00,1,61,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372838,42.408279,6.39,2013-07-24 19:19:58+00,1,67,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372835,42.408263,5.89,2013-07-24 19:19:57+00,1,56,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372835,42.408251,5.59,2013-07-24 19:19:56+00,1,19,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.17
-85.372835,42.408249,4.49,2013-07-24 19:19:55+00,1,10,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372835,42.408249,3.49,2013-07-24 19:19:54+00,1,41,180,6.6,0,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372835,42.40826,2.49,2013-07-24 19:19:53+00,1,62,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372835,42.408276,1.89,2013-07-24 19:19:52+00,1,75,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372833,42.408293,3.39,2013-07-24 19:19:51+00,1,82,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372833,42.408313,3.29,2013-07-24 19:19:50+00,1,89,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.5
-85.372833,42.408332,3.29,2013-07-24 19:19:49+00,1,90,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.92
-85.372833,42.408354,4.39,2013-07-24 19:19:48+00,1,90,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.92
-85.372833,42.408376,3.39,2013-07-24 19:19:47+00,1,90,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.92
-85.372833,42.408396,3.89,2013-07-24 19:19:46+00,1,89,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.92
-85.372833,42.408418,3.39,2013-07-24 19:19:45+00,1,98,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.92
-85.372833,42.40844,3.39,2013-07-24 19:19:44+00,1,103,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),953.92
-85.372833,42.408465,2.99,2013-07-24 19:19:43+00,1,111,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),954.33
-85.372833,42.40849,2.99,2013-07-24 19:19:42+00,1,108,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),954.33
-85.372833,42.408515,2.99,2013-07-24 19:19:41+00,1,108,180,6.6,1,141,5648,F1: Lysimeter,L1: ,Wheat (Sft Rd Wtr),954.33
```

Location Dataset

- Location data such as city, state and country is obtained from OpenWeather Geocoding API by constructing API calls with longitude and latitude coordinates.
- The location data from the Geocoding API is in JSON format, containing fields like latitude, longitude, city, state, country which is then converted to CSV.
- File size 45.9 MB
- Record count 678,411


```
[{'name': 'Ross Township',
  'lat': 42.37264055,
  'lon': -85.3570973784533,
  'country': 'US',
  'state': 'Michigan'}]
```


Sample Raw **Location Dataset** in **JSON Format**

INDEX	longitude	latitude	name	state	country
0	-85.372875	42.410102	Ross Township	Michigan	United States
1	-85.373105	42.409535	Ross Township	Michigan	United States
2	-85.373105	42.409618	Ross Township	Michigan	United States
3	-85.373208	42.408202	Ross Township	Michigan	United States
4	-85.373358	42.409088	Ross Township	Michigan	United States
5	-85.373451	42.408667	Ross Township	Michigan	United States
6	-85.373453	42.409317	Ross Township	Michigan	United States
7	-85.373478	42.40957	Ross Township	Michigan	United States
8	-85.373726	42.408717	Ross Township	Michigan	United States
9	-85.373737	42.409076	Ross Township	Michigan	United States
10	-85.37342	42.409703	Ross Township	Michigan	United States
11	-85.37342	42.408156	Ross Township	Michigan	United States
12	-85.372021	42.407062	Ross Township	Michigan	United States
13	-85.372009	42.407954	Ross Township	Michigan	United States
14	-85.371685	42.407116	Ross Township	Michigan	United States

Sample Raw **Location Dataset** in **CSV Format**

Weather Dataset

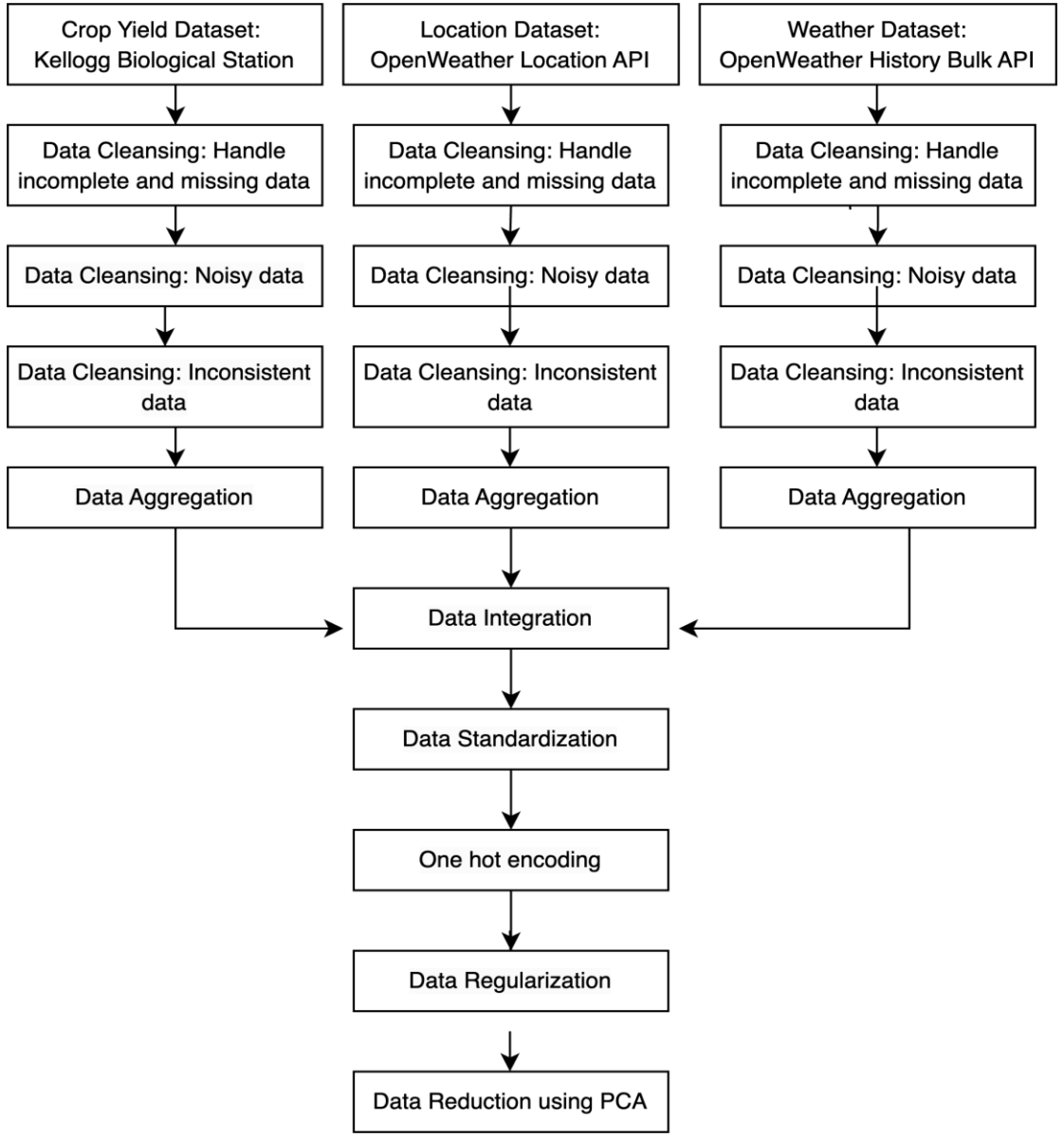
- Weather data including historical hourly data is obtained from OpenWeather Weather API by providing them the city name through their bulk historic weather data download feature.
- The raw weather data is downloaded from the API in CSV format with various weather features like date, city, wind, rain etc. for analysis.
- File Size of 48.9 MB.
- Record count 275,169

1	dt	dt_iso	timezc	city_name	lat	lon	temp	visibility	dew_pc	feels_lil	temp_n
9557	820454400	1996-01-01 00:0	-18000	Ross Township	40.53677	-80.02	276.11	10000	272.72	274.03	275.85
9558	820458000	1996-01-01 01:0	-18000	Ross Township	40.53677	-80.02	276.06		274.11	276.06	275.78
9559	820461600	1996-01-01 02:0	-18000	Ross Township	40.53677	-80.02	275.7		274.38	275.7	275.4
9560	820465200	1996-01-01 03:0	-18000	Ross Township	40.53677	-80.02	275.08	8000	273.92	275.08	274.69
9561	820468800	1996-01-01 04:0	-18000	Ross Township	40.53677	-80.02	275.83		274.2	275.83	275.42
9562	820472400	1996-01-01 05:0	-18000	Ross Township	40.53677	-80.02	275.74	4800	275.17	275.74	275.46
9563	820472400	1996-01-01 05:0	-18000	Ross Township	40.53677	-80.02	275.74	4800	275.17	275.74	275.46
9564	820476000	1996-01-01 06:0	-18000	Ross Township	40.53677	-80.02	275.58	3200	274.86	274.17	274.96

K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
temp_n	temp_n	pressur	sea_lev	grnd_le	humidit	wind_s	wind_d	wind_g	rain_1h	rain_3h	snow_1	snow_3	clouds_	weathe	weathe	weathe	weathe
275.85	276.36	1012			78	2.1	110		0.44				100	500	Rain	light rain	10n
275.78	276.31	1011			87	1.04	121		0.5				99	500	Rain	light rain	10n
275.4	275.93	1011			91	0.88	110		0.83				100	500	Rain	light rain	10n
274.69	275.46	1011			92	0	0		0.78				100	500	Rain	light rain	10n
275.42	276.17	1012			89	0.81	114		0.91				100	500	Rain	light rain	10n
275.46	275.92	1011			96	0	0		0.49				100	741	Fog	fog	50n
275.46	275.92	1011			96	0	0		0.49				100	500	Rain	light rain	10n
274.96	275.77	1011			95	1.5	60		0.19				100	741	Fog	fog	50n
274.96	275.77	1011			95	1.5	60		0.19				100	300	Drizzle	light intens	09n

Sample Raw **Weather Dataset**

Data Pre-process Flow



Missing Data Handling for Historical Crop Yield, Location, and Weather Datasets

Before and After Handling of Missing Values in **Crop Yield Dataset**

```
Missing values before cleaning:
 longitude              0
latitude                0
crop_flow_lb_s          0
datetime                0
duration_               0
distance_in             0
swth_wdth_in            0
moisture                0
status              29250
pass_num                0
serial_number           0
field                   0
dataset                 0
product                 0
elevation_ft            0
dtype: int64
```

```
Missing values after cleaning:
 longitude          0
latitude            0
crop_flow_lb_s      0
datetime            0
duration_           0
distance_in         0
swth_wdth_in        0
moisture            0
status              0
pass_num            0
serial_number       0
field               0
dataset             0
product             0
elevation_ft        0
dtype: int64
```

No Missing Values in **Location Dataset**

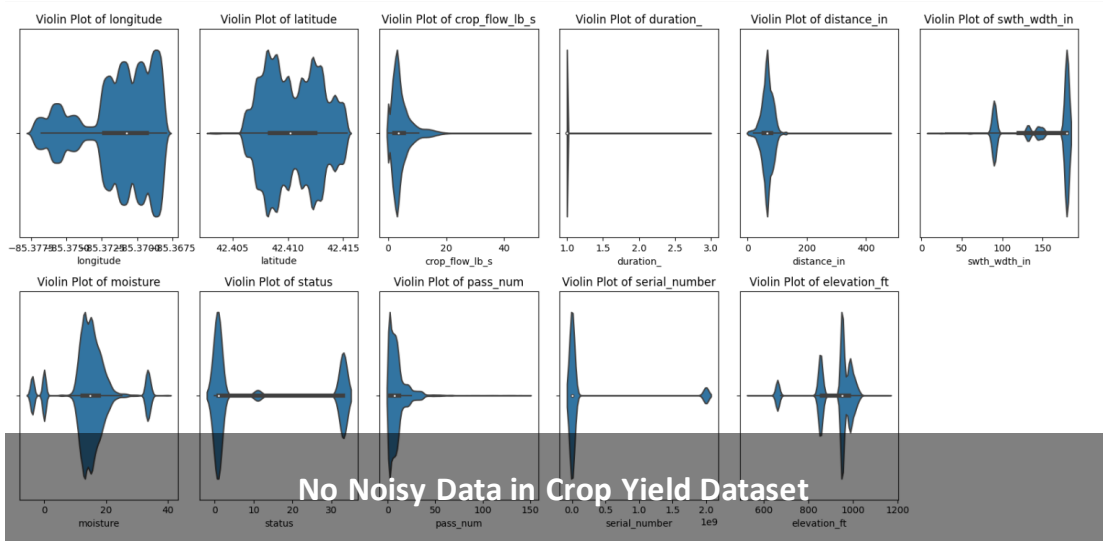
```
Missing values before cleaning:
 INDEX        0
longitude     0
latitude      0
name          0
state         0
country       0
dtype: int64
```

Before and After Handling of Missing Values in **Weather Dataset**

```
dt_iso                      0
timezone                    0
city_name                   0
lat                         0
lon                         0
temp                        0
visibility              29312
dew_point                   1
feels_like                  0
temp_min                    0
temp_max                    0
pressure                    0
sea_level              275168
grnd_level             275168
humidity                    0
wind_speed                  0
wind_deg                    0
wind_gust              226848
rain_1h                227159
rain_3h                275132
snow_1h                265022
snow_3h                275168
clouds_all                  0
weather_id                  0
weather_main                0
weather_description         0
weather_icon                0
dtype: int64
```

```
dt_iso                  0
timezone                0
city_name               0
lat                     0
lon                     0
temp                    0
visibility              0
dew_point               0
feels_like              0
temp_min                0
temp_max                0
pressure                0
humidity                0
wind_speed              0
wind_deg                0
clouds_all              0
weather_id              0
weather_main            0
weather_description     0
weather_icon            0
dtype: int64
```

Noisy Data Handling for Historical Crop Yield and Location Dataset



No Noisy Data in Crop Yield Dataset

```
location['name'].value_counts()
```

```
Ross Township    678411
Name: name, dtype: int64
```

```
location['state'].value_counts()
```

```
Michigan    678411
Name: state, dtype: int64
```

```
location['country'].value_counts()
```

```
United States    678411
Name: country, dtype: int64
```

No Noisy Data in Location Dataset

Noisy Data Handling for Weather Dataset

	timezone	lat	lon	temp	visibility	dew_point	feels_like	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	clouds_all	weather_id
count	275168.000000	275168.000000	2.751680e+05	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000
mean	-15806.766775	40.536772	-8.001996e+01	283.963250	9061.165963	278.738308	282.451308	283.219646	284.884600	1016.517695	72.203527	3.165662	177.430021	67.642851	725.966184
std	1756.524649	0.000000	5.684352e-14	21.841422	2093.190631	10.201194	22.714034	21.858512	10.356082	7.216410	18.969923	2.246658	113.299781	36.537921	125.733979
min	-18000.000000	40.536772	-8.001996e+01	-9725.890000	100.000000	11.810000	-9732.890000	-9726.040000	-2.270000	969.000000	0.000000	0.000000	0.000000	0.000000	200.000000
25%	-18000.000000	40.536772	-8.001996e+01	275.670000	9061.165963	271.070000	272.880000	274.960000	276.450000	1012.000000	59.000000	1.500000	80.000000	40.000000	701.000000
50%	-14400.000000	40.536772	-8.001996e+01	284.695000	10000.000000	279.250000	283.830000	283.900000	285.490000	1017.000000	75.000000	3.100000	200.000000	75.000000	801.000000
75%	-14400.000000	40.536772	-8.001996e+01	292.450000	10000.000000	287.170000	292.410000	291.740000	293.260000	1021.000000	88.000000	4.600000	270.000000	100.000000	803.000000
max	-14400.000000	40.536772	-8.001996e+01	309.810000	10000.000000	302.260000	316.170000	309.930000	317.260000	1043.000000	100.000000	23.100000	360.000000	100.000000	804.000000

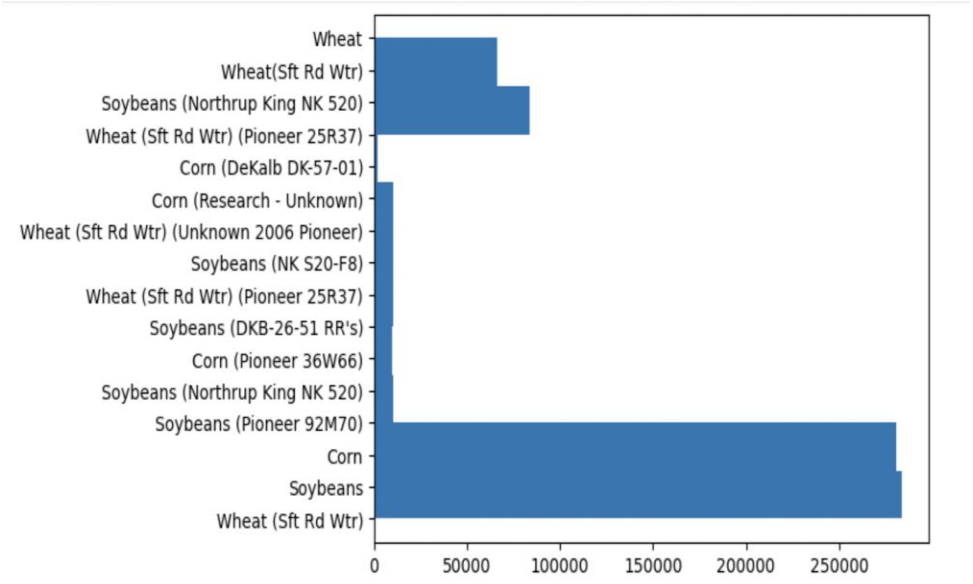
Weather data issue observed with temp values below zero Kelvin scale.

	timezone	lat	lon	temp	visibility	dew_point	feels_like	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	clouds_all	weather_id
count	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000	275168.000000
mean	-15804.804338	40.532353	-80.011232	283.997341	9060.337145	278.736042	282.486061	283.255011	284.856342	1016.406337	72.194605	3.165460	177.409452	67.632930	725.894145
std	1764.111901	0.423241	0.835482	10.672604	2095.044690	10.248443	12.330011	10.676350	10.759496	12.834197	18.983934	2.246839	113.310897	36.543175	125.957559
min	-18000.000000	0.000000	-80.019956	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	-18000.000000	40.536772	-80.019956	275.670000	9061.165963	271.070000	272.880000	274.960000	276.450000	1012.000000	59.000000	1.500000	80.000000	40.000000	701.000000
50%	-14400.000000	40.536772	-80.019956	284.695000	10000.000000	279.250000	283.830000	283.900000	285.490000	1017.000000	75.000000	3.100000	200.000000	75.000000	801.000000
75%	-14400.000000	40.536772	-80.019956	292.450000	10000.000000	287.170000	292.410000	291.740000	293.260000	1021.000000	88.000000	4.600000	270.000000	100.000000	803.000000
max	0.000000	40.536772	-80.019956	309.810000	10000.000000	302.260000	316.170000	309.930000	317.260000	1043.000000	100.000000	23.100000	360.000000	100.000000	804.000000

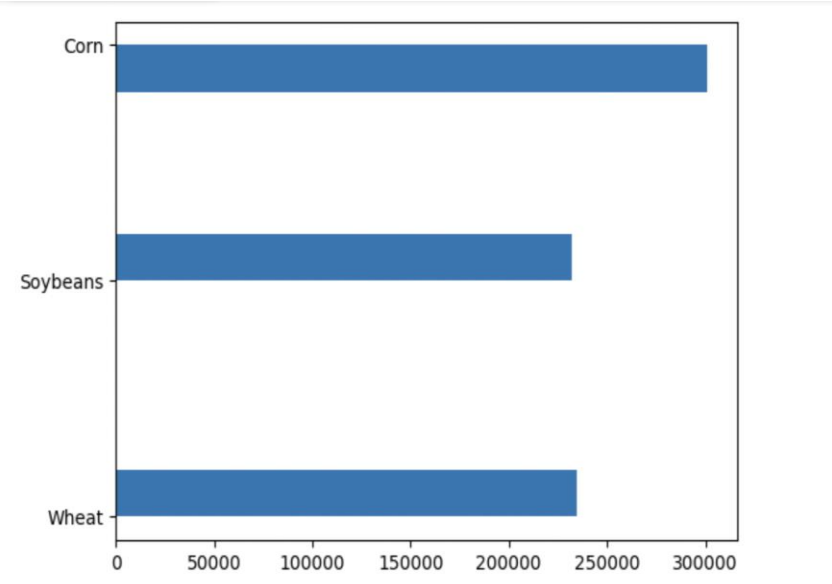
Negative values set to zero for consistency, and eliminating noisy data

Inconsistent Data Handling for Crop Yield Dataset

- The 'product' column containing inconsistent crop values.
- Handled by binning the values into consistent 'Corn', 'Soybeans', 'Wheat' categories.
- The cleaned 'product' column is visualized in a histogram.



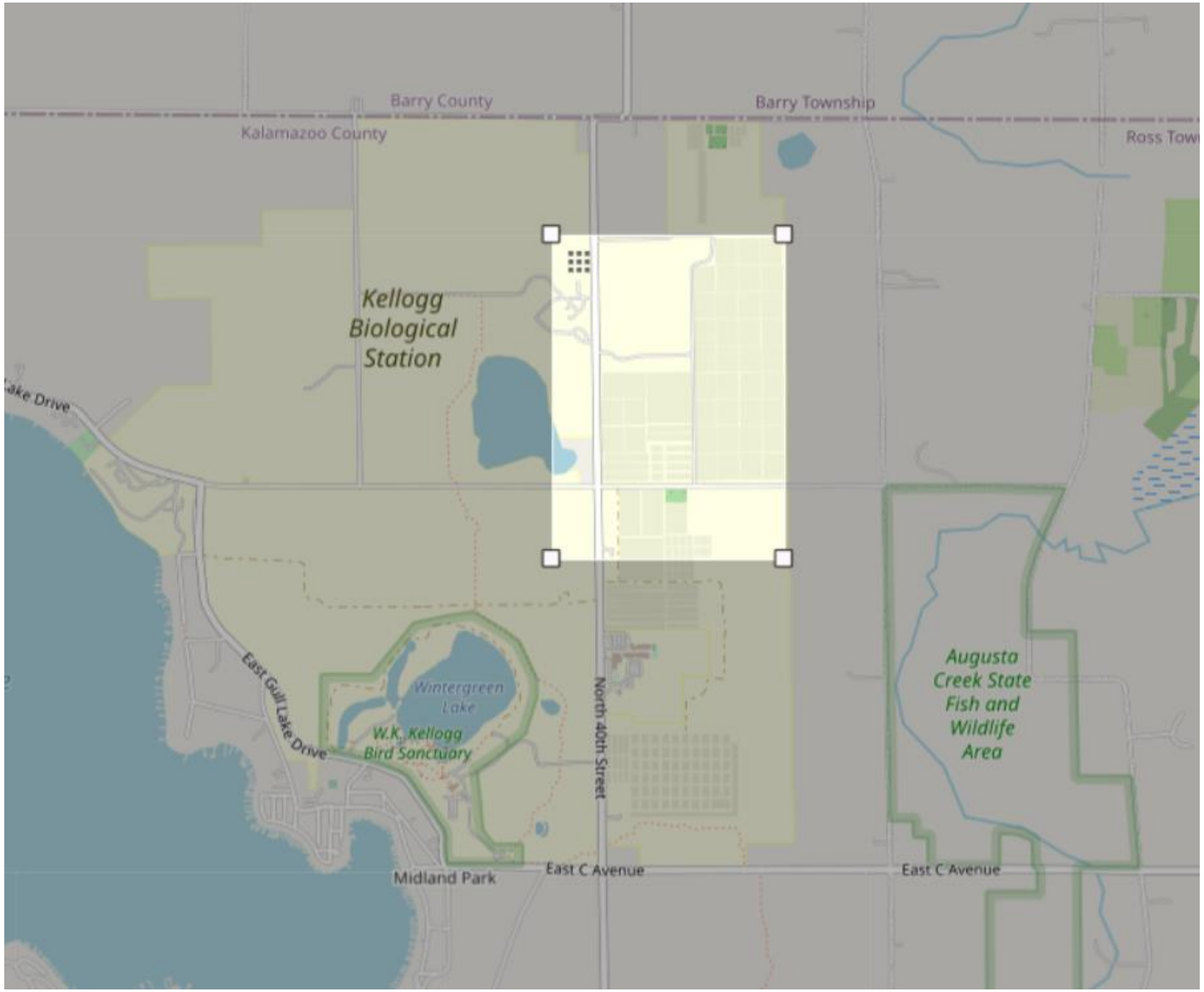
Before



After

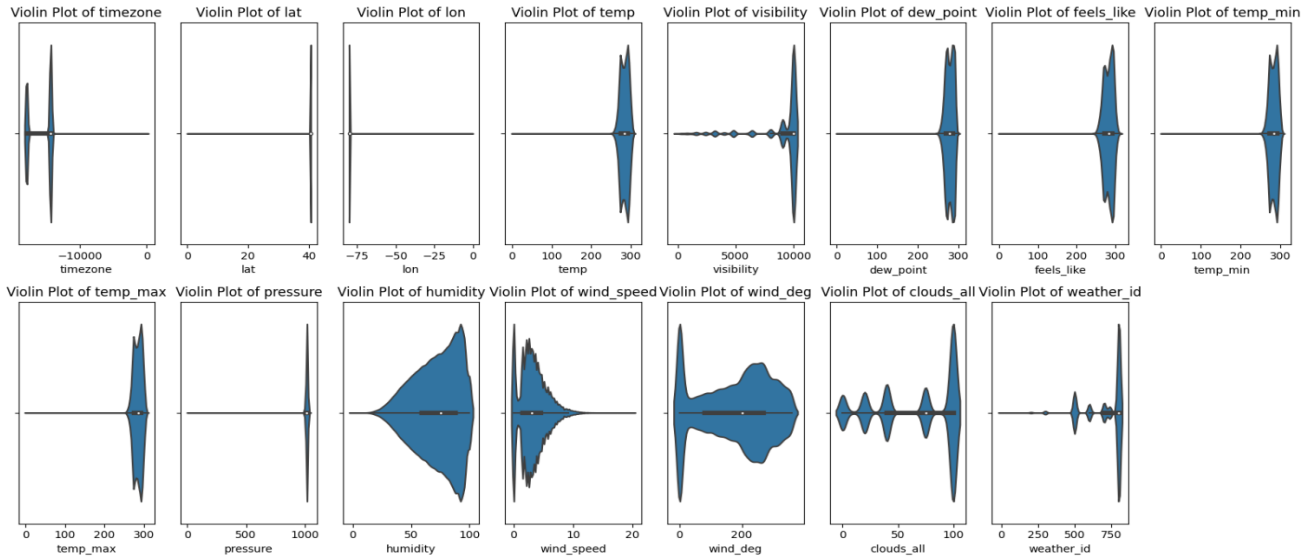
Inconsistent Data Handling for Location Dataset

The location data passed quality checks for data types and coordinates falling within the expected Kellog Biological Station area.



Inconsistent Data Handling for Weather Dataset

Weather dataset has no irregularities, and continuous variables exhibiting expected ranges and distributions in violin plots.



Weather dataset fields showed consistency with no data needing binning.

weather_id	weather_main	weather_description	weather_icon
804	Clouds	overcast clouds	04n
801	Clouds	few clouds	02n
800	Clear	sky is clear	01n
600	Snow	light snow	13d
721	Haze	haze	50n
520	Rain	light intensity s...	09n
601	Snow	snow	13d
701	Mist	mist	50n
620	Snow	light shower snow	13d
520	Rain	light intensity s...	09d
301	Drizzle	drizzle	09n
771	Squall	proximity squalls	50n
202	Thunderstorm	thunderstorm with...	11n
521	Rain	shower rain	09d
511	Rain	freezing rain	13d
771	Squall	squalls	50d
500	Rain	light rain	10d
521	Rain	shower rain	09n
620	Snow	light shower snow	13n
202	Thunderstorm	thunderstorm with...	11d

Data Aggregation to ensure same granularity level

Crop Yield data (second-level granularity)

longitude	latitude	crop_flow_lb_s	datetime	duration_	distance_in	swth_wdth_in	moisture	status	pass_num	serial_number	field	dataset	product	elevation_ft
-85.372824	42.408276	3.29	2013-07-24 19:20:04+00:00	1	56.0	180.0	6.6	0.0	141	5648	F1: Lysimeter	L1:	Wheat	953.5
-85.372838	42.408288	2.89	2013-07-24 19:20:03+00:00	1	47.0	180.0	6.6	0.0	141	5648	F1: Lysimeter	L1:	Wheat	953.5

Weather data (hour-level granularity)

dt_iso	timezone	city_name	lat	lon	temp	visibility	dew_point	feels_like	temp_min	...
1995-01-01 00:00:00 +0000 UTC	-18000	Ross Township	40.536772	-80.019956	278.57	1600.0	277.83	276.47	277.92	...
1995-01-01 01:00:00 +0000 UTC	-18000	Ross Township	40.536772	-80.019956	278.92	NaN	278.33	276.47	278.35	...

Crop yield data aggregated to hour level to match the granularity of the weather dataset

crop_flow_lb_s	datetime	moisture	field	product	elevation_ft
0.00	2009-10-16 14:00:00	16.70	F1: T2R2-West	Soybeans	997.42
4.80	2009-10-16 14:00:00	15.36	F1: T2R2-West	Soybeans	995.00

Merged dataset

crop_flow_lb_s	datetime	moisture	field	product	elevation_ft	name	dt_iso	city_name	temp	visibility	dew_point	feels_like	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	clouds_all
0.00	2009-10-16 14:00:00	16.70	F1: T2R2-West	Soybeans	997.42	Ross Township	2009-10-16 14:00:00	Ross Township	276.26	8047.0	274.47	274.21	275.95	276.45	1012	88	2.10	80	100
4.80	2009-10-16 14:00:00	15.36	F1: T2R2-West	Soybeans	995.00	Ross Township	2009-10-16 14:00:00	Ross Township	276.26	8047.0	274.47	274.21	275.95	276.45	1012	88	2.10	80	100
4.68	2009-10-16 14:00:00	15.65	F1: T2R5-West	Soybeans	1015.92	Ross Township	2009-10-16 14:00:00	Ross Township	276.26	8047.0	274.47	274.21	275.95	276.45	1012	88	2.10	80	100
3.98	2009-10-16 14:00:00	15.65	F1: T2R5-West	Soybeans	1016.25	Ross Township	2009-10-16 14:00:00	Ross Township	276.26	8047.0	274.47	274.21	275.95	276.45	1012	88	2.10	80	100
2.98	2009-10-16 14:00:00	15.55	F1: T2R5-West	Soybeans	1019.58	Ross Township	2009-10-16 14:00:00	Ross Township	276.26	8047.0	274.47	274.21	275.95	276.45	1012	88	2.10	80	100
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Merged **crop yield** and **weather** dataset on the datetime column

Data Standardization

Before:

crop_flow_lb_s	datetime	moisture	field	product	elevation_ft	name	dt_iso	city_name	temp	visibility	dew_point	feels_like	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	clouds_all
0.00	2009-10-16 14:00:00	16.70	F1: T2R2-West	Soybeans	997.42	Ross Township	2009-10-16 14:00:00	Ross Township	276.26	8047.0	274.47	274.21	275.95	276.45	1012	88	2.10	80	100
4.80	2009-10-16 14:00:00	15.36	F1: T2R2-West	Soybeans	995.00	Ross Township	2009-10-16 14:00:00	Ross Township	276.26	8047.0	274.47	274.21	275.95	276.45	1012	88	2.10	80	100

After:

	moisture	elevation_ft	temp	visibility	dew_point	feels_like	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	clouds_all
0	0.603556	0.949969	-1.590173	-0.908470	-1.349404	-1.605309	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179
1	0.323661	0.847196	-1.590173	-0.908470	-1.349404	-1.605309	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179
2	0.384235	1.735626	-1.590173	-0.908470	-1.349404	-1.605309	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179
3	0.384235	1.749640	-1.590173	-0.908470	-1.349404	-1.605309	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179
4	0.363348	1.891059	-1.590173	-0.908470	-1.349404	-1.605309	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179
...	...	...	...	...	...	...	...	...	...	...	...	...	...
350440	-0.467983	-0.840905	1.762612	0.333798	1.103366	1.764699	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626
350441	-0.467983	-0.840905	1.762612	0.333798	1.103366	1.764699	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626
350442	-0.467983	-0.840905	1.762612	0.333798	1.103366	1.764699	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626
350443	-0.467983	-0.840905	1.762612	0.333798	1.103366	1.764699	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626
350444	-0.467983	-0.826891	1.762612	0.333798	1.103366	1.764699	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626

350445 rows × 13 columns

Numerical data is standardized to a common scale using a Standard Scaler, which changes all numeric features to have a mean of zero and standard deviation of one.

One Hot Encoding

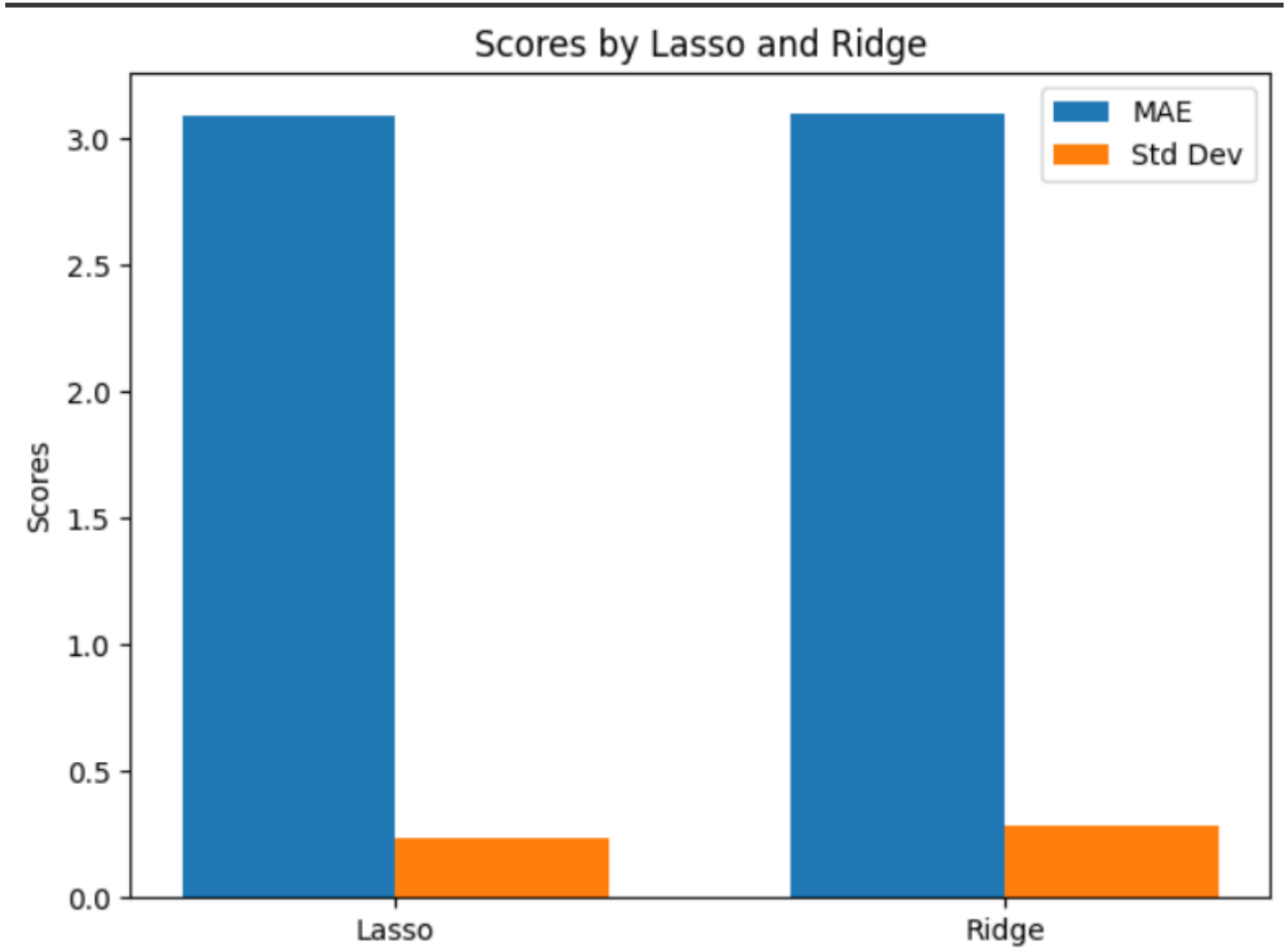
	crop_flow_lb_s	datetime	moisture	field	product	elevation_ft	name	dt_iso	city_name	temp	...	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	clouds_all	product_Corn	product_Soybeans	product_Wheat
0	0.00	2009-10-16 14:00:00	0.603556	F1: T2R2-West	Soybeans	0.949969	Ross Township	2009-10-16 14:00:00	Ross Township	-1.590173	...	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179	0	1	0
1	4.80	2009-10-16 14:00:00	0.323661	F1: T2R2-West	Soybeans	0.847196	Ross Township	2009-10-16 14:00:00	Ross Township	-1.590173	...	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179	0	1	0
2	4.68	2009-10-16 14:00:00	0.384235	F1: T2R5-West	Soybeans	1.735626	Ross Township	2009-10-16 14:00:00	Ross Township	-1.590173	...	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179	0	1	0
3	3.98	2009-10-16 14:00:00	0.384235	F1: T2R5-West	Soybeans	1.749640	Ross Township	2009-10-16 14:00:00	Ross Township	-1.590173	...	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179	0	1	0
4	2.98	2009-10-16 14:00:00	0.363348	F1: T2R5-West	Soybeans	1.891059	Ross Township	2009-10-16 14:00:00	Ross Township	-1.590173	...	-1.547585	-1.620763	-1.10452	1.067368	-0.373003	-0.722042	1.112179	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
350440	10.60	2007-07-09 18:00:00	-0.467983	F1: T2R2-East	Wheat	-0.840905	Ross Township	2007-07-09 18:00:00	Ross Township	1.762612	...	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626	0	0	1
350441	9.90	2007-07-09 18:00:00	-0.467983	F1: T2R2-East	Wheat	-0.840905	Ross Township	2007-07-09 18:00:00	Ross Township	1.762612	...	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626	0	0	1
350442	10.40	2007-07-09 18:00:00	-0.467983	F1: T2R2-East	Wheat	-0.840905	Ross Township	2007-07-09 18:00:00	Ross Township	1.762612	...	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626	0	0	1
350443	9.70	2007-07-09 18:00:00	-0.467983	F1: T2R2-East	Wheat	-0.840905	Ross Township	2007-07-09 18:00:00	Ross Township	1.762612	...	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626	0	0	1
350444	9.80	2007-07-09 18:00:00	-0.467983	F1: T2R2-East	Wheat	-0.826891	Ross Township	2007-07-09 18:00:00	Ross Township	1.762612	...	1.668514	1.695972	-0.74677	-1.549791	0.175205	0.513915	-0.580626	0	0	1

350445 rows × 23 columns

One hot encoding is performed on the categorical 'product' column to convert the crop types like Corn, Soybeans, and Wheat into numeric indicator columns that the machine learning algorithms can comprehend.

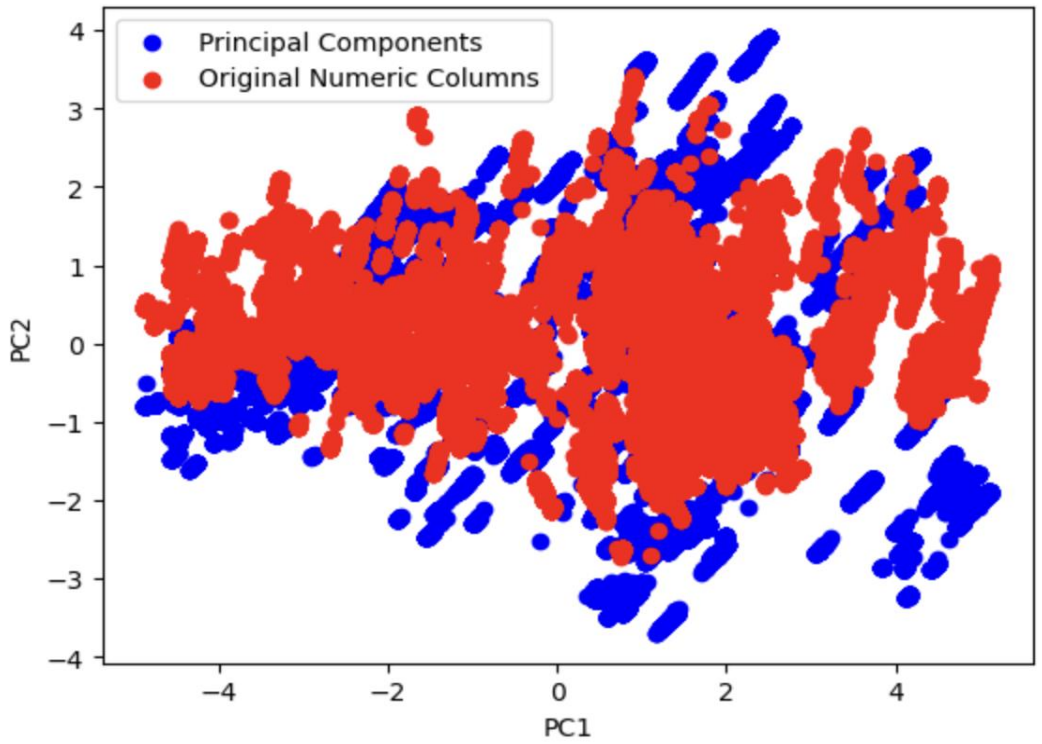
Data Regularization

- L1 regularization (Lasso) adds absolute coefficient penalties to shrink some coefficients to zero, while L2 regularization (Ridge) adds squared coefficient penalties to shrink coefficients toward but not exactly to zero.
- Lasso (L1) exhibits a slightly lower mean absolute error (MAE) of 3.086 compared to Ridge (L2) with a MAE of 3.102, suggesting a marginal performance advantage for Lasso, although the difference is relatively small.



Data Reduction using PCA

PCA, or Principal Component Analysis transforms high-dimensional data into a lower-dimensional space while preserving variance, by finding principal components which are linear combinations of the original features.



Transformed Dataset: After Dimensionality Reduction using PCA

	crop_flow_lb_s	datetime	field	city_name	product_Corn	product_Soybeans	product_Wheat	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
0	0.00	2009-10-16 14:00:00	F1: T2R2-West	Ross Township	0	1	0	3.781658	1.597206	-0.231569	0.327539	-0.422582	0.499634	-0.862528	-0.309784	0.121862
1	4.80	2009-10-16 14:00:00	F1: T2R2-West	Ross Township	0	1	0	3.713158	1.502763	-0.156346	0.495098	-0.623663	0.532004	-0.868154	-0.311718	0.115603
2	4.68	2009-10-16 14:00:00	F1: T2R5-West	Ross Township	0	1	0	3.872774	1.673078	-0.514877	0.147105	-0.545803	1.158249	-0.604078	-0.316033	-0.028574
3	3.98	2009-10-16 14:00:00	F1: T2R5-West	Ross Township	0	1	0	3.875117	1.675503	-0.520415	0.142061	-0.545247	1.168495	-0.599825	-0.316109	-0.030929
4	2.98	2009-10-16 14:00:00	F1: T2R5-West	Ross Township	0	1	0	3.894926	1.694252	-0.573709	0.100911	-0.554342	1.279905	-0.555002	-0.317068	-0.056445
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
350440	10.60	2007-07-09 18:00:00	F1: T2R2-East	Ross Township	0	0	1	-4.100974	-0.358484	0.172790	-0.145473	0.453026	0.124937	-0.427422	-0.429035	0.355361
350441	9.90	2007-07-09 18:00:00	F1: T2R2-East	Ross Township	0	0	1	-4.100974	-0.358484	0.172790	-0.145473	0.453026	0.124937	-0.427422	-0.429035	0.355361
350442	10.40	2007-07-09 18:00:00	F1: T2R2-East	Ross Township	0	0	1	-4.100974	-0.358484	0.172790	-0.145473	0.453026	0.124937	-0.427422	-0.429035	0.355361
350443	9.70	2007-07-09 18:00:00	F1: T2R2-East	Ross Township	0	0	1	-4.100974	-0.358484	0.172790	-0.145473	0.453026	0.124937	-0.427422	-0.429035	0.355361
350444	9.80	2007-07-09 18:00:00	F1: T2R2-East	Ross Township	0	0	1	-4.098632	-0.356059	0.167252	-0.150516	0.453582	0.135182	-0.423169	-0.429112	0.353006

350445 rows x 16 columns

Final version of Pre-processed Dataset

	crop_flow_lb_s	datetime	field	city_name	product_Corn	product_Soybeans	product_Wheat	PC1	PC2	PC3	PC4	PC5	PC6	PC7	month	year
0	0.00	2009-10-16 14:00:00	F1: T2R2-West	Ross Township	0	1	0	3.781658	1.597206	-0.231569	0.327539	-0.422582	0.499634	-0.862528	10	2009
1	4.80	2009-10-16 14:00:00	F1: T2R2-West	Ross Township	0	1	0	3.713158	1.502763	-0.156346	0.495098	-0.623663	0.532004	-0.868154	10	2009
2	4.68	2009-10-16 14:00:00	F1: T2R5-West	Ross Township	0	1	0	3.872774	1.673078	-0.514877	0.147105	-0.545803	1.158249	-0.604078	10	2009
3	3.98	2009-10-16 14:00:00	F1: T2R5-West	Ross Township	0	1	0	3.875117	1.675503	-0.520415	0.142061	-0.545247	1.168495	-0.599825	10	2009
4	2.98	2009-10-16 14:00:00	F1: T2R5-West	Ross Township	0	1	0	3.894926	1.694252	-0.573709	0.100911	-0.554342	1.279905	-0.555002	10	2009
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
350440	10.60	2007-07-09 18:00:00	F1: T2R2-East	Ross Township	0	0	1	-4.100974	-0.358484	0.172790	-0.145473	0.453026	0.124937	-0.427422	7	2007
350441	9.90	2007-07-09 18:00:00	F1: T2R2-East	Ross Township	0	0	1	-4.100974	-0.358484	0.172790	-0.145473	0.453026	0.124937	-0.427422	7	2007

All the numerical columns have been deduced to seven principal components resembling 95% of the variance. The product column is one-hot encoded into product_Corn, product_Soybeans and product_wheat. Regularization has been done to prevent overfitting of the model.

Test Dataset Preparation

The transformed dataset is split into 80% for training, 10% for validation, and 10% for final testing.

Prepared Datasets

```
Training dataset:    (280356, 15) (280356,)
Validation dataset:  (35044, 15) (35044,)
Test dataset:        (35045, 15) (35045,)
```

Sample Input Features

datetime	field	product_Corn	product_Soybeans	product_Wheat	PC1	PC2	PC3	PC4	PC5	PC6	PC7	month	year
1255701600	17	0	1	0	3.781657783	1.597205594	-0.2315690747	0.3275387478	-0.4225822826	0.4996341939	-0.8625280189	10	2009
1255701600	17	0	1	0	3.713157571	1.502763114	-0.1563461879	0.4950978877	-0.6236625416	0.5320040184	-0.8681535571	10	2009
1255701600	23	0	1	0	3.872773963	1.673078297	-0.5148770968	0.1471049395	-0.5458026675	1.158249204	-0.6040777938	10	2009
1255701600	23	0	1	0	3.875116611	1.675503207	-0.5204145282	0.1420613234	-0.5452470253	1.168494838	-0.5998248964	10	2009
1255701600	23	0	1	0	3.894926143	1.694251878	-0.5737090319	0.1009108772	-0.5543419976	1.279905318	-0.5550016555	10	2009
1255701600	23	0	1	0	3.894926143	1.694251878	-0.5737090319	0.1009108772	-0.5543419976	1.279905318	-0.5550016555	10	2009
1255701600	23	0	1	0	3.860350833	1.660218923	-0.4855119303	0.1738513886	-0.5487492546	1.103916299	-0.6266310375	10	2009
1255701600	17	0	1	0	3.778676231	1.594119344	-0.2245214348	0.3339578956	-0.4232894635	0.4865942966	-0.8679407974	10	2009
1255701600	17	0	1	0	3.735661185	1.526056951	-0.2095390896	0.4466486057	-0.6183250095	0.6304241963	-0.8272999671	10	2009

Sample Target Features

```
321199      6.74
123683      3.48
306520      3.70
255744      7.19
79336       4.53
            ...
233239     12.28
53716       1.61
115659     13.86
90388      13.12
179040      4.76
Name: crop_flow_lb_s, Length: 280356, dtype: float64
```

Model Selection and Updates

Machine Learning models for crop yield prediction:
LSTM (Long Short-Term Memory) Networks
ARIMA (Autoregressive Integrated Moving Average)
XGBoost Regressor
LightGBM (Light Gradient-Boosting Machine)

Models are trained and hyperparameters are tuned on validation data. Final model selection based on the evaluation metrics like RMSE, MSE, MAE, and R2 score.
Models can also be updated periodically as new yield data becomes available.

Model Comparison and Justification

Model	Strengths	Considerations	Justification
Long Short-Term Memory (LSTM) Networks	Well-suited for sequential data and time-series forecasting and can capture complex patterns	Good for predicting sequential time-series data	Chosen for its ability to capture complex temporal patterns in crop yield data.
XGBoost Regressor	Strong predictive performance.	Good ensemble techniques and efficient tree construction strategies.	Included for its robustness and accuracy in handling complex relationships.
LightGBM	Efficient and scalable for large datasets.	Efficient to be used on tabular data with many relationships and is a good ensemble technique	Selected for its efficiency, scalability, and ability to handle categorical features.
ARIMA (Autoregressive Integrated Moving Average)	Well-established for time-series forecasting.	Good for predicting sequential time-series data and recognizing seasonality pattern as well.	Included for simplicity and interpretability, especially with straightforward data.

Conclusion

Best Model

- Based on the test results, it was concluded that the LSTM model is the best-performing model with a 0.78 R2 score, which was identified as the most significant metric for evaluation.
- The model also demonstrated its effectiveness with a MSE of 6.75, a MAE of 1.8, and a RMSE of 2.59.

Future Scope

- Combine the strengths of LSTM and ARIMA algorithms for hybrid model development.
- Improve predictive accuracy by addressing individual limitations of LSTM and ARIMA, leading to more robust predictions.
- Geographical diversity in data collection.

References

Aruvansh, N., Garg, S., Agrawal, A., & Agrawal, P. (2019). Crop Yield Prediction Using Machine Learning Algorithms. 2019 Fifth International Conference on Image Information Processing (ICIIP). https://doi.org/10.1109/iciip47207.2019.8985951

Babbar, N., Kumar, A., & Verma, V. K. (2023). Forecasting Wheat Yield Using Long Short-Term Memory Considering Soil and Metrological Parameters. 2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT). https://doi.org/10.1109/icct56969.2023.10076090

Begum, M. B., Sivakannu, G., Eindhumathy, J., Priya, J. S., Mahendran, M., & Kumar, R. R. (2023). Enhancing agricultural productivity with data-driven crop recommendations. 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS). https://doi.org/10.1109/icaiss58487.2023.10250657

Behera, S., Menon, D., Shenoy, G. V., & Suresh, J. (2023). Suggestion of Appropriate Crops Based on Rainfall and Underground Water Analysis. 2023 3rd International Conference on Intelligent Technologies (CONIT). https://doi.org/10.1109/conit59222.2023.10205821

Bramantoro, A., Suhaili, W. S., & Siau, N. Z. (2022). Precision Agriculture Through Weather Forecasting. 2022 International Conference on Digital Transformation and Intelligence (ICDI). https://doi.org/10.1109/icdi57181.2022.10007299

Choudhary, N. K., Chukkapalli, S. S. L., Mittal, S., Gupta, M., Abdelsalam, M., & Joshi, A. (2020). YieldPredict: A Crop Yield Prediction Framework for Smart Farms. 2020 IEEE International Conference on Big Data (Big Data). https://doi.org/10.1109/bigdata50022.2020.9377832

Dwivedi, S. A., Attry, A., Parekh, D., & Singla, K. (2021). Analysis and forecasting of Time-Series data using S-ARIMA, CNN and LSTM. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). https://doi.org/10.1109/icccis51004.2021.9397134

Garg, S., Pundit, P., Jindal, H., Saini, H., & Garg, S. (2021). Towards a Multimodal System for Precision Agriculture using IoT and Machine Learning. 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt51525.2021.9579646

Geetha, M., Suganthe, R. C., Nivetha, S. K., Anju, R., Anuradha, R., & Haripriya, J. (2022). A Time-Series Based Yield Forecasting Model Using Stacked Lstm To Predict The Yield Of Paddy In Cauvery Delta Zone In Tamilnadu. 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). https://doi.org/10.1109/iceeict53079.2022.9768441

Haque, K., Islam, M. K., & Sattar, A. (2022). Wheat Production Forecasting in Bangladesh Using Deep Learning Techniques. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt54827.2022.9984464

Indira, D. N. V. S. L. S., Sobhana, M., Swaroop, A. H. L., & Kumar, V. P. (2022). KRISHI RAKSHAN - A Machine Learning based New Recommendation System to the Farmer. 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). https://doi.org/10.1109/iciccs53718.2022.9788221

Islam, A., Khair, I., Hossain, S., Ifty, R. A., Arefin, M. N., & Patwary, M. J. A. (2023). Ensemble Machine Learning Approach For Agricultural Crop Selection. 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE). https://doi.org/10.1109/ecce57851.2023.10101585

Kandan, M., Niharika, G., Lakshmi, M. J., Manikanta, K., & Bhavith, K. (2021). Implementation of Crop Yield Forecasting System based on Climatic and Agricultural Parameters. 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT). https://doi.org/10.1109/icissgt52025.2021.00051

Krishna, V., Reddy, T., Harsha, S., Ramar, K., Hariharan, S., & Bhanuprasad, A. (2022). Analysis of Crop Yield Prediction using Machine Learning algorithms. 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT). https://doi.org/10.1109/cisct55310.2022.10046581

Mariadass, D. A., Moung, E. G., Sufian, M. M., & Farzamnia, A. (2022). Extreme Gradient Boosting (XGBoost) Regressor and Shapley Additive Explanation for Crop Yield Prediction in Agriculture. 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE). https://doi.org/10.1109/iccke57176.2022.9960069

Mateo-Sanchis, A., Adsuara, J. E., Piles, M., Muñoz-Marí, J., Pérez-Suay, A., & Camps-Valls, G. (2023). Interpretable Long Short-Term Memory Networks for crop yield estimation. IEEE Geoscience and Remote Sensing Letters, 20, 1–5. https://doi.org/10.1109/lgrs.2023.3244064

Mehrmolaei, S., & Keyvanpour, M. R. (2016). Time series forecasting using improved ARIMA. 2016 Artificial Intelligence and Robotics (IRANOPEN). https://doi.org/10.1109/rios.2016.7529496

OpenWeatherMap.org. (n.d.-a). Geocoding API - OpenWeatherMap. Retrieved October 3, 2023, from https://openweathermap.org/api/geocoding-api

OpenWeatherMap.org. (n.d.-b). Historical weather API - OpenWeatherMap. Retrieved October 3, 2023, from https://openweathermap.org/history

Rai, S., Nandre, J., & Kanawade, B. (2022). A Comparative Analysis of Crop Yield Prediction using Regression. 2022 2nd International Conference on Intelligent Technologies (CONIT). https://doi.org/10.1109/conit55038.2022.9847783

Ranjani, J., Kalaiselvi, V., Sheela, A., Deepika, D., & Janaki, G. (2021). Crop yield prediction using a machine learning algorithm. 2021 4th International Conference on Computing and Communications Technologies (ICCCT). https://doi.org/10.1109/icoct53315.2021.9711853

Robertson, G. (2016). Precision Agriculture Yield Monitoring in Row Crop Agriculture at the Kellogg Biological Station, Hickory Corners, MI (1996 to 2013) ver 23. Environmental Data Initiative. https://doi.org/10.6073/pasta/423c07d6ea3317c545beabb4b8e502c8 (Accessed 2023-09-16).

Saini, P., & Nagpal, B. (2022). Deep-LSTM model for wheat crop yield Prediction in India. 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT). https://doi.org/10.1109/ccict56684.2022.00025

Seireg, H. R., Omar, Y. M. K., El-Samie, F. E. A., El-Fishawy, A. S., & Elmahalawy, A. (2022). Ensemble machine learning techniques using computer simulation data for wild Blueberry yield prediction. IEEE Access, 10, 64671–64687. https://doi.org/10.1109/access.2022.3181970

Shuvessa, S. K., Ryan, A. A., Mamun, S., Nabi, N., & Ahamed, M. S. (2022). An Approach using Machine Learning to Determine Bangladeshi Jute Yield relying on Weather Patterns. 2022 32nd International Conference on Computer Theory and Applications (ICCTA). https://doi.org/10.1109/iccta58027.2022.10206263

Sneha, V., & Bhavana, V. (2023). Sugarcane Yield and Price Prediction Using Forecasting Models. 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF). https://doi.org/10.1109/iceconf57129.2023.10084094

Venkatesh, A., & Saravanan, M. (2022). An Efficient Method for Predicting Linear Regression with Polynomial Regression. 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC). https://doi.org/10.1109/icosec54921.2022.9952049

Thank you !!