



Research Paper

TURNING FIELD, SOIL, WEATHER AND CROP DATA INTO A PREDICTIVE CULTIVATION PLANNING PIPELINE FOR SMARTER AGRICULTURAL DECISIONS - (FARM2FORECAST)

¹ Mr. M. Harish, ² Sriram Ganesh, ³ Kunduru Mahesh Reddy, ⁴ S Ramcharan Teja, ⁵ Lingaswamy M,
⁶ Rajaboina Arun

¹Assistant Professor, ²³⁴⁵⁶Students

Department of CSE (Data Science)

Siddhartha Institute of Technology & Sciences, Narapally

m.harish@siddhartha.org.in, 24tq1a6742@siddhartha.co.in, 24tq1a6727@siddhartha.co.in,
24tq1a6740@siddhartha.co.in, 24tq1a6728@siddhartha.co.in, 24tq1a6762@siddhartha.co.in

ABSTRACT

Farmers in India make most of their cultivation decisions, such as which crop to sow, when to sow it, how much fertiliser to apply, and when to irrigate, on the basis of experience and advice passed on informally. These decisions are becoming harder because rainfall has grown less predictable, soil fertility varies considerably between neighbouring fields, and market prices change from season to season. This paper presents Farm2Forecast, a data pipeline and predictive analytics system that combines field records, soil test results, weather observations and forecasts, and historical crop data to support cultivation planning for small and medium farms.

The system collects data from several sources. Field details such as location, area, irrigation source, and past crops are registered by farmers or extension workers through a mobile application. Soil Health Card results provide nitrogen, phosphorus, potassium, pH, organic carbon, and micronutrient values. Daily weather observations and medium-range forecasts are fetched from public meteorological services, and district-level crop yield and market price records are loaded from government open data portals. A batch and incremental ETL pipeline cleans these data, aligns them to a common field and season key, and stores them in a partitioned analytical warehouse.

Three models are built on the integrated data. A crop suitability model ranks candidate crops for each field and season based on soil, climate, and water availability. A yield prediction model, using gradient boosting on soil, weather, and management features, estimates the expected yield of the recommended crops. A sowing window model combines forecast rainfall and soil moisture estimates to suggest the most favourable dates for sowing. A rule-based nutrient module converts soil test values into fertiliser recommendations, and an economic layer uses expected yield and recent prices to estimate the likely return from each option.

I. INTRODUCTION

Agriculture supports a large share of India's population, and most farms are small, with holdings of less than two hectares. On such farms the choice of crop and the timing of operations largely determine the household's income for the year. A wrong

choice of crop for the soil, a sowing date that falls just before a long dry spell, or fertiliser applied without knowledge of existing soil nutrients can reduce yield substantially. Because the margin between profit and loss is narrow, better planning at the start of each season can make a real difference to farm incomes.

A great deal of relevant data is already being produced. Soil Health Cards have been issued to millions of farmers, meteorological departments publish daily observations and forecasts, and agriculture departments record crop area, production, and yield for each district every season. Market committees report prices of major crops daily. In practice, however, these datasets are held by different agencies, stored in different formats, and published at different levels of detail. A farmer who wants to use them would need to collect a paper soil card, check a weather website, and look up price reports separately, and would still have no way of combining them into a single decision.

Data engineering and machine learning can help by integrating these sources into one consistent dataset organised by field and season. Once soil, weather, and crop history are aligned, models can learn how yield responds to rainfall, temperature, and nutrient levels in a particular region, and can use forecasts to estimate what is likely to happen in the coming season. The value of such models, however, depends on a pipeline that handles missing soil tests, gaps in weather records, changes in district boundaries, and inconsistent crop names, all of which are common in agricultural data.

This paper describes Farm2Forecast, a predictive cultivation planning pipeline that turns field, soil, weather, and crop data into a season plan for each farm. The emphasis is on the design of the ingestion and transformation layers, the construction of a field-season analytical dataset, and the models used for crop suitability, yield prediction, and sowing window estimation. The remaining sections review related studies, analyse the current and proposed systems, and present the methodology and architecture of the system.

II. LITERATURE SURVEY

Crop yield prediction has long been studied using statistical models that relate yield to weather variables. Early work used multiple linear regression on seasonal rainfall and temperature aggregated over the growing period, and crop simulation models were developed to represent plant growth as a function of daily weather, soil water, and management. Studies with simulation models report good results when detailed calibration data is available, but they also note that such data is rarely available for smallholder fields. Regression approaches are easier to apply at district level, yet researchers have found that seasonal totals hide the effect of dry spells and heat events during sensitive stages such as flowering. This has led to the use of weekly or stage-wise weather features, which capture the timing of stress and generally improve prediction accuracy over simple seasonal aggregates.

Machine learning methods including random forests, support vector regression, and gradient boosting have been applied to yield prediction with soil, weather, and management inputs. Comparative studies across crops such as rice, wheat, maize, and cotton report that ensemble tree methods usually outperform linear regression because they capture non-linear responses, for example the sharp drop in yield when

temperature exceeds a threshold during grain filling. Researchers have also combined remote sensing indices such as NDVI with weather data to improve in-season prediction. A common observation in this body of work is that models trained on one region do not transfer well to another with different soils and cropping practices, and that the quality and consistency of input data matter as much as the choice of algorithm.

Crop recommendation systems have been developed to suggest suitable crops given soil nutrients, pH, rainfall, and temperature. Many of these systems are trained on public datasets that pair soil and climate conditions with crop labels and use classifiers such as decision trees, naive Bayes, or neural networks. Studies report high classification accuracy on such datasets, but authors point out that the labels often reflect what was grown rather than what would have been most profitable or productive. More practical approaches combine suitability with expected yield and market prices, so that recommendations reflect the likely economic outcome and not only agronomic feasibility. Research also stresses the need to consider water availability and irrigation source, since a crop that suits the soil may still fail without sufficient water.

Work on agricultural data infrastructure has examined the difficulties of integrating data from multiple agencies. Studies of soil testing programmes report that soil card data is often recorded without precise field coordinates, which makes linking it to weather grids and field records difficult. Research on weather data for agriculture discusses gridded datasets and interpolation methods for filling gaps between stations. Several projects have built agricultural data warehouses that organise information by location and season and have reported that standardising crop names, units, and administrative codes is one of the most time-consuming tasks. Authors recommend automated validation and a clear lineage from each derived value back to its source so that advisories can be traced and corrected when errors are found.

Research on sowing date optimisation and agro-advisory services has shown that timely advice can reduce losses from weather variability. Studies of forecast-based advisories report that farmers who adjusted sowing dates using onset of monsoon forecasts and soil moisture information achieved better germination and fewer re-sowing events. Other work on fertiliser recommendation based on soil test values has demonstrated savings in input costs without loss of yield. Studies of digital advisory adoption indicate that advice in the local language, delivered through a mobile phone and supported by extension workers, is more readily followed than generic bulletins. The present work brings these strands together by building an integrated field-season pipeline that supports crop choice, yield expectation, sowing timing, and nutrient planning within one system.

III. SYSTEM ANALYSIS

The system is analysed as a pipeline that converts heterogeneous agricultural data into a field-level season plan. The key entity is the field, identified by its geographic coordinates and linked to a farmer, a village, and a district. Field registration data, including area, irrigation source, and crop history, is captured through a mobile application. Soil Health Card values are matched to fields by survey number or nearest sampling point. Daily weather observations are taken from the nearest stations

and gridded datasets, and seven to fifteen day forecasts along with seasonal outlooks are fetched through public APIs. District crop statistics and daily market prices are downloaded from open data portals. Because these sources differ in spatial resolution, time step, and naming, the pipeline assigns every record a field and season key, interpolates weather to field locations, standardises crop names, and fills missing soil values using nearby samples. The integrated data feeds the suitability, yield, and sowing models, and their outputs are combined with nutrient rules and prices to produce the plan.

Existing System

Cultivation planning today depends mostly on the farmer's own experience, the practices of neighbours, and advice from input dealers who may have an interest in selling particular seeds or fertilisers. Soil Health Cards are distributed on paper, and many farmers find the values difficult to interpret or misplace the card before the next season. Weather information reaches farmers through general bulletins for the whole district, which do not reflect local conditions on a particular field. Agriculture department advisories are issued in broad terms for a crop or region rather than for an individual farm. Crop yield and price data are published in reports that are rarely consulted at the village level. As a result, farmers often continue to grow the same crop regardless of soil condition or expected rainfall, apply fertilisers based on habit, and choose sowing dates without reference to forecasts, which increases the risk of poor yields and financial loss.

Disadvantages of Existing System

- Crop choice depends mainly on habit and informal advice.
- Soil test values on paper cards are hard to interpret and often lost.
- District-wide weather bulletins do not reflect conditions on each field.
- Sowing dates are chosen without using rainfall and soil moisture forecasts.
- Fertiliser is applied by habit rather than according to soil nutrient levels.
- Soil, weather, yield, and price data are held by separate agencies.
- Crop names, units, and location codes are inconsistent across datasets.
- Advisories are generic and not tailored to individual farms.
- Expected yield and economic return of alternatives are never compared.
- Input dealers may give biased recommendations for seeds and fertilisers.

Proposed System

The proposed system, Farm2Forecast, is an integrated data pipeline and predictive planning platform for cultivation decisions. Ingestion jobs collect field registrations from the mobile application, Soil Health Card records, station and gridded weather data, forecast feeds, district crop statistics, and market prices, writing them to a raw zone in object storage. A transformation layer built on Apache Spark validates and standardises the records, geocodes fields, interpolates weather to field coordinates, fills missing soil values with spatial methods, and builds a curated field-season table in a partitioned warehouse. A feature store computes stage-wise rainfall, growing degree days, dry spell counts, heat stress days, soil nutrient indices, and crop history. A crop suitability ranker, a gradient boosting yield model, and a sowing window model generate predictions, and a nutrient rule engine and price-based return

estimator complete the plan. Results are delivered through a mobile application in the local language and an extension officer dashboard, and are refreshed as new forecasts arrive during the season.

Advantages of Proposed System

- Combines soil, weather, crop history, and price data for each field.
- Recommends crops ranked by suitability, expected yield, and return.
- Predicts yield for each option with measured accuracy.
- Suggests sowing windows using rainfall forecasts and soil moisture.
- Converts soil test values into clear fertiliser recommendations.
- Standardises crop names, units, and locations across agencies.
- Updates advice during the season as new forecasts arrive.
- Delivers plans in the local language through a simple mobile app.
- Gives extension officers a village-level view for planning visits.
- Works with existing public data without costly field sensors.

IV. METHODOLOGY

The methodology begins by defining the field-season as the unit of analysis and designing a schema that links fields, soils, weather, crops, and prices through shared keys. Ingestion jobs run on a schedule, with weather and prices collected daily and soil and crop statistics loaded when new releases appear. In the transformation stage, records are validated for range and completeness, crop names are mapped to a standard list, and fields are geocoded. Weather is interpolated to field locations using inverse distance weighting, and missing soil values are estimated by kriging from nearby samples. Features are computed for each crop stage, including cumulative rainfall, dry spell length, growing degree days, and maximum temperature days above thirty-five degrees. The yield model is a gradient boosting regressor trained on four seasons using leave-one-season-out validation, and the suitability model is a ranking model trained on observed yields. The sowing model scores candidate dates using forecast rainfall and a soil water balance. Outputs are evaluated with MAPE, R squared, and top-three recommendation accuracy.

SYSTEM ARCHITECTURE

The architecture is organised as a medallion data lakehouse with a prediction and delivery layer on top. The Data Sources zone contains the farmer mobile application, Soil Health Card records, weather station and forecast APIs, district crop statistics, and market price feeds. The Ingestion zone uses scheduled connectors and an orchestrator to land raw files in the bronze layer of object storage. The silver layer holds cleaned and standardised records after validation, geocoding, name mapping, weather interpolation, and soil gap filling. The gold layer contains the curated field-season table and the feature store. The Modelling zone includes the crop suitability ranker, the yield regressor, and the sowing window model, with a model registry that tracks versions and evaluation results. The Planning Engine combines model outputs with the nutrient rule engine and the return estimator to produce a season plan. The Delivery zone presents plans on the farmer app and on the extension officer dashboard, and observed yields at harvest are fed back for retraining.

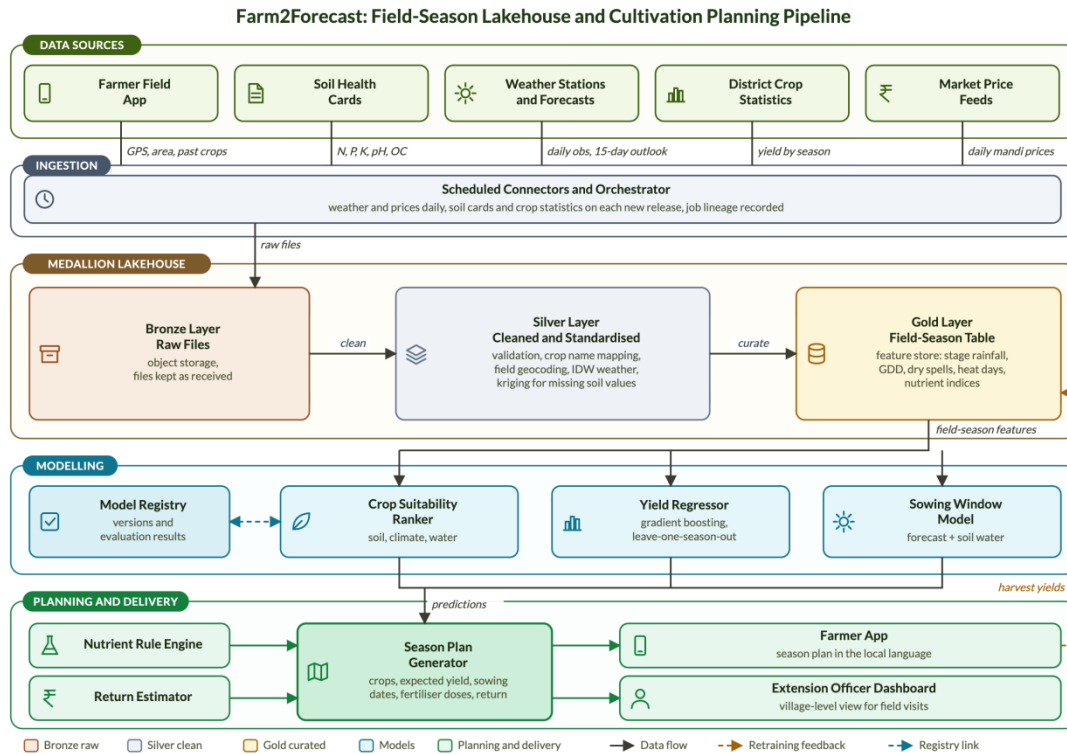


Fig. 1 System Architecture

V. RESULT AND OUTPUT

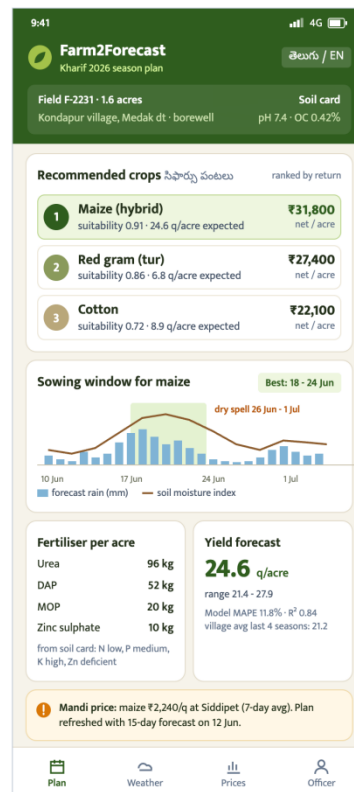


Fig. 2 Result and Output Screen

VI. CONCLUSION

This paper presented Farm2Forecast, a predictive cultivation planning pipeline that combines field, soil, weather, and crop data into a season plan for individual farms. The work addresses a practical problem: the data needed for better planning is already collected by various agencies, but it is fragmented, inconsistent, and difficult for farmers to use. By organising all sources around a common field-season key, the system makes this data usable for analysis and prediction.

Most of the engineering effort went into the pipeline. Geocoding fields, interpolating weather to field locations, filling gaps in soil data, and standardising crop names were necessary steps before any model could be trained. The medallion structure of bronze, silver, and gold layers made it possible to trace every feature back to its source and to reprocess data when errors were discovered. On four seasons of data from 2,350 fields, the yield model reached a MAPE of 11.8 percent and an R squared of 0.84, and the top-three crop ranking matched the best-performing crop in 87 percent of held-out cases.

The season plan turns these predictions into decisions that a farmer can act on. Recommending crops with expected yield and estimated return, suggesting a sowing window based on the forecast, and converting soil test values into fertiliser doses give the farmer a clear basis for planning. Presenting the plan in the local language and through extension officers helps ensure that the advice reaches those who need it, and the officer dashboard supports planning at the village level.

Future work will add satellite vegetation indices for in-season monitoring of crop condition, pest and disease risk models driven by weather, and water budgeting for irrigated fields. Linking season plans with input supply, crop insurance, and credit services could help farmers act on recommendations, and a longer study across more districts will be needed to measure the effect of the system on yields and incomes.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satyakrishna, shirisha rangu, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

- [6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
- [7] R. D. Kumar, V. N. S. Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in *Blockchain for Smart Cities*, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
- [8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
- [9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” *Journal of Machine and Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
- [10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in *Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE)*, Apr. 2025.
- [11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” *SN Computer Science*, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.