

MACHINE LEARNING BASED CROP PRICE PREDICTION AND MARKET TREND ANALYSIS USING PYTHON

¹ Dr.V.Vijay Bhaskar, ² Kadudoori Abhishek Reddy, ³ Botla Vishnuvardhan, ⁴ Kandikatla Srihari, ⁵ Chinimilla Umavathi

¹Assistant Professor, ²³⁴⁵Students

Department of ECE

Siddhartha Institute of Technology & Sciences, Narapally

vijay.valiki@siddhartha.co.in, 24tq1a0420@siddhartha.co.in, 24tq1a0404@siddhartha.co.in,

24tq1a0421@siddhartha.co.in, 24tq1a0409@siddhartha.co.in

ABSTRACT

Agricultural commodity prices in Indian markets change from day to day and from one market yard to another, and these changes directly affect the income of farmers, the cost of food for consumers, and the planning of traders and storage operators. A farmer who sells tomatoes or onions on the wrong day can receive a price far below what was available a week earlier or later. This paper presents a machine learning based crop price prediction and market trend analysis system developed in Python that forecasts short-term modal prices for selected crops and presents market trends in a form that can be understood without technical training.

The system uses historical daily arrival and price records published for regulated agricultural markets, covering the minimum, maximum, and modal price along with the quantity of produce that arrived on each day. These records are combined with rainfall and temperature data for the growing districts and with a calendar of festivals and harvest seasons. The raw data contains missing days, market holidays, spelling differences in crop and variety names, and occasional entry errors, so a preprocessing stage standardises names, removes impossible values, fills short gaps, and aligns all sources to a common daily index.

Feature engineering converts the cleaned data into inputs that describe recent market behaviour. Lagged prices, rolling averages over one, two, and four weeks, arrival volume and its change, month and week of the year, and rainfall accumulated over the growing period are computed for each crop and market. Several regression models are then trained and compared, including Linear Regression, Random Forest, Gradient Boosting, XGBoost, and a seasonal ARIMA baseline. The models are evaluated with a time-ordered split so that the test period always lies after the training period, and performance is measured using RMSE, MAE, MAPE, and the coefficient of determination.

The results show that tree-based ensemble models, particularly Gradient Boosting and XGBoost, predict next-week modal prices more accurately than the linear and ARIMA baselines, with arrival quantity and recent price lags being the most influential features. A web dashboard built with Streamlit allows the user to select a crop and market, view past prices with seasonal patterns highlighted, see the forecast for the coming days with an uncertainty band, and compare prices across nearby markets to identify where selling may be more profitable.

The system is intended to help farmers, farmer producer organisations, and small traders make better-informed decisions about when and where to sell their produce. It does not claim to remove uncertainty from agricultural markets, which are affected by policy decisions, transport disruptions, and sudden weather events, but it provides a consistent and data-driven view of likely price movement. Future work includes adding more crops and markets, using satellite-based crop condition indices, and delivering forecasts through SMS and voice messages in regional languages.

I. INTRODUCTION

Agriculture supports a large share of the population, and the price a farmer receives at the market decides whether a season ends in profit or loss. Prices of perishable crops such as tomato, onion, and potato are known to swing sharply within a few weeks. A good harvest across many districts can flood the markets and push prices down, while unseasonal rain or transport problems can reduce arrivals and cause prices to rise suddenly. Farmers usually have limited information about these movements and often sell soon after harvest because they lack storage or cannot predict whether waiting will be worthwhile.

Price information for regulated markets is published daily through government portals, but it appears in the form of large tables that are difficult for an individual farmer to interpret. A table showing today's minimum, maximum, and modal price does not reveal whether prices are rising or falling, how the current level compares with the same period last year, or whether a neighbouring market is paying more. Traders with experience and access to many markets can read these patterns, which places small farmers at a disadvantage when negotiating a sale. Decisions about sowing, harvesting, and storing are therefore frequently made on the basis of word of mouth rather than on evidence.

Machine learning can learn relationships between past prices, arrival quantities, weather, and seasonal factors, and use them to estimate prices in the near future. Python provides mature libraries for this work, including pandas for data handling, scikit-learn and XGBoost for modelling, statsmodels for classical time-series methods, and Streamlit for building interactive dashboards. With these tools it is possible to build a complete pipeline, from data collection to forecast display, within the scope of an undergraduate project while still following sound practices such as time-ordered validation and comparison against simple baselines.

This paper describes a crop price prediction and market trend analysis system that collects historical market data, prepares it through careful cleaning and feature engineering, trains and compares several forecasting models, and presents the results through a simple web dashboard. The system focuses on short-term forecasts of one to two weeks, which are the most useful for deciding when to sell, and on trend views that summarise seasonal patterns and differences between markets. The overall goal is to make market information more accessible and to support better timing of sales by farmers and small traders.

II. LITERATURE SURVEY

Forecasting of agricultural commodity prices has been studied for several decades, and early work relied mainly on classical statistical time-series methods. Autoregressive integrated moving average models and their seasonal extensions were applied to monthly and weekly price series of cereals, pulses, and vegetables, and were found to capture trend and seasonal cycles reasonably well for crops with stable production patterns. Exponential smoothing methods such as Holt-Winters were also used because they are simple to fit and interpret. Studies in this area noted, however, that these models assume the underlying relationships remain constant over time and that the series is approximately linear after differencing. Vegetable prices, in particular, show sudden jumps caused by weather events and supply shocks, which classical models tend to smooth over. Researchers also observed that univariate models cannot make use of useful external information such as arrival quantities, rainfall, or policy announcements, which limits their accuracy during unusual periods.

To overcome these limitations, many studies introduced machine learning regression models that can include several explanatory variables. Support vector regression, random forests, and gradient boosting machines were trained on lagged prices together with arrivals, weather variables, and calendar features. Comparative studies across different crops frequently reported that ensemble tree methods achieved lower error than both linear regression and ARIMA, especially for highly volatile vegetables. Feature importance analysis in these studies consistently placed recent price lags and market arrival volumes among the most useful predictors, confirming the basic supply-and-demand relationship in the data. Some authors pointed out that tree-based models do not extrapolate beyond the range of prices seen during training, so they may underestimate unprecedented price spikes, and recommended combining them with trend-aware methods or retraining frequently as new data arrives.

Deep learning approaches have also been applied to crop price forecasting. Recurrent neural networks, particularly Long Short-Term Memory networks, were used to model long sequences of daily prices and were reported to capture nonlinear temporal patterns better than classical methods in several studies. Hybrid models combining signal decomposition methods, such as wavelet transforms or empirical mode decomposition, with neural networks were proposed to separate the long-term trend from short-term fluctuations before modelling each component. While some of these studies reported impressive accuracy, others noted that deep models require long and clean histories, careful tuning, and considerable computing resources, and that their gains over well-tuned gradient boosting models were often small for short horizons. The difficulty of explaining deep model predictions to farmers and market officials was also mentioned as a practical barrier to adoption.

A significant part of the literature concerns data quality and validation practice. Market price records often contain missing days due to holidays, inconsistent naming of crop varieties, duplicate entries, and prices recorded in different units. Studies that paid attention to cleaning and aligning these records reported noticeably better model performance than those that used raw data. Several authors warned against random train-test splits for time-series problems, since they allow information from the future to leak into training and produce overly optimistic results. Walk-forward or rolling-origin validation, in which the model is repeatedly trained on past data and tested on

the following period, was recommended as a more honest estimate of real forecasting ability. Error measures such as RMSE, MAE, and MAPE were widely used, with MAPE preferred when comparing crops whose prices differ greatly in magnitude.

Recent work has emphasised the delivery of forecasts to end users through decision-support tools. Mobile applications, web dashboards, and SMS services have been developed to provide farmers with price information and short-term forecasts, and user studies indicate that clear visual presentation of trends and comparisons between markets is valued more than detailed numerical output. Some systems include recommendations about the best market to sell in after accounting for transport cost, and others show forecast ranges rather than single values to communicate uncertainty honestly. Researchers have also highlighted the need for regional language support and simple interfaces for users with limited digital literacy. The present system builds on these findings by combining careful data preparation, time-aware validation, comparison of statistical and machine learning models, and a dashboard that presents forecasts, uncertainty, and market comparisons in a simple form.

III. SYSTEM ANALYSIS

The system is analysed as a data pipeline that transforms published market records into short-term price forecasts and trend summaries. Daily records for selected crops and markets are downloaded from the public agricultural marketing portal, each containing the date, market, commodity, variety, arrival quantity, and minimum, maximum, and modal prices. District rainfall and temperature records and a calendar of festivals and harvest seasons are collected separately. The preprocessing module standardises commodity and variety names, converts all prices to rupees per quintal, removes records with impossible values, and aligns the data to a continuous daily index, filling short gaps and flagging market holidays. The feature module computes lagged prices, rolling means and standard deviations, arrival changes, seasonal indicators, and accumulated rainfall. Several regression and time-series models are trained using walk-forward validation, and their errors are compared. The selected model produces forecasts for the next seven to fourteen days, and the dashboard displays these forecasts together with historical trends, seasonal profiles, and comparisons between nearby markets.

Existing System

At present, farmers and small traders obtain price information mainly from other farmers, commission agents at the market yard, local newspapers, and occasionally from government portals and mobile applications that list the prices of the current day. This information describes only what has already happened and gives no indication of where prices are likely to move. Published tables are long and are organised by market and commodity in a way that makes comparisons across time and across markets difficult for a non-specialist. Some advisory services issue broad seasonal outlooks, but these are general statements that do not address a particular crop in a particular market over the coming week. Commission agents and larger traders, who observe many transactions, have a clear information advantage over individual farmers. As a result, farmers often sell immediately after harvest regardless of the

market situation, and decisions about storage, choice of market, and timing of sale are made largely on guesswork or advice from intermediaries whose interests may differ from their own.

Disadvantages of Existing System

- Available price information describes only the current day and gives no forecast.
- Published market tables are long and difficult for farmers to interpret.
- Comparison of prices across nearby markets must be done manually.
- Seasonal patterns and year-on-year changes are not presented clearly.
- Farmers depend on commission agents whose interests may differ from their own.
- Advisory outlooks are general and do not cover specific crops and markets.
- Arrival quantities and weather information are not linked to price movement.
- Decisions about storage and timing of sale are made largely on guesswork.
- Uncertainty in future prices is not communicated to the user in any form.
- Historical data is not reused to learn from past market behaviour.

Proposed System

The proposed system is a Python-based crop price prediction and market trend analysis platform that forecasts short-term modal prices and presents market behaviour through an interactive dashboard. Historical daily records of arrivals and prices for selected crops and markets are collected and combined with district rainfall, temperature, and a seasonal calendar. A preprocessing module standardises names and units, removes erroneous entries, fills short gaps, and aligns all sources to a common daily index. Lagged prices, rolling statistics, arrival changes, seasonal indicators, and accumulated rainfall are computed as features. Linear Regression, Random Forest, Gradient Boosting, XGBoost, and seasonal ARIMA models are trained and compared using walk-forward validation, with RMSE, MAE, MAPE, and R-squared as measures. The best model for each crop is used to forecast prices for the following one to two weeks, together with an uncertainty range. The Streamlit dashboard shows historical trends, seasonal profiles, forecasts, and price comparisons across nearby markets so users can decide when and where to sell.

Advantages of Proposed System

- Provides short-term price forecasts for specific crops and markets.
- Combines price history with arrival quantity, weather, and seasonal information.
- Compares several models and selects the most accurate one for each crop.
- Uses walk-forward validation to give a realistic estimate of forecast accuracy.
- Shows an uncertainty range along with each forecast value.
- Presents seasonal patterns and year-on-year trends in simple charts.
- Compares prices across nearby markets to help choose where to sell.
- Reduces dependence on intermediaries for market information.
- Built entirely with open-source Python libraries and public data.

- Can be extended easily to new crops, markets, and data sources.

IV. METHODOLOGY

The methodology begins with the selection of crops and markets for which long and reasonably complete price histories are available. Daily records covering several years are downloaded and merged with district-level rainfall and temperature data and a calendar of major festivals and harvest periods. Data cleaning standardises commodity and variety names, converts units, removes records where the minimum price exceeds the maximum or where values are far outside the normal range, and reindexes each series to a continuous daily calendar, interpolating short gaps. Features are created for each crop and market, including price lags of one, seven, and fourteen days, rolling means and standard deviations, percentage change in arrivals, week of year, month, and cumulative rainfall. The data is split in time order, and walk-forward validation is used to train and test each model over successive periods. Hyperparameters of the ensemble models are tuned using time-series cross-validation. RMSE, MAE, MAPE, and R-squared are computed, feature importance is examined, and the chosen model is saved and served through the dashboard.

SYSTEM ARCHITECTURE

The architecture of the system is organised into layers that lead from raw public data to forecasts displayed on the dashboard. The Data Collection Layer downloads daily market arrival and price records and gathers weather and calendar information. The Data Cleaning Layer standardises names and units, removes erroneous entries, and handles missing days and holidays. The Data Integration Layer aligns prices, arrivals, weather, and calendar data to a common daily index for each crop and market. The Feature Engineering Layer computes lags, rolling statistics, arrival changes, seasonal indicators, and rainfall features. The Model Training Layer fits Linear Regression, Random Forest, Gradient Boosting, XGBoost, and seasonal ARIMA models. The Evaluation and Selection Layer compares models using walk-forward validation and chooses the best one for each crop. The Forecasting Layer produces predictions with uncertainty ranges for the coming days. The Trend Analysis Layer computes seasonal profiles and market comparisons, and the Dashboard Layer presents forecasts, trends, and comparisons through an interactive web interface.

Machine learning based crop price prediction and market trend analysis using python



Fig. 1 System Architecture

V. RESULT AND OUTPUT



VI. CONCLUSION

This paper has presented a machine learning based crop price prediction and market trend analysis system developed in Python. The system collects public market records, combines them with weather and seasonal information, and converts them into short-term price forecasts and clear trend views. It addresses a practical problem faced by farmers and small traders, who often have to decide when and where to sell with very limited information about future price movement.

Careful data preparation was found to be essential. Standardising names and units, removing erroneous records, and aligning several data sources to a common daily index produced a much cleaner basis for modelling. Among the models compared, Gradient Boosting and XGBoost gave the lowest forecast errors for most crops, and recent price lags and arrival quantities emerged as the strongest predictors. The use of walk-forward validation ensured that the reported accuracy reflects how the models would perform on genuinely unseen future data.

The dashboard makes the results accessible to users who are not familiar with data analysis. It presents price history, seasonal patterns, forecasts with uncertainty ranges, and comparisons across nearby markets in a simple layout. By showing a range rather than a single number, the system communicates honestly that forecasts can be wrong, particularly when prices are affected by sudden weather events, policy changes, or transport disruptions that are not visible in the historical data.

Future work can extend the system to more crops and markets, include satellite-based vegetation indices and sowing area estimates as early indicators of supply, and account for transport cost when recommending a market. Delivering forecasts through SMS and voice messages in regional languages would allow farmers without smartphones to benefit, and regular automatic retraining would help the models adapt as market conditions change from season to season.

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.