Machine Learning Based Agricultural Profitability Recommendation Systems: A Paradigm Shift in Crop Cultivation

Nilesh P. Sable[1], Rajkumar V. Patil[2], Mahendra Deore[3], Ratnmala Bhimanpallewar[4], Parikshit N. Mahalle[5] *

[1] Department of Computer Science & Engineering (Artificial Intelligence), Bansilal Ramnath Agarwal Charitable Trust's Vishwakarma Institute of Information Technology, Pune (India)
[2] MIT Art, Design & Technology University, Pune (India)
[3] Department of Computer Engineering, MKSSS's Cummins College of Engineering for Women, Pune-411052, Maharashtra, (India)
[4] Department of Information Technology, Bansilal Ramnath Agarwal Charitable Trust's Vishwakarma Institute of Information Technology, Pune (India)
[5] Professor and Dean R&D, Bansilal Ramnath Agarwal Charitable Trust's, Vishwakarma Institute of Technology, Pune (India)

* Corresponding author: drsablenilesh@gmail.com (N. P. Sable), rajkumar.v.patil30@gmail.com (R. V. Patil), mdeore83@gmail.com (M. Deore), ratnmalab@gmail.com (R. Bhimanpallewar), aalborg.pnm@gmail.com (P. N. Mahalle)

Received 24 August 2023 | Accepted 16 July 2024 | Published 30 October 2024

unir
LA UNIVERSIDAD EN INTERNET

Abstract

In India, the demand for fruits and vegetables has been consistently increasing alongside the rising population, making crop production a crucial aspect of agriculture. However, despite the growing demand and potential profitability, farmers have been slow to transition from traditional food grain crops to fruits and vegetables. In this paper, we explore the changing demands of food categories in India, highlighting the shift towards increased consumption of fruits and vegetables. Despite the potential benefits, farmers face various challenges and uncertainties associated with cultivating these crops. To address this, we propose the use of Machine Learning (ML) and Deep Learning (DL) techniques to analyze historical market price data for fruits and vegetables from 2016 to 2021 and predict future prices. This accurate prediction system will aid farmers in deciding which crops to grow and when to harvest, ultimately maximizing profits.

Keywords

Agriculture, Cultivation, Data Analysis, Machine Learning, Regression.

DOI: 10.9781/ijimai.2024.10.005

I. Introduction

AGRICULTURE, the world's oldest and most important sector, has always been essential for supplying food, fibers, and fuel to humanity. Archaeological evidence places the origins of farming at about 10,000 years ago, when people began to depend on it for their food [1]. Agriculture plays a crucial role in the development of civilizations by cultivating the soil and raising livestock. Nevertheless, throughout millennia, agricultural growth progressed gradually. Using fire to regulate plant development was a common practice in early agricultural techniques since people had seen how well-established vegetation was following wildfires. Farmers gradually started tilling the soil and producing crops on tiny pieces of land by hand using simple equipment. As time went on, productive farming implements were developed, and yield-boosting irrigation methods were mastered [2].

As per the statistics of Annual crop production in India since 2003-04 fruit and vegetable production keeps on increasing [3]. With the continuous growth in population and changing food consumption patterns in India, there is an increasing demand for fruits and vegetables. However, farmers have been hesitant to shift from traditional food grain crops to fruits and vegetables due to various reasons. This paper aims to provide a solution by using ML and DL algorithms to analyze historical market price data of Mumbai Agricultural Produce Market Committee (Mumbai APMC) and predict fruit and vegetable prices, assisting farmers in making informed decisions about crop selection and harvesting.

Changes in Food Consumption Patterns: According to Dr. Richa Govil of Ashoka India, there has been a noteworthy rise in the consumption of fruits and vegetables despite a minor decline in the consumption of cereals like wheat and rice. Farmers have been sluggish to adopt fruit and vegetable crops despite this change [4]. Farmers face a number of difficulties, some of which have been highlighted

Please cite this article as: N. P. Sable, R. V. Patil, M. Deore, R. Bhimanpallewar, P. N. Mahalle. Machine Learning Based Agricultural Profitability Recommendation Systems: A Paradigm Shift in Crop Cultivation, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 1, pp. 39-54, 2024, http://dx.doi.org/10.9781/ijimai.2024.10.005

by a survey of those in the agricultural industry [4]. Farmers are reluctant to grow fruits and vegetables for a variety of reasons. Firstly, compared to food grains, these crops are riskier to grow since they are more reliant on environmental factors. The cultivation of fruits and vegetables also takes more labour, and developing nations like India have difficulties due to the lack of automation. There is necessity of a reliable prediction system. We suggest using Machine Learning (ML) and Deep Learning (DL) approaches to examine historical market price data for fruits and vegetables from 2016 to 2021 in order to address these issues. Farmers may decide which crops to cultivate and when to harvest them profitably by using prediction of future pricing for each month and year.

A. Key Highlight of Research

- Price Prediction for Fruits and Vegetables: The paper focuses on predicting market prices for fruits and vegetables to aid farmers in making informed selling decisions.
- A literature survey was done to analyze a wide range of methodologies for analyzing time series data, with a special focus on predicting fruit market prices.
- Analysis of ML and DL Algorithms: Conduct a comprehensive examination of ML and DL techniques that are applicable to predicting fruit market prices.
- Analysis of Regression Algorithms: An analysis of several regression methods and their mathematical models to determine their suitability for predicting prices.
- Evaluation Metrics: Discussion on diverse evaluation metrics applicable to regression algorithms, accompanied by their respective mathematical formulations.
- Determining the Optimal Regression Model: Applying evaluation metrics to identify the optimal regression model for precise price prediction.
- Predicting the optimal month for farmers to harvest their crops, maximizing the possible selling price in the market.

The paper is structured as follows: Section II provides a comprehensive literature survey on Time Series Data Analysis, as well as on ML and DL based approaches in regression algorithms. Section III outlines the Experimental Methodology employed in this study. In Section IV, we discuss the results obtained from our experiments Also we discuss future work and limitations of research. Finally, in Section V, we conclude the study and present future scope for further research.

II. Literature Survey

The agriculture industry as well as the economy at large heavily rely on predictions of fruit and vegetable prices in market yards. Farmers, sellers, and policymakers may make informed decisions about production, distribution, and marketing strategies due to accurate price projections. To predict the pricing of fruits and vegetables in market yards, researchers have used a variety of strategies and procedures throughout the years. This review of the literature seeks to provide readers an overview of the current research and methodology used to predict the prices of fruits and vegetables. Researcher used Time Series Analysis, ML and Artificial Intelligence (AI), Data Mining and Big Data Analytics. For accurate prediction, one must carefully examine historical data and take into account a number of variables that have an impact on market dynamics. To tackle this problem, researchers have used a variety of methods, such as time series analysis, ML, and data mining. Each methodology has its strengths and limitations, and the choice of a technique depends on the available data, research objectives, and computational resources.

A. Literature Survey on Time Series Analysis

Villaren M. Vibas et al. concentrated on the development of a mathematical model to analyze the retail price changes of essential agricultural commodities, particularly fruits (mango and banana) and vegetables (cabbage, pechay and tomato) in the Philippines' National Capital Region (NCR) [5]. The Philippine Statistics Authority (PSA) provided the study with data that covered the ten-year period from 2009 to 2018. The data was analyzed and predictive models were created using time series modelling approaches as ARIMA, SARIMA, and ARIMAx. The study's conclusions showed that during the span of 10 years, the monthly prices of all the items under investigation had increased. When projecting monthly retail prices of fruit commodities, the ARIMAX (5, 2, 2, x=mango) model was shown to be the most accurate, whereas the ARIMAX (2, 2, 1, x=banana) model excelled for bananas. The study suggested utilizing the ARIMAX (3, 2, 1, x=pechay) model for cabbage, the SARIMA (1, 1, 1)(1, 1, 1)12 model for pechay, and the SARIMA (2, 1, 1)(2, 1, 1)12 model for tomatoes for calculating monthly prices for vegetable commodities. The aim is to help consumers, farmers, traders, business owners, and policymakers make wise decisions about economic issues and long-term planning involving basic agricultural commodities in the NCR region.

Sarker Rakhal, et al. examines the dynamics of price transmission in the Canadian orange and apple markets. The analysis makes use of orange and apple import and retail prices of each month from 1996 to 2017 [6]. The author examines the amount, direction, and speed of price transmission between the upstream (import) and downstream (retail) levels using co-integration and error correction modelling techniques. According to the results, both commodities' import and retail prices have a single, long-term relationship, with the import price having an impact on the retail price. Additionally, the findings show that apples have asymmetric price transmission, whereby the margin corrects more quickly when it is constricted than when it is expanded.

Ali Jahangir et al. carried out research to examine the pricing and arrival patterns of apple produced at Jammu's Narwal market [7]. The Directorate of Horticulture, Planning and Marketing in Narwal provided ten years' worth of monthly secondary data on apple pricing and arrivals (from 2007–08 to 2016–17). Linear regression was used by author to find insight and pattern for apple. Moreover, seasonal indices were computed to investigate the cyclical changes in company activity linked to the year cycle. The results showed a favorable trend in both apple pricing and arrivals, with an anticipated yearly rise of Rs. 220.06 per quintal and arrivals of 15,969.42 quintals of apples. The primary period for apple arrivals in Narwal market was from August through January. Prices for apples ranged from the lowest in April to the highest in August. The seasonal indices showed that apple arrivals peaked in October and peaked at their lowest in April, whereas the seasonal indicator for pricing peaked in August and was lowest in April. Although all above time series analysis studies offer useful information for agricultural market decision-making, none of them particularly address price prediction or recommend the ideal month for farmers to harvest their crops for the highest potential price.

B. Literature Survey on Machine Learning (ML) and Deep Learning (DL)

L. Nassar et al. compare the effectiveness of deep learning (DL) models for predicting the prices of fresh produce (FP) markets with statistical and conventional ML models [8]. Two datasets are used: one from a website that lists daily crop prices in Taiwanese marketplaces, and the other, for daily strawberry transactions over a seven-year period, comes from a confidential source in Canada. The findings demonstrate that traditional ML models perform better than statistical models like ARIMA. Gradient Boosting performs well among the ML models, although simple and compound DL models meet it.

Convolutional Long Short-Term Memory Recurrent Neural Network (CNN-LSTM) with attention, a compound DL model, performs the best and can predict FP prices up to three weeks in advance.

Ifeanyi Okwuchi discusses the difficult challenge of predicting Fresh Produce (FP) pricing, taking into account elements like the produce's limited shelf life, inability to be stored for extended periods of time, and outside impacts like weather and climate change [9]. The goal of the project is to build machine learning-based models for FP yield and price prediction, including both traditional and deep learning models. A variety of Californian data are used, including weather, strawberry output, farm-gate pricing, and store purchase prices. To evaluate the different prediction models, the author suggests a brand-new aggregated error metric (AGM) that incorporates mean absolute error, mean squared error, and R2 coefficient of determination. In order to further enhance the predictions, stacking ensemble approaches are used, such as voting regressor and stacking using Support Vector Regression (SVR).

Razat Agarwal used machine learning techniques to classify and predict fruit images using a large dataset. Five supervised learning models were created and evaluated for their effectiveness in identifying fruit, including Random Forest (RF), Naive Bayes, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM). According to the experimental results, SVM performed better than the other methods for both large (95 fruits, 48905 images) and small (18 fruits, 8846 images) datasets. Additionally, it was observed that reducing the number of fruits i.e. small dataset negatively impacted the accuracy of each algorithm [10].

R. Dharavath and E. Khosla address the growing concern of inflation in India, particularly in the region of Bengaluru, Karnataka. The study's objective is to evaluate local fruit and vegetable prices and use seasonal ARIMA to predict future prices. Policymakers and people may prevent price rises by acting proactively by anticipating prices. Strategies might be developed to cut the price of fruits and vegetables, maintaining affordability for all residents, if predicted prices show a rise in the months to come [11].

C. Sharma et al. discuss the problem of agricultural profitability in India. They take into consideration factors like temperature, humidity, pH, rainfall, and other climatic conditions that have an impact on crop yields and, in turn, have an effect on the pricing of fruits, vegetables, and pulses. The authors are aware of the fact that farmers lack information on the choice of crops and anticipated pricing. This issue is addressed by the proposed approach, which predicts crop prices based on previous data trends. Utilizing Decision Tree Regression as a supervised machine learning approach, the method takes into account a number of variables like pH, humidity, precipitation, temperature, and market price [12].

M. Kankar and M.A. Kumar highlight the necessity for technical developments in India's agriculture industry. They propose employing deep learning models like LSTM and BiLSTM as well as deep neural networks like CNN to predict the price of agricultural products. The goal of the project is to reduce market volatility and increase the precision of price prediction for fruits and vegetables by using LSTM models for predicting market prices and a CNN model to classify photos based on variety and quality [13].

R.K. Paul et al. focused on the price prediction of vegetables, particularly Brinjal, at 17 important marketplaces in Odisha, India [14]. The Generalized Neural Network (GRNN), SVR, Random Forest (RF), and Gradient Boosting Machine (GBM) were machine learning methods that were compared with conventional statistical models like the Autoregressive integrated moving average (ARIMA). The findings showed that, when compared to ARIMA, ML approaches, notably GRNN, often displayed higher prediction accuracy. To prove that the ML models outperformed the conventional strategy, the study also used a variety of accuracy metrics and statistical tests. Table I gives details of related literature papers [5]-[14].

TABLE I. Literature Survey

Paper	Research Area	Description
[5] 2019	Time Series	The study focuses on developing mathematical models to analyze price fluctuations in essential agricultural products, specifically mangoes, bananas, cabbage, pechay, and tomatoes in the Philippines' National Capital Region over ten years. These models, including ARIMA and SARIMA, help predict and manage price changes, benefiting consumers, farmers, traders, and policymakers.
[6] 2021	Time Series	The study investigates price transmission dynamics in the Canadian orange and apple markets from 1996 to 2017. It reveals a single, long-term relationship between import and retail prices, with import prices influencing retail prices. Apples exhibit asymmetric price transmission, adjusting faster during constraints than expansions.
[7] 2018	Time Series	The study analyzed ten years of apple pricing and arrivals data from Jammu's Narwal market, revealing a favorable trend with annual price increases of Rs. 220.06 per quintal and peak arrivals from August to January. However, the research did not provide specific price predictions or optimal harvest months for farmers.
[8] 2020	DL	The study compares DL, statistical, and conventional ML models for fresh produce market price prediction. Traditional ML outperforms ARIMA, with CNN-LSTM performing best, predicting prices three weeks ahead.
[9] 2020	ML	The author addresses the complex challenge of forecasting Fresh Produce (FP) pricing, considering factors like limited shelf life, weather impact, and climate change. They aim is to create machine learning models, both traditional and deep learning, using Californian data. A novel aggregated error metric (AGM) is proposed, which blends various evaluation metrics. Stacking ensemble methods, including voting regressor and SVR stacking, are employed to improve predictions.
[10] 2019	ML	The study employed machine learning to classify and predict fruit images with five models. SVM outperformed the others for large and small datasets, highlighting the impact of reduced fruit diversity on algorithm accuracy.
[11] 2019	ML	The author focuses on the rising inflation concern in Bengaluru, Karnataka, India, with a study aiming to assess local fruit and vegetable prices. Seasonal ARIMA models predict future prices to enable proactive price control strategies for affordability.
[12] 2023	ML	The authors examine agricultural profitability in India, considering climate factors (temperature, humidity, pH, rainfall) affecting crop yields and pricing. They address farmers' information gaps by using Decision Tree Regression to predict crop prices based on historical data, including pH, humidity, precipitation, temperature, and market prices.
[13] 2022	DL	The authors underscore the need for technological advancements in India's agriculture sector, advocating the use of LSTM, BiLSTM, and CNN deep learning models to enhance price prediction accuracy, reduce market volatility, and classify fruits and vegetables based on variety and quality in a research project.
[14] 2022	ML/DL	The study focused on predicting Brinjal prices in 17 key markets in Odisha, India, comparing machine learning techniques like GRNN, SVR, RF, and GBM with traditional ARIMA models. Results demonstrated superior prediction accuracy with ML methods, supported by various accuracy metrics and statistical test.

C. Gap Analysis

In the literature studies [5]-[14], various Time Series Analysis, ML and DL, Data Mining and Big Data Analytics, were employed to predict market prices for specific fruits and fresh produce. These studies compared different models, including traditional machine learning, deep learning, and ensemble approaches, to assess their effectiveness in price prediction. The findings indicated that certain models, such as Decision tree regression, SVR and Gradient Boosting, performed well in predicting fruit and fresh produce prices. However, none of the studies specifically addressed the recommendation of the ideal month for farmers to harvest their crops for the highest potential price. This suggests that further research is needed to explore this aspect and provide insights for farmers regarding optimal harvest timing for maximizing profits in the market. While mean absolute error and root mean square error were employed as accuracy measurements in some research, more standardized assessment metrics are required to compare various prediction algorithms. The development of reliable assessment criteria that take predicting accuracy and economic ramifications into account might be a key area of future study.

D. Objective

- Evaluate the effectiveness of several Time Series Analysis, ML and DL regression methods in predicting market prices for certain fruits and vegetables in the Mumbai APMC and Maharashtra, India.
- To identify the most effective regression algorithm among the studied ML and DL techniques for accurately predicting market prices of fruits and vegetables in the specified region.
- Create a predictive model that can accurately estimate prices for fruits and vegetables. This model will help farmers determine the four months with the greatest pricing, allowing them to make smart decisions about when to harvest their crops for maximum profitability in the market.

This paper aims to assist farmers under Mumbai APMC and Maharashtra, India by utilizing and comparing various ML and DL algorithms to determine the optimal harvest timing for maximizing profits. In addition, a standardized metric is employed to compare the performance of all the algorithms. By considering these factors, the research offers valuable guidance for decision-making in the agricultural industry and contributes to maximizing profitability in the market.

III. Experimental Method

The Experimental Method section of this research paper outlines the step-by-step process employed to predict fruit and vegetable prices in the Maharashtra market. To start, an appropriate dataset made up of past price records is gathered. For the purpose of ensuring data integrity, the dataset is then put through a comprehensive cleaning procedure. In order to construct and evaluate models, the dataset is then split into training and test sets. On the training set, a variety of machine learning regression methods are used, and their effectiveness is measured using evaluation metrics like RMSE, MSE, and MAE. These criteria are used to determine the most effective algorithm, which is then used to predict prices. After then, the accuracy of the estimations is checked against current market values. Fig. 1 illustrates a system model at a higher level - Level 0, while Fig. 2 shows an additional representation of system model - low level system model level 1.

The experiment was conducted using Python, a versatile programming language widely used in data science and machine learning.

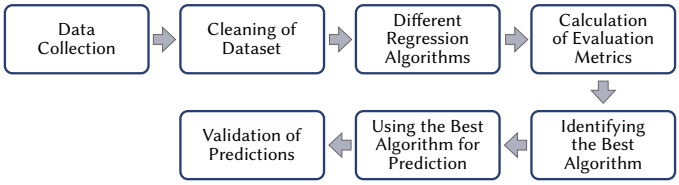

Fig. 1. Level 0 System Model: Higher level.

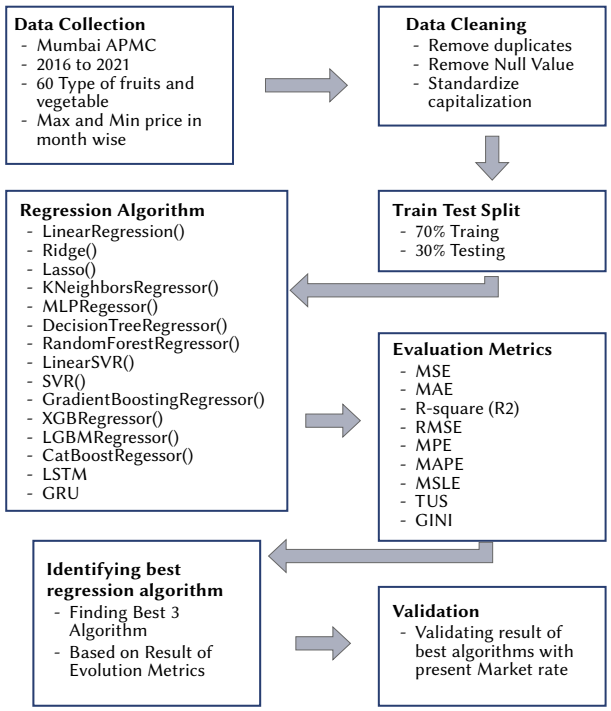

Fig. 2. Level 1 System Model: Low level.

The following Python libraries were utilized:

- **pandas** for data manipulation and analysis.
- **numpy** for numerical operations.
- **sklearn (scikit-learn)** for implementing machine learning models and data preprocessing.
- **xgboost** for the XGBoost model.
- **lightgbm** for the LightGBM model.
- **catboost** for the CatBoost model.

The code was executed on a Jupyter notebook with on Windows 10, 64 bit, NVIDIA GeForce 920M environment. Python version 3.8 was used. All necessary packages were installed via pip or conda package managers.

A. Data Collection

The first step in our study involved collecting relevant data, specifically historical records of fruit and vegetable prices in the Maharashtra market, focusing on a specific time period. We obtained a dataset of fruit and vegetable prices from Mumbai APMC, spanning from April 2016 to March 2021. This dataset includes information on 60 different types of fruits and vegetables.

The collected data consists of the maximum price recorded for each commodity within a given month, as well as the minimum price during the same period. Among the fruits and vegetables included in the dataset are BOR, LIME, GUAVA, KESAR Mangos, PAPAYA, GAJAR (Carate), VANGI (Eggplant), KANDA (Onion), and TAMBATE (Tomato), among others. Table II shows some rows of the datasets.

TABLE II. DATASET

Item	Month	Year	Min Price	Max Price
BOR	April	2016	1600	1900
POMEGRANATE	October	2017	7671	7988
GUAVA	September	2019	3170	4723
TAMBATE	July	2020	3496	4084
PAPAYA	August	2016	968	1432

B. Cleaning of Dataset

After collecting the dataset, we proceeded with a comprehensive cleaning process to address any inconsistencies, errors, or missing values present [15]-[17]. The data cleaning phase involved several tasks, including the removal of duplicate entries, handling missing data through imputation or deletion, and resolving formatting or labeling inconsistencies. This step aimed to ensure the dataset's suitability for subsequent analysis and modeling. To begin, we eliminated any rows containing null values. During non-seasonal periods or months, certain produce items are not available in the market, making their absence understandable. Consequently, removing these null value rows helped maintain data integrity and accuracy. Furthermore, we took measures to address any duplication within the dataset. Duplicate entries can skew analyses and lead to inaccurate results. By identifying and eliminating duplicated data, we ensured that each observation within the dataset was unique and representative. Finally, we devoted attention to resolving inconsistencies in the dataset. This involved rectifying variations in formatting or labeling that could hinder analysis efforts. Overall, through the rigorous data cleaning process, we successfully prepared the dataset for further analysis and modeling by eliminating null values, removing duplicates, and resolving inconsistencies.

We have manually converted the dataset, originally handwritten, to CSV format. Then we used Python programming to clean the data. We used dropna () method of Pandas Data Frame to remove Null data. The Month column of the dataset, which originally included month names as strings, changed into numerical values to enhance computational efficiency. This was done via the replace function in the Pandas library. Regarding feature and target selection, the features (X) used for training include Item, Month, and Year, and the target variables (Y) include Min_Price and Max_Price. Separate target variables Y_min and Y_max were also defined for more specific model training. With the help of scikit-learn's ColumnTransformer and OneHotEncoder, the categorical variable Item was converted into a numerical format that is appropriate for machine learning models through the process of one-hot encoding.

C. Regression Algorithms

We utilized the sklearn library in Python to employ and evaluate various regression algorithms for predicting fruit and vegetable prices in the market. By comparing the results of different algorithms, we aimed to identify the most suitable model for accurate price predictions. Sections 3.3.1 to 3.3.12 present the detailed technical algorithms for the twelve regression techniques referenced in this paper [18]-[21].

1. Linear Regression

The Linear Regression function from sklearn library is a well-liked implementation of this approach in machine learning libraries. Linear regression is a fundamental statistical technique used for predictive modelling [22]-[23]. The provided dataset, which covers the years 2016 through 2021, may be used to estimate fruit and vegetable prices, and linear regression can be a useful technique in this regard. With the use of features like "Fruit or vegetable", "month," and "year," linear regression may determine a connection between these elements and the associated "max price" and "min price" values. Linear regression mathematical relationship between independent and dependent variable is shown in equation (1) [22]-[23], where independent variable are Fruit or vegetable, month, and year and dependent variable are Max price and Min price.

$$y_0 = b_0 + b_1x_1 + b_2x_2 + \cdots + x_nb_n \quad (1)$$

Where: y is still the dependent variable, x_1, x_2,---, x_n are the independent variables, b_o is the intercept term, b_1, b_2, ----, b_n are the coefficients associated with each independent variable. Algorithm 1 gives technical implantation of Linear Regression.

Algorithm 1: Linear Regression

1. Initialize weights w and bias b randomly
2. Set learning rate α and number of iterations N
3. For $i = 1$ to N:
4. For each training sample (x, y):
5. Predict $\hat{y} = w * x + b$
6. Compute error $e = \hat{y} - y$
7. Update weights: $w = w - \alpha * e * x$
8. Update bias: $b = b - \alpha * e$
9. Return final weights w and bias b

2. Ridge Regression (L2 Regularization)

Ridge regression adds a penalty term to the traditional linear regression model as shown in equation (2), forcing the model to not only fit the data but also minimize the sum of squared coefficients [24]. As a result of minimizing the effects of multicollinearity among the features, this regularization strategy stabilizes the model and helps prevent overfitting. Ridge regression can successfully manage the possible strong correlation between months and years, which may affect the price variations of fruits and vegetables. Algorithm (2) shows Ridge Regression (L2 Regularization).

$$\Sigma_{i=1}^{n}\left(y_i - \Sigma_{j=1}^{p} x_{ij}\,\beta_j\right)^2 - \lambda\Sigma_{j=1}^{n}\beta_j^2 \quad (2)$$

Where y is dependent variable, x is independent variable, β is coefficient vector to be estimated, λ is the penalty parameter, also known as the tuning parameter, controlling the strength of the penalty term.

Algorithm 2: Ridge Regression (L2 Regularization)

1. Input: Feature matrix X, target vector y, regularization parameter λ
2. Append a column of ones to X for the intercept term
3. Compute $X\text{^}T * X$ and $X\text{^}T * y$
4. Add $\lambda * I$ to $X\text{^}T * X$ (where I is the identity matrix)
5. Compute the inverse of $(X\text{^}T * X + \lambda * I)$
6. Compute the ridge coefficients: $\beta = (X\text{^}T * X + \lambda * I)\text{^}(-1) * X\text{^}T * y$
7. Output: Ridge coefficients β

3. Lasso Regression (L1 Regularization)

Lasso regression, on the other hand, adds a penalty term based on the absolute values of coefficients [25] as shown in equation (3). By setting some coefficients to absolutely zero, it accomplishes feature selection, thereby removing less important characteristics. In the context of predicting fruit and vegetable prices, Lasso can pinpoint the characteristics that have the greatest bearing on price changes over time, perhaps emphasizing seasonal trends and particular elements that are key in determining price.

$$\Sigma_{i=1}^{n}\left(y_i - \Sigma_{j=1}^{p} x_{ij}\,\beta_j\right)^2 - \lambda\Sigma_{j=1}^{n}|\beta_j| \quad (3)$$

Where y is dependent variable, x is independent variable, β is coefficient vector to be estimated, λ is the penalty parameter, also known as the tuning parameter, controlling the strength of the penalty term. Algorithm (3) shows Lasso Regression (L1 Regularization).

Algorithm 3: Lasso Regression (L1 Regularization)

1. Initialize weights (coefficients) randomly or with zeros.
2. Define a loss function (e.g., Mean Squared Error) and regularization strength (lambda).
3. Perform gradient descent or coordinate descent:
   for each iteration {
   a. Compute predictions using current weights.
   b. Compute gradients of the loss function with respect to weights.
   c. Update weights using gradients and regularization term:
   weight = weight - (learning_rate * (gradient + lambda * sign(weight)))
   }
4. Repeat until convergence or maximum iterations reached.

We can take benefit of the advantages of both techniques by incorporating Ridge and Lasso regularization techniques into linear regression models in order to produce predictions of fruit and vegetable prices that are more reliable and accurate while also reducing the risk of overfitting and dealing with potentially irrelevant features in the dataset.

4. K-Nearest Neighbors (KNN) Regression

Based on historical data, the non-parametric supervised learning technique K-Nearest Neighbours (KNN) regression is used to predict fruit and vegetable prices. KNN Regressor may be used to estimate price trends with a dataset covering the years 2016 through 2021 and include variables such as fruit or vegetable kind, month, year, and maximum and minimum prices. The model estimates prices for certain products at a particular moment by computing the average of the 'k' closest data points with comparable attributes. Because of its simplicity and capacity to identify regional trends in the data, KNN is particularly helpful for predicting the prices of fruits and vegetables, which may experience seasonal or regional variations [26]. The predicted value for a new input x is computed as the average (or weighted average) of the target values of the k nearest neighbors of x and given by equation (4) [26].

$$\hat{y}(x) = \frac{1}{k}\sum_{i=1}^{k} y_i \quad \textbf{(4)}$$

Where x is the input feature vectors (independent variables), y is Target values (dependent variable), k is the number of nearest neighbors, $\hat{y}$ is the predicted target value for the input x. yi is the target value of the ith nearest neighbor of x. Algorithm 4 shows K-Nearest Neighbors (KNN) regression.

Algorithm 4: K-Nearest Neighbors (KNN) regression

1. function KNN_Regression(X_train, y_train, X_test, k):
2. for each test sample x_test in X_test:
3. Calculate distances d between x_test and all samples in X_train
4. Sort distances d in ascending order
5. Select the top k samples from X_train based on smallest distances
6. Predict y_test for x_test as the average of y_train values for the selected k samples
7. return predicted y_test values

5. Multi-Layer Perceptron Regressor

The Multi-Layer Perceptron (MLP) Regressor is a class of neural network used for regression problems, we use MLP Regressor function from sklearn. neural network library to use MLP [24]. This model may be used to project future price patterns based on previous data in the context of projecting fruit and vegetable prices. The 2016–2021 dataset, which includes characteristics like fruit or vegetable kind, month, year, maximum price, and minimum price, enables the MLPRegressor to learn intricate patterns and correlations within the data and provide price prediction. By addressing non-linear correlations between input characteristics and target pricing effectively, this strategy enables companies to make well-informed decisions, optimize their supply chains, and anticipate market changes. For a single-hidden-layer MLP regressor with h neurons in the hidden layer, the output of the hidden layer can be calculated as shown in equation (5) [27].

$$Z = \emptyset(XW^{(1)} + b^{(1)}) \quad (5)$$

Where $W^{(1)}$ is the weight matrix connecting the input layer to the hidden layer, with dimensions (n, h), $b^{(1)}$ is the bias vector for the hidden layer, with dimensions $(1, h)$, ϕ is the activation function applied element-wise to the weighted sum.

The output of the MLP Regressor can then be calculated as shown in equation (6).

$$\hat{Y} = ZW^{(2)} + b^{(2)} \quad (6)$$

Where, $W^{(2)}$ is the weight matrix connecting the hidden layer to the output layer, with dimensions $(h, 1)$, $b^{(2)}$ is the bias for the output layer. Algorithm 5 shows Multi-Layer Perceptron Regressor.

Algorithm 5: Multi-Layer Perceptron Regressor

1. Initialize weights randomly
2. Define activation function (e.g., sigmoid, ReLU)
3. Define learning rate (alpha), number of layers (L), number of neurons per layer (N)
4. For each epoch:
   For each training example (X, y):
   Forward pass:
   Calculate output of each neuron in each layer using current weights
   Apply activation function to each neuron's output
   Backward pass:
   Calculate error derivative with respect to output
   Update weights using gradient descent
5. Repeat step 4 until convergence or maximum number of epochs reached
6. Output the trained MLP model

6. Decision Tree Regression

The Decision Tree Regressor is a regression technique that uses historical data from 2016 to 2021 and a Decision Tree algorithm to predict fruit and vegetable prices. The dataset is recursively partitioned using variables like type (fruit or vegetable), month, and year to create a tree-like model. Price, as indicated by maximum and minimum prices, is the dependent variable to be predicted. The Decision Tree can manage non-linear patterns in price swings because it can record complicated correlations between characteristics and the target [28]-[29]. The mathematical equation for decision tree regression is expressed in equation (7).

$$\hat{y} = \sum_{i=1}^{N} w_i . I(x \in R_i) \quad (7)$$

Where, $\hat{y}$ represents the predicted output. N is the number of leaf nodes, W_i is the predicted value at leaf node, R_i denotes the region defined by the ith leaf node. $I(x \in R_i)$ is an indicator function that returns 1 if the input x belongs to the region R_i, and 0 otherwise. Algorithm 6 shows Decision Tree Regression.

Algorithm 6: Decision Tree Regression

```
1. DecisionTreeRegression(data, max_depth, min_samples_split):
2.     if max_depth == 0 or len(data) < min_samples_split:
3.         return Leaf Node (mean(data. target))
4.     else:
5.         best_split = find_best_split(data)
6.         if best_split == None:
7.             return LeafNode(mean(data.target))
8.         left_data, right_data = split(data, best_split)
9.         left_subtree = Decision Tree Regression (left_data, max_
           depth - 1, min_samples_split)
10.        right_subtree = Decision Tree Regression (right_data, max_
           depth - 1, min_samples_split)
11.        return Decision Node (best_split, left subtree, right_subtree)
12. find_best_split(data):
13.    best_split = None
14.    best_mse = infinity
15.    for each feature in data.features:
16.        for each threshold in unique(data[feature]):
17.            left_data, right_data = split(data, (feature, threshold))
18.            mse = weighted_mse(left_data, right_data)
19.            if mse < best_mse:
20.                best_mse = mse
21.                best_split = (feature, threshold)
22.    return best_split
```

In algorithm 6 Lines 1-11 the main recursive function for building the decision tree is defined. It stops if the maximum depth is reached or the number of samples is too small. Otherwise, it finds the best split and recursively builds left and right subtrees. Lines 12-22 define the function to find the best split based on the minimum mean squared error (MSE). This function evaluates all possible splits and returns the one with the lowest MSE.

7. Random Forest Regression

Random Forest Regressor is a popular ML algorithm based on the Random Forest ensemble method. It works for regression jobs and has the ability to handle both categorical and numerical data. It becomes a useful tool for solving a variety of regression issues by integrating numerous decision trees, which decreases overfitting and increases prediction accuracy [30]. The model can efficiently analyze features like the type of fruit or vegetable, month, year, maximum price, and minimum price. The model can predict future price changes by using historical information, which helps producers, suppliers, and customers make wise decisions and adjust to market changes. The mathematical equation for Random Forest Regression can be summarized as given in equation (8).

$$\hat{y} = \frac{1}{2}\sum_{i=1}^{N} f(x_i) \quad (8)$$

Where $\hat{y}$ is the predicted output, N is the number of trees in the forest, $f(x_i)$ represents the output (prediction) of the i^{th} decision tree in the forest for input x. Algorithm 7 shows Random Forest Regression.

Algorithm 7: Random Forest Regression

```
1. Input: Training data (X_train, y_train), number of trees N, max
   depth D, sample size S
2. Initialize an empty list of trees: forest = [ ]
3. for i = 1 to N do
4.     Sample with replacement S data points from (X_train, y_train)
5.     Build a Decision Tree Regression T on the sampled data with
       max depth D
6.     Add tree T to the forest
7. end for
8.
9. Function Predict(X_test):
10.    Initialize predictions = [ ]
11.    for each tree T in forest do
12.        pred = T.predict(X_test)
13.        Add pred to predictions
14.    end for
15.    return average(predictions)
16. end Function
```

8. Support Vector Machine: Linear Kernel and Radial Basis Function Kernel

Multidimensional data may be handled by Support Vector Machines (SVM) using linear and RBF (Radial Basis Function) kernels, making them appropriate for analyzing features like fruit or vegetable, month, year, maximum price, and minimum price. In order to approximate linear correlations between features and prices, LinearSVR function from sklearn.svm library uses a linear kernel that performs well for datasets with linearly separable classes. However, SVR with an RBF kernel may detect non-linear patterns in the data, which is crucial for detecting intricate interactions and seasonal changes that may have an impact on the pricing of fruits and vegetables [31]. The decision function for the linear kernel SVM is given by equation (9).

$$f(x) = sign(w.x + b) \quad (9)$$

Where w is the weight vector, x is the input vector, b is the bias term, sign(·) is the sign function, returning -1 for negative values and 1 for non-negative values. The decision function for the RBF kernel SVM is given by equation (10).

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b) \quad (10)$$

Where α_i are the Lagrange multipliers obtained during training, y_i are the class labels of the training data, x_i are the support vectors, $K(x_i, x)$ is the kernel function, which in the case of RBF kernel is given by equation (11).

$$K(x_i, x) = exp(-\gamma \|x_i - x\|^2) \quad (11)$$

Where γ is the kernel parameter and $\|\cdot\|$ denotes the Euclidean distance. Algorithms 8 and 9 show SVM Liner Kernel and SVM Radial Basis function.

Algorithm 8: Support Vector Machine: Linear Kernel

1. Input: Training data (X, y), regularization parameter C
2. Initialize: weights w, bias b
3. while not converged:
4. for each (xi, yi) in (X, y):
5. if $yi * (w \cdot xi + b) < 1$:
6. $w = w + \eta * (yi * xi - 2 * \lambda * w)$
7. $b = b + \eta * yi$
8. else:
9. $w = w - \eta * 2 * \lambda * w$
10. Output: weights w, bias b

Algorithm 9: Support Vector Machine: Radial Basis Function Kernel

1. Input: Training data (X, y), regularization parameter C, kernel parameter γ
2. Initialize: Lagrange multipliers α, bias b
3. while not converged:
4. for each (xi, yi) in (X, y):
5. compute kernel: $K(xi, xj) = exp(-\gamma * ||xi - xj||^2)$
6. if $yi * (\Sigma \alpha j\, yj\, K(xj, xi) + b) < 1$:
7. $\alpha i = \alpha i + \eta * (1 - yi * (\Sigma \alpha j\, yj\, K(xj, xi) + b))$
8. $b = b + \eta * yi$
9. else:
10. $\alpha i = \alpha i$
11. Output: Lagrange multipliers α, bias b

9. Gradient Boosting

Gradient Boosting Regressor is a popular machine learning algorithm used for regression tasks. It systematically builds several weak learners (often decision trees), with each tree attempting to fix the flaws of the one before it [32]. Gradient Boosting can successfully handle complicated interactions between data like month, year, max price, and min price in the context of fruit and vegetable price prediction. It can predict for various fruits and vegetables across the 2016–2021 dataset by capturing non-linear patterns. The mathematical equation for Gradient Boosting Regressor can be represented by equation (12).

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{12}$$

Where $F(x)$ is the final prediction function, M is the number of weak learners (trees) in the ensemble, $h_m(x)$ represents the output of the m^{th} weak learner, typically a decision tree, γ_m is the weight (or learning rate) assigned to the m^{th} weak learner. Algorithm 10 shows gradient boosting Regressor.

Algorithm 10: Gradient Boosting Regressor

1. Initialize $F\_0(x) = y_{mean}$
2. For $m = 1$ to M do:
3. Compute residuals $r\_m = y - F_{(m\text{-}1)}(x)$
4. Fit a base learner $h_{m(x)}$ to residuals r_m
5. Compute optimal step size γ_m
6. Update model: $F_{m(x)} = F_{(m\text{-}1)}(x) + \gamma_m\, h_m(x)$
7. End For
8. Output final model $F_{M(x)}$

10. XGBoost

The XGBRegressor is an optimized implementation of Gradient Boosting. Due to its regularization methods, parallel processing, and the handling of missing data, it delivers improved performance and efficiency [33]. XGBoost can use the dataset's temporal properties (month and year) to capture the impacts of seasonality and provide accurate price estimates for fruit and vegetable prices. The mathematical equation for predicting the target variable y using XGBRegressor is shown in equation (13).

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{k} f_k(x_i), \qquad f_k \in F \tag{13}$$

Where $\hat{y}_i$ is the predicted value for the i^{th} instance, x_i represents the features of the i^{th} instance, K is the number of trees in the ensemble, f_k is an individual decision tree from the ensemble. F is the space of all possible decision trees. Algorithm 11 shows XGBoost regressor

Algorithm 11: XGBoost Regressor

1. Initialize model with a constant value: $\emptyset(x_i) = \lambda$
2. For each iteration $t = 1$ to T:
3. Compute the negative gradient of the loss function for each data point: $g\_i = \partial L(y_i, \emptyset(x_i))/\partial \emptyset(x_i)$
4. Fit a regression tree to the negative gradients: $T_t = TreeFit(\{(x_i, g_i)\})$
5. Update the model: $\emptyset(xi) = \emptyset(xi) + \eta * Tt(xi)$
6. End for

11. Light GBM

LGBMRegressor is another gradient boosting algorithm known for its high speed and memory efficiency. By splitting trees into leaf-wise rather than levels, it uses less processing power [34]. LightGBM is a useful model for estimating price trends because it can effectively manage the temporal features of the dataset and accommodate fluctuations in pricing for various produce over time in the context of fruit and vegetable price prediction. LGBMRegressor Mathematical equation is given by equation (14).

$$Y = \text{BaseTree}(x) - \text{lr}.\text{Tree}_1(x) - \text{lr}.\text{Tree}_2(x) - \text{lr}.\text{Tree}_3(x) \tag{14}$$

Where Y represents the predicted target value, Base Tree(x) is the output of the base tree, which is essentially the initial prediction made by the model. $\text{Tree}_1(x)$, $\text{Tree}_2(x)$, $\text{Tree}_3(x)$ are the contributions from individual trees. lr denotes learning rate. Algorithm 12 shows Light Gradient Boosting Model Regressor.

Algorithm 12: LGBMRegressor

1. Initialize dataset D, number of trees T, learning rate lr
2. Initialize model: $Y = BaseTree(x)$
3. For $t = 1$ to T:
4. Compute residuals $r = Y - lr * Tree_t(x)$
5. Fit regression tree to residuals: $Tree_t(x)$
6. Update model: $Y = Y - lr * Tree_t(x)$
7. End For
8. Output: Final model Y

12. CatBoost

CatBoost Regressor is a popular algorithm for regression tasks, designed to handle both numerical and categorical features [34]. The fruit and vegetable price prediction dataset, which contains categorical variables like month, is a good fit for it since it can handle categorical data without the requirement for explicit encoding. The model is effective for large-scale prediction tasks when the verbose option

is set to 0, which guarantees that it operates silently and prevents unnecessary output during training and prediction. It is given by equation (15) and algorithm 13 .

$$L(t,a) = \frac{\sum_{i=0}^{N} w_i |a_i - t_i|}{\sum_{i=0}^{N} w_i} \quad (15)$$

Where, $L(t, a)$ represents the loss function. t_i is the true target value for the i[th] sample, a_i is the predicted value for the i[th] sample, w_i denotes the weight assigned to the i[th] sample (usually equal to 1).

The range of regression algorithms discussed above offers a diverse set of tools for predicting fruit and vegetable prices based on historical data spanning 2016 to 2021. Fundamental methods for modelling variations in prices include linear regression, ridge, and lasso, while K-Nearest Neighbours captures regional and seasonal patterns. MLPRegressor and other neural network techniques uncover complex patterns in the data. Support Vector Machines handle multidimensional data and linear/RBF patterns, whereas Decision Tree and Random Forest handle non-linear connections. Gradient Boosting, XGBoost, LightGBM, and CatBoost optimize performance, scalability, and efficiency, making them suitable for large datasets.

Algorithm 13: CatBoost Regressor

1. Initialize ensemble of decision trees
2. For each tree:
3. Initialize leaves with average target value
4. For each feature:
5. For each split point:
6. Calculate loss reduction using **equation 15**
7. Choose the best split based on loss reduction
8. Update leaf values based on targets within each leaf
9. Train until convergence or maximum number of iterations
10. Output ensemble of decision trees

D. Evaluation Metrics for Regression Algorithms

Regression algorithms use evaluation metrics to measure the effectiveness and precision of the model's predictions in relation to the actual target values. Mean Absolute Error (MAE) [35]-[36], Mean Squared Error (MSE) [35]-[36], and R-squared (R2) or Coefficient of Determination [36] are the three most often used assessment metrics for regression models employed in related studies [9]-[34]. In this paper we have also compared different regression algorithm with Root Mean Square Error (RMSE) [35]-[36], Mean Percentage Error (MPE) [36], Mean Absolute Percentage Error (MAPE) [36], Huber Loss (HL) [37], Mean Squared Logarithmic Error (MSLE) [38], Theil's U Statistic (TUS) [39], Gini Coefficient (Gini) [40]. The evaluation of each regression model was carried out with the help of the evaluation metrics mentioned above, and we utilized numpy and the sklearn. metrics package in Python.

1. Mean Absolute Error (MAE)

MAE calculates the average absolute difference between the predicted values and the actual target values. It measures the average magnitude of errors without considering their direction [35]. Equation (16) gives MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (16)$$

Where, n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i.

2. Mean Squared Error (MSE)

The average of the squared differences between the predicted values and the actual target values is computed using MSE. It is frequently applied in different regression techniques and penalizes larger errors more severely than MAE [35]. Equation (17) gives MSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (17)$$

Where n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i.

3. R-squared (R2) or Coefficient of Determination

R-squared calculates the proportion of the target's variance that can be predicted from the model's independent variables (features). The value ranges from 0 to 1, with 0 denoting that the model explains no variation and 1 denoting a perfect match [35]. R-squared is computed using equation (18).

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}} \quad (18)$$

Where $SS_{residuals}$ is the sum of squared residuals (difference between actual and predicted values), SS_{total} is the total sum of squares (variance of the actual target values).

4. Root Mean Square Error (RMSE)

RMSE quantifies the average discrepancy between the projected and actual prices. The calculation involves finding the square root of the average of the squared discrepancies between the actual (y_i) and predicted ($\hat{y}_i$) prices as shown in equation (19).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (19)$$

Here 'n' represents the number of predictions. A lower RMSE indicates higher prediction accuracy.

5. Mean Percentage Error (MPE)

MPE is a metric employed to assess the precision of predictions. The function computes the mean percentage deviation between projected and real prices. The equation (20) represents MPE.

$$MPE = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)}{y_i} * 100 \quad (20)$$

Where n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i.

6. Mean Absolute Percentage Error (MAPE)

MAPE is a metric that quantifies the accuracy of predictive models. The algorithm computes the mean absolute percentage deviation between predicted and observed prices. The equation (21) shows MAPE.

$$MPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{(y_i - \hat{y}_i)}{y_i}\right| * 100 \quad (21)$$

Where n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i. MAPE expresses error as a percentage, which simplifies its interpretation. A smaller MAPE signifies a higher level of accuracy in predictions.

7. Mean Squared Logarithmic Error (MSLE)

The MSLE quantifies the average squared discrepancy between the natural logarithm of the predicted values and the actual values. It is commonly used in fruit prediction of prices, where the estimates can vary greatly in magnitude. MSLE is computed using equation (22).

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (log(y_i + 1) - log(\hat{y}_i + 1))^2 \quad (22)$$

Where n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i. This measure imposes equal fines for both underestimation and overestimation, rendering it appropriate for datasets that exhibit skewness. A lower MSLE value suggests more accuracy in forecasting fruit prices.

8. Loss (HL)

The Huber loss function, commonly used in regression applications such as fruit price prediction, combines the resilience of mean squared error (MSE) with the responsiveness of mean absolute error (MAE). It reduces the influence of extreme values on the model's performance. Equation (23) shows HL and is given by $L_\delta(y - \hat{y})$.

$$L_\delta(y - \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & |y - \hat{y}| \le \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta), & Otherwise \end{cases} \quad (23)$$

Where y is actual target value, $\hat{y}$ is the predicted value, δ is a parameter that determines the threshold beyond which the loss becomes linear rather than quadratic. We have use default value δ as 1. The Huber loss function offers a well-balanced method by penalizing significant errors in a linear manner within a specified tolerance (δ), and in a quadratic manner beyond that tolerance. This guarantees enhanced robustness against outliers values while also achieving efficient model optimization.

9. Theil's U Statistic (TUS)

Theil's U Statistic (TUS) is a metric used in econometrics to evaluate the accuracy of predictions, specifically in the context of predicting fruit prices. The process involves comparing the observed values with the expected values, taking into account both bias and variability. TUS equation is shown in equation (24).

$$TUS = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n} y_i}} \quad (24)$$

Where, n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i. A TUS number near to 0 suggests accurate predictions, whereas higher levels indicate more prediction errors.

10. Gini Coefficient (Gini)

The Gini Coefficient (Gini) is a statistical indicator that measures the level of inequality in a given distribution. It is commonly employed in the field of economics for predicting fruit prices. The scale spans from 0, indicating complete equality, to 1, representing extreme disparity. Equation (25) shows Gini Coefficient.

$$Gini = \frac{\sum_{i=1}^{n}(2i - n - 1)(y_i - \hat{y}_i)}{\sum_{i=1}^{n} y_i} \quad (25)$$

Where n is the number of data points, y_i is the actual target value for data point i, $\hat{y}_i$ is the predicted value for data point i. A higher Gini coefficient indicates a higher level of inequality, suggesting the presence of possible price differences in the fruit market.

To obtain a comparison evaluation of the regression model's performance, it is crucial to combine these measures. R-squared provide information on how effectively the model accounts for the variance in the target variable, whereas MAE and MSE give a sense of the absolute error size. When working with complicated models and huge datasets, R-squared must be utilized with caution because it may occasionally be deceptive and should be combined with additional metrics [35]-[36].

E. Identifying and Validating Best Regression Algorithm

After evaluating and comparing various regression algorithms, we will select the most suitable one for predicting fruit and vegetable prices. By analyzing the data, we will identify the top 4 months with the highest prices for specific produce. This valuable information will empower farmers to plan their crop harvesting and cultivation strategically, ensuring they obtain the best possible prices for their products. Additionally, policymakers can benefit from this prediction to make informed decisions and support the agricultural sector. The accuracy of this prediction will be validated against the current market price of tomatoes, ensuring its reliability and usefulness in real-world scenarios.

Next section shows basic experimentation and data analysis results which include analysis of price of tomato and mosambi, Regression Algorithms and prediction of price from best regression algorithms.

IV. Result and Discussion

In this section, we present the predicted prices of fruits and vegetables in the Maharashtra market for the next few year based on our research and analysis. We obtained a dataset of fruit and vegetable prices from Vasai Market, spanning from April 2016 to March 2021. This dataset includes information on 60 different types of fruits and vegetables. Detail about dataset is given in section 3. The predictions aim to provide insights into potential price fluctuations and aid market stakeholders in making informed decisions.

A. Analysis of Price of Tomato and Mosambi From April 2016 to March 2021

In this section, we examine historical pricing patterns for the previous few years for tomatoes and mosambi (sweet lime). The results are based on the dataset discussed in section 3, which shows the rise in particular months and fall in particular months in prices for tomatoes and mosambi.

Fig. 3 shows historical tomato pricing data over the last several years. Different time periods are shown by the x-axis, while matching prices in the Mumbai APMC are shown on the y-axis. Our examination of the data shows a yearly increase tendency in tomato prices in the months of June, July, and August. This increase can be linked to a number of things, including changes in the dynamics of supply and demand, climatic variables that impact growing, the cost of transportation, and competition.

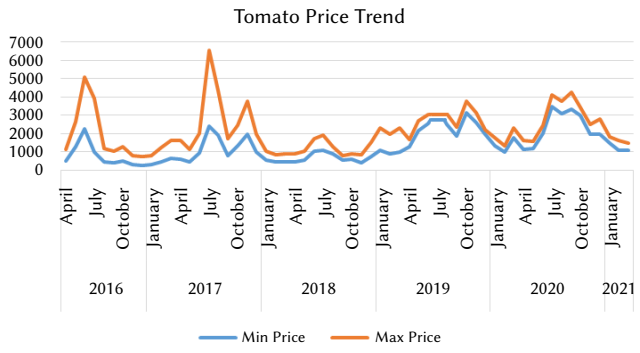

Fig. 3. Tomato Price Trend.

Fig. 4 shows the historical price changes for Mosambi during the given time period. The graph depicts a rising tendency in Mosambi pricing, similar to the pattern in tomato prices. Mosambi prices might rise as a result of market demand, supply chain interruptions, shifting customer tastes, and changes in farming practices. Additionally, outside variables like climatic changes and the dynamics of the worldwide market might have influenced price developments. Consumers, companies, and agricultural stakeholders may be impacted by the steady increase in Mosambi prices. It could have an impact on consumer choices, dynamics of export and import, and fruit growers' financial success.

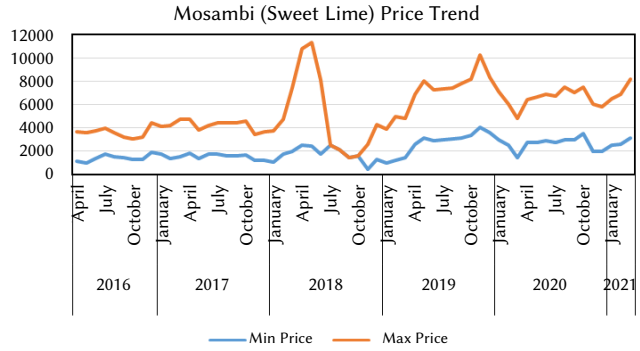

Fig. 4. Mosambi (Sweet Lime) Price Trend.

B. Regression Algorithms

In Section 3 of the research, the dataset preprocessing and regression analysis pipeline were comprehensively detailed. The dataset consists of 60 distinct fruits and vegetables collected over a 5-year period, resulting in a total of 3,600 data entries, with each year contributing 12 rows of data per fruit/vegetable type. First, all null values were removed, resulting in a refined dataset containing 3,115 entries. Following this, the categorical variable "Month" was transformed. Originally represented by the names of the months (e.g., January, February), this variable was converted into numerical ranging from 1 to 12. To handle categorical features effectively during the regression analysis, the one-hot encoding technique was applied. This transformation converted categorical features into binary vectors, making them compatible with various regression algorithms [41]. The dataset was then divided into training and testing sets. Specifically, 70% of the data (2,180 entries) was allocated for training purposes, while the remaining 30% (935 entries) was reserved for testing the trained models. This separation ensures that the models are evaluated on unseen data, providing a more accurate assessment of their generalization performance. To further enhance the robustness of the regression models, k-fold cross-validation was employed during the training process. This technique involves dividing the training dataset into 'k' folds or subsets. The model is then trained 'k' times, each time using a different fold as the testing set and the remaining fold for training. The final performance metric is calculated by averaging the results from each iteration. In this research, we chose a value of k =5. Additionally, to ensure uniformity in the scale of input features, the dataset underwent standardization using the Standard Scaler [42]. This preprocessing step is particularly essential for regression algorithms that are sensitive to variations in the scale of input features, promoting stable and reliable model training across different features. By incorporating k-fold cross-validation and standardization into the training process, the research aims to provide a more robust evaluation of the regression models' performance, accounting for potential variations and improving their generalization capabilities.

All 13 listed regression algorithms were employed, and the training data was used as input for each of them. We have also added 2 more DL algorithm LSTM (Long Short-Term Memory) and GRU (Gate Recurrent Unit). The performance of the models was evaluated using three key metrics: MSE, MAE, and R-square (R^2). These metrics provide insights into the accuracy and goodness-of-fit of the regression models. Table III and Table IV presented the performance metrics of all the algorithms, providing a comprehensive comparison of their predictive capabilities. The results shed light on the strengths and weaknesses of each algorithm in capturing the underlying patterns and relationships in the dataset.

TABLE III. Performance Metrics of All the Algorithms (MSE, MAE, R-Square, RMSE, MPE)

Model	MSE	MAE	R-square (R^2)	RMSE	MPE
LinearRegression()	33.97	44.54	0.56	5.83	-17.89
Ridge()	34.03	44.53	0.56	5.83	-19.84
Lasso()	34.00	44.58	0.56	5.83	19.54
KNeighborsRegressor()	20.81	34.20	0.73	4.56	-21.76
MLPRegressor()	80.62	85.55	0.00	8.98	-40.65
DecisionTreeRegressor()	0.00	0.00	1.00	0.00	0.00
RandomForestRegressor()	3.94	11.53	0.95	1.99	-3.28
LinearSVR()	94.28	100.0	0.00	9.71	-24.48
SVR()	100.0	90.10	NA	10.00	-50.13
GradientBoostingRegressor()	16.55	49.86	0.82	4.07	-39.98
XGBRegressor()	2.67	18.52	0.97	1.63	-7.18
LGBMRegressor()	31.06	45.05	0.67	5.57	-18.91
CatBoostRegressor()	5.17	26.92	NA	2.27	-12.39
LSTM	60.24	20.48	NA	7.76	-18.21
GRU	55.54	20.08	NA	7.45	-17.82

TABLE IV. Performance Metrics of All the Algorithms (MAPE, HL, MSLE, TUS, GINI)

Model	MAPE	HL	MSLE	TUS	GINI
Linear Regression ()	39.23	51.67	0.27	0.31	0.39
Ridge()	39.86	51.67	0.27	0.32	0.39
Lasso()	39.78	51.72	0.26	0.32	0.38
K Neighbors Regressor ()	35.53	39.68	0.18	0.28	0.41
MLPRegressor()	52.89	100.00	0.83	0.98	0.31
Decision Tree Regressor ()	0.00	0.00	0.00	0.00	0.04
Random Forest Regressor ()	8.35	13.05	0.02	0.06	0.04
Linear SVR ()	74.83	82.52	0.67	0.55	0.45
SVR()	88.80	80.29	0.65	0.70	0.76
Gradient Boosting Regressor ()	56.65	47.05	0.28	0.45	0.38
XGBRegressor()	17.38	17.52	0.07	0.14	0.04
LGBMRegressor()	35.94	44.78	0.18	0.29	0.39
CatBoost Regressor ()	26.60	26.63	0.11	0.21	0.42
LSTM	64.26	87.26	0.52	0.47	0.59
GRU	63.85	86.13	0.51	0.43	0.57

For simplicity of comparison, the MAE values have been scaled down to a range of 0 to 100 and the MSE values to a range of 0 to 1000. Fig. 5 shows comparison of the different algorithms, all metrics are scaled down to 0 to 1 in the comparison chart.

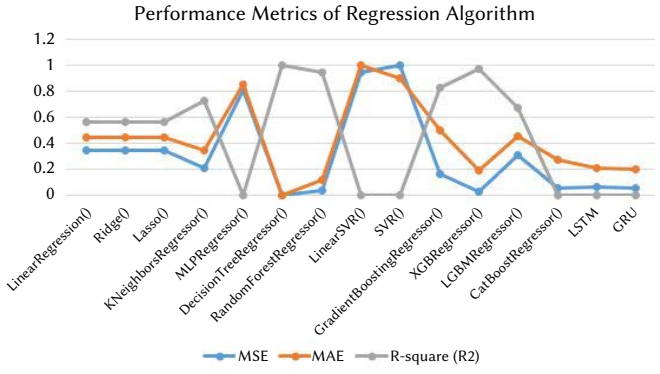

Fig. 5. Performance Metrics of regression Algorithm.

1. Linear Regression (), Ridge (), and Lasso ()

The MSE, MAE, and R2 values obtained from these three linear regression methods were comparable, demonstrating equivalent

performance on the dataset. The models' estimated R2 value of 0.564 indicates that they account for around 56.4% of the variation in the target variable. This indicates a moderate level of predictive power, implying that the models capture a substantial portion of the variability present in the data.

2. KNeighbors Regressor ()

With reduced MSE and MAE and a higher R2 value of 0.728, the KNeighbors Regressor beat the linear regression models, indicating that this algorithm provides a better fit to the data, potentially due to its ability to capture nonlinear relationships and complex patterns.

3. MLP Regressor ()

Dataset was divided in batch size of 64 and total iteration completed by algorithm was 100. In comparison to other models, the MLPRegressor displayed much higher MSE and MAE values, indicating that it underperformed on this dataset. Additionally, the target variable's variation is only partially explained by this model, as seen by the low R2 value of 0.001.

4. Decision Tree Regressor ()

The Decision Tree Regressor performed remarkably well, achieving an MSE and MAE of 0.000, which is likely due to overfitting. The perfect R2 value of 1.000 indicates that this model perfectly fits the data, but it might not generalize well to new data. Let's denote the true target values as y_i and the predicted values by the decision tree as $\hat{y}_i$. The total sum of squares (TSS) represents the total variance in the target variable. TSS is given by equation (26).

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (26)$$

Where $\bar{y}$ is the mean of the true target values. The residual sum of squares (RSS) measures the unexplained variance by the model. RSS is given by equation (27).

$$RSS = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (27)$$

The R-squared value is given by equation (28).

$$R^2 = 1 - \frac{RSS}{TSS} \quad (28)$$

For a perfect fit, (R^2= 1), indicating that the model explains all the variance. Since the Decision Tree Regressor () has an R^2 value of 1.00, it means that it perfectly fits the training data.

5. Random Forest Regressor ()

The Random Forest Regressor produced low MSE and MAE values, indicating good performance. The relatively high R2 value of 0.947 suggests that this model explains a significant portion of the variance in the target variable.

6. Linear SVR () and SVR ()

Both Support Vector Regressor algorithms showed relatively high MSE and MAE values, with very low R2 values. The R2 values, which measure the proportion of variance in the dependent variable that is predictable from the independent variables, were very low. These findings suggest that the SVR models struggled to capture the underlying patterns in the dataset effectively, leading to subpar regression performance.

7. Gradient Boosting Regressor ()

The Gradient Boosting Regressor achieved a reasonable MSE and MAE, along with a relatively high R2 value of 0.825, indicating good performance and a relatively better fit compared to some other models. This performance underscores the effectiveness of Gradient Boosting techniques in handling complex relationships within the data and making accurate predictions.

8. XGB Regressor ()

The XGBRegressor showed a low MSE and MAE, as well as a high R2 value of 0.972, indicating excellent performance and a strong ability to explain variance in the target variable. This level of performance indicates that the XGBRegressor successfully captures complex relationships within the data, providing reliable predictions.

9. LGBM Regressor ()

The LGBMRegressor had a moderate MSE and MAE, along with an R2 value of 0.671, suggesting a reasonable fit to the data. Overall, these results imply that the LGBMRegressor is capturing a significant portion of the underlying patterns in the data.

10.CatBoost Regressor ()

The CatBoost Regressor performed quite well, with relatively low MSE and MAE values, indicating good performance on the dataset. The model's adeptness in capturing intricate relationships within the data translates to accurate predictions. Its boosted decision tree architecture contributes to its prowess by iteratively improving upon weaknesses, enhancing predictive accuracy with each iteration. Overall, the CatBoost Regressor emerges as a formidable choice for tasks requiring precise regression, showcasing its reliability and effectiveness in real-world applications.

11. LSTM

The dataset was divided in batch size of 32 and total iteration completed by algorithm was 50. The LSTM model achieved a mean squared error (MSE) of 60.235 and a mean absolute error (MAE) of 20.480. These relatively low error values suggest that the LSTM model provides reasonably accurate predictions.

12. GRU

The dataset was divided in batch size of 32 and total iteration completed by algorithm was 50. The GRU model also performed well with an MSE of 55.54 and an MAE of 20.081. The low MSE and MAE values indicate that the GRU model is capable of making relatively accurate predictions.

Based on the performance metrics, the Decision Tree Regressor () stands out as the top-performing model, with an MSE and MAE of 0.000, indicating a perfect fit to the data. While this might seem impressive at first glance, it is essential to keep in mind that such a result could be a sign of overfitting, where the model memorizes the training data but fails to generalize to new, unseen data.

The Random Forest Regressor () and XGB Regressor are strong contenders for a more balanced selection. Both models exhibited quite low MSE and MAE values, indicating strong performance and some degree of prediction accuracy. The strong R2 values demonstrated their ability to fit the data accurately by explaining a substantial amount of the variance in the target variable.

When examined more closely, the XGBRegressor emerges as a highly advantageous choice. It demonstrated the best accuracy and precision in predicting the target variables (max price and min price) with the lowest MSE and MAE values after Decision Tree Regressor among all other models. Furthermore, the XGBRegressor proves to be a solid option for this regression task, evident from its high R2 value of 0.972, indicating that it can explain a sizable percentage of the variance in the target variable.

Considering the overall performance and the balance between accuracy and generalization, the XGBRegressor appears to be the most appropriate regression method for this particular dataset and research challenge. It stands out as the recommended option due to its capacity to precisely anticipate the target variables while still maintaining good generalization.

Overall, the findings presented valuable insights for understanding the predictive power of various models, facilitating better decision-making and potential applications in real-world scenarios.

C. Predicting Values From Best Regression Algorithm

In this study, we employed three top-performing algorithms, XGBRegressor, Random Forest, and Decision Tree, to predict tomato prices for the year 2023. Our proposed method effectively identified the four most lucrative months in terms of market prices. This information can greatly assist farmers in making informed decisions about the optimal timing for tomato harvesting. The results presented in Table V highlight the months with the highest price potential, empowering farmers to maximize their profits and optimize crop cultivation strategies.

TABLE V. Tomato Price Prediction

Sr. No.	XGBRegressor		Random Forest		Decision Tree	
	Month	Price	Month	Price	Month	Price
1	9	3138.57	9	3906.23	9	4258
2	7	3133.95	8	3647.96	7	4084
3	10	3131.60	7	3560.89	8	3744
4	8	2942.82	10	3488.16	10	3400

Result in Table V shows that all three algorithm predicted the same highest four months in which farmer can get more profit. This means that prediction is good by comparing best three regression algorithm. This implies that our prediction is robust. Additionally, Table VI shows the average of the predicted prices for these four months, providing valuable insights for farmers to capitalize on potentially more profitable periods.

TABLE VI. Predicted Price Trends: Average of Top 3 Algorithm

Month	Average
7	3592.95
8	3444.93
9	3767.60
10	3339.92

To validate the prediction that tomatoes became expensive in India from the month of July, we refer to the publication by Biswas [43]. The article titled "Tomato prices are on fire — and will not come down soon. Here is why" by P. Biswas in The Indian Express highlights the surge in tomato prices and the reasons behind it. According to the article published on June 29, 2023, the cost of tomatoes has witnessed a significant increase and is expected to remain high for an extended period.

To assess the adaptability (flexibility in adjusting to new data and maintaining performance) of the proposed method, we examined the results for tomatoes using real-time data from 2023, sourced from the Annual Report of the "Agricultural Produce Market Committee, Pune (Krushi Utpanna Bazar Samiti, Pune)" [44]. The maximum rate for tomatoes in July was 4500 INR, with an average rate of 3800 INR. In August, the maximum rate was 4000 INR, with an average rate of 3600 INR. For September, the maximum rate remained 4000 INR, with an average rate of 3500 INR. These results demonstrate that our proposed approach is adaptable.

These findings can empower stakeholders in the market, particularly farmers, to make informed decisions on optimal harvesting and crop cultivation strategies, maximizing their profits during these high-price periods. Overall, our research contributes valuable insights for better decision-making in the fruit and vegetable market in Maharashtra.

D. Future Scope

In terms of future prospects, our research aims to develop a farmer-centric application specifically tailored for predicting and assisting in crop planning within Maharashtra. This application will be continuously updated with the latest datasets, ensuring that farmers have access to the most relevant and accurate information for decision-making. Fig. 6 shows how application should be developed and work.

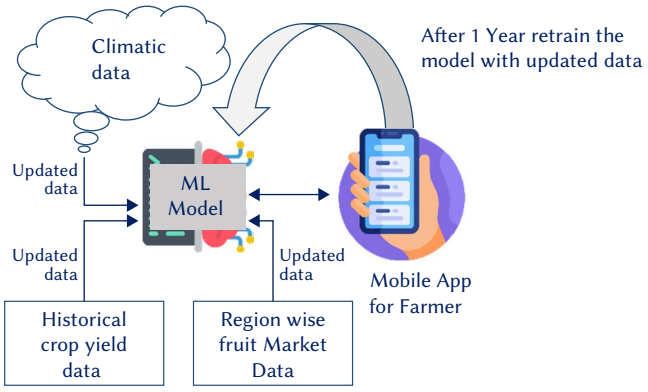

Fig. 6. Future working of Mobile Application for farmer.

By extending the scope of future application to encompass not just Maharashtra, but the entire nation of India and eventually the worldwide farming community, we have the thrilling chance to transform agricultural methods on a global level. Our future goal is to use cutting-edge technology and data-driven analysis to give farmers from different countries the tools and information necessary to enhance their agricultural operations and increase crop yields.

Moreover, a potential direction for future study and development is to improve the prediction capabilities of our algorithm by integrating supplementary features into the Regressor model. To enhance the accuracy and usefulness of the insights supplied to farmers, we may incorporate parameters such as weather conditions, soil nutrient levels, and geographical variances.

However, it is important to confront certain challenges, such as apprehensions over privacy and confidentiality, particularly when handling sensitive agricultural data. Considering that the data may be spread out across several sources and places, it becomes crucial to apply privacy-preserving mechanisms. An effective strategy to address these issues is by implementing federated learning, which involves training the model collectively using decentralized data sources while ensuring data privacy is maintained. Studying and analyzing the viability of such methods will be a crucial component of future research.

The future scope of our research encompasses the development of a comprehensive, farmer-centric application for predictive analytics and crop planning, with a focus on scalability, accessibility, and privacy preservation. Through the utilization of cutting-edge technology and inventive approaches, our goal is to make substantial contributions to the worldwide agricultural community and enable farmers to succeed in a progressively intricate and ever-changing environment.

1. Limitation of Research – Future Work

In future study, it is necessary to solve various limitations.

- Quality and availability of data: Gathering large and high-quality agricultural data, particularly in rural regions, poses a significant challenge.
- Privacy and Confidentiality: Safeguarding sensitive farmer data while ensuring its utility for analysis presents legal and ethical hurdles.

- Establishing a balance between complex prediction models and easily understandable interpretations is essential for building trust and comprehension among farmers.
- Scalability and generalization: Validating and adapting models to account for regional variances is necessary when extending them from local to global dimensions.
- Giving appropriate access to advanced technology such as cloud computing and mobile applications continues to pose financial and logistical difficulties.
- Encouraging widespread adoption among farmers and stakeholders requires overcoming challenges such as technology literacy and cultural disagreement.

To tackle these difficulties, it is necessary to have collaboration between different fields of study, use creative approaches, and maintain continuous involvement with farming communities.

V. Conclusion

The influence of technology is driving significant changes in every field worldwide, including the agricultural sector in India. To bolster its development and growth, the Indian farming industry requires more technological support. Accurate price prediction of agricultural products is crucial to ensure fair returns for farmers and to help them recover their investments. Our proposed method offers a valuable framework for predicting fruit and vegetable prices in the Maharashtra market, leveraging various ML and DL algorithms. This approach provides critical insights for decision-making in the Indian farming sector, empowering farmers and policymakers with data-driven cultivation strategies, distribution optimization, and effective marketing. The analysis and evaluation of several regression algorithms revealed XGBRegressor, Random Forest, and Decision Tree as the most suitable models, boasting high R2 scores close to 1 and low MSE and MAE. Future prospects include creating a farmer-centric application for forecasting and crop planning in Maharashtra, with regularly updated datasets. Scaling the application to cover India and the world represents an exciting opportunity to revolutionize global farming practices and benefit farmers across borders.

Data Availability

The data used for this study are available from the authors on request.

Conflicts of Interest

The authors have nothing to declare as conflicts of interest.

Funding Statement

The authors did not receive any financial support for conducting the research, writing the article, or publishing it.

References

[1] Q. Zhang, "Opinion paper: Precision agriculture, smart agriculture, or digital agriculture," *Computers and Electronics in Agriculture*, vol. 211, pp. 107982, 2023. doi:10.1016/j.compag.2023.107982.

[2] Y. Huang, Q. Zhang, "Agricultural Cybernetics", *Springer: Berlin/Heidelberg, Germany*, 2021.

[3] Dvara Research, "Why don't Indian farmers grow more fruits and vegetables?," *Dvara Research Blog*. Jan. 30, 2013 [Online]. Available: https://www.dvara.com/research/blog/2013/01/30/why-dont-indian-farmers-grow-more-fruits-and-vegetables/

[4] J. Cheruku and V. Katekar, "Digitalisation of Agriculture in India: The case for doubling farmers' income," *Indian Institute of Public Administration*, pp. 194-205, 2023.

[5] M. Vibas and A. R. Raqueño, "A Mathematical Model for Estimating Retail Price Movements of Basic Fruit and Vegetable Commodities Using Time Series Analysis," *International Journal of Advance Study and Research Work*, vol. 2, no. 7, pp. 1–5, 2019. doi: 10.5281/zenodo.3333529.

[6] S. Rakhal and C. Brianne, "Price Transmission in Canadian Fresh Fruit Market: A Time Series Analysis", *International Journal of Food and Agricultural Economics (IJFAEC)*, vol. 9, no. 3, pp. 175-189, 2021. doi: 10.22004/ag.econ.313363.

[7] A. Jahangir, K. Jyoti, B. Deep Ji, and B. Anil, "Analysis of Prices and Arrivals of Apple Fruit in Narwal Market of Jammu", *Economic Affairs*, vol. 63, no. 1, pp. 107-111, March 2018, doi: 10.30954/0424-2513.2018.00150.13

[8] L. Nassar, I. E. Okwuchi, M. Saad, F. Karray and K. Ponnambalam, "Deep Learning Based Approach for Fresh Produce Market Price Prediction," *2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK*, 2020, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9207537.

[9] I. Okwuchi, "Machine Learning based Models for Fresh Produce Yield and Price Forecasting for Strawberry Fruit," *M.S. thesis, Univ. of Waterloo*, 2020. [Online]. Available: http://hdl.handle.net/10012/15976.

[10] R. Agarwal and Prof. P. Sagar, "A Comparative Study of Supervised Machine Learning Algorithms for Fruit Prediction", *Journal of Web Development and Web Designing*, vol. 4, no. 1, pp. 14–18, Apr. 2019, doi: 10.5281/zenodo.2621205.

[11] R. Dharavath and E. Khosla, "Seasonal ARIMA to Forecast Fruits and Vegetable Agricultural Prices," *2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Rourkela, India*, 2019, pp. 47-52, doi: 10.1109/iSES47678.2019.00023.

[12] C. Sharma, R. Misra, M. Bhatia and P. Manani, "Price Prediction Model of fruits, Vegetables and Pulses according to Weather," *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India*, 2023, pp. 347-351, doi: 10.1109/Confluence56041.2023.10048880.

[13] M. Kankar and M. A. Kumar, "Price Prediction of Agricultural Products Using Deep Learning," *Advanced Machine Intelligence and Signal Processing*, D. Gupta, K. Sambyo, M. Prasad, and S. Agarwal, Eds. *Singapore: Springer*, 2022, vol. 858, *Lecture Notes in Electrical Engineering*, pp. 495-506. doi: 10.1007/978-981-19-0840-8_38.

[14] R. K. Paul, M. Yeasin, P. Kumar, P. Kumar, M. Balasubramanian, H. S. Roy, et al., "Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India," *PLoS ONE*, vol. 17, no. 7, p. e0270553, Jul. 2022, doi: 10.1371/journal.pone.0270553.

[15] C. Chai, J. Wang, Y. Luo, Z. Niu and G. Li, "Data Management for Machine Learning: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 4646-4667, 1 May 2023, doi: 10.1109/TKDE.2022.3148237.

[16] Z. Luo, C. Fang, C. Liu and S. Liu, "Method for Cleaning Abnormal Data of Wind Turbine Power Curve Based on Density Clustering and Boundary Extraction," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 2, pp. 1147-1159, April 2022, doi: 10.1109/TSTE.2021.3138757.

[17] F. Ridzuan and W. M. N. W. Zainon, "A Review on Data Cleansing Methods for Big Data," *Procedia Computer Science*, vol. 161, pp. 731-738, ISSN 1877-0509, 2019, doi: https://doi.org/10.1016/j.procs.2019.11.177.

[18] Y. Nieto, V. García-Díaz, C. Montenegro, C. C. González and R. González Crespo, "Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions," *IEEE Access*, vol. 7, pp. 75007-75017, 2019, doi: 10.1109/ACCESS.2019.2919343.

[19] D. P. Kumar, T. Amgoth, and C. S. R. Annavarapu, "Machine learning algorithms for wireless sensor networks: A survey," *Information Fusion*, vol. 49, pp. 1-25, 2019, doi: 10.1016/j.inffus.2018.09.013.

[20] Y. Nieto, V. García-Díaz, C. Montenegro, et al., "Supporting academic decision making at higher educational institutions using machine learning-based algorithms," *Soft Computing*, vol. 23, no. 12, pp. 4145-4153, 2019, doi: 10.1007/s00500-018-3064-6

[21] M. Ganesan, A. Suruliandi, S. P. Raja, and E. Poongothai, "An Empirical Evaluation of Machine Learning Techniques for Crop Prediction," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 96-104, 2023, doi: 10.9781/ijimai.2022.12.004.

[22] T. Ivanovski, G. Zhang, T. Jemrić, M. Gulić and M. Matetić, "Fruit

firmness prediction using multiple linear regression," *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia*, 2020, pp. 1306-1311, doi: 10.23919/MIPRO48935.2020.9245213.

[23] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning", *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 140-147, Dec. 2020. doi: 10.38094/jastt1457

[24] A. Tsigler and P. L. Bartlett, "Benign overfitting in ridge regression," *Journal of Machine Learning Research*, vol. 24, no. 123, pp. 1-76, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-1398.html

[25] H. Xu, C. Caramanis and S. Mannor, "Robust Regression and Lasso," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3561-3574, July 2010, doi: 10.1109/TIT.2010.2048503.

[26] M. Kück and M. Freitag, "Forecasting of customer demands for production planning by local k-nearest neighbor models," *IEEE Transactions on Engineering Management*, vol. 231, p. 107837, 2021, ISSN: 0925-5273, doi: 10.1016/j.ijpe.2020.107837

[27] I. N. Yulita, A. S. Abdullah, A. Helen, S. Hadi, A. Sholahuddin, and J. Rejito, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java," *Journal of Physics: Conference Series*, vol. 1722, no. 1, p. 012021, Jan. 2021. doi: 10.1088/1742-6596/1722/1/012021.

[28] H. Luo, F. Cheng, H. Yu and Y. Yi, "SDTR: Soft Decision Tree Regressor for Tabular Data," *IEEE Access*, vol. 9, pp. 55999-56011, 2021, doi: 10.1109/ACCESS.2021.3070575.

[29] E. Pekel, "Estimation of soil moisture using decision tree regression," *Theoretical and Applied Climatology*, vol. 139, no. 3, pp. 1111–1119, Mar. 2020, doi: 10.1007/s00704-019-03048-8.

[30] H. Wang, Q. Yilihamu, M. Yuan, H. Bai, H. Xu, and J. Wu, "Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: A comparison of regression and random forest," *Ecological Indicators*, vol. 119, p. 106801, 2020. doi: 10.1016/j.ecolind.2020.106801.

[31] M. Alida and M. Mustikasari, "Rupiah Exchange Prediction of US Dollar Using Linear, Polynomial, and Radial Basis Function Kernel in Support Vector Regression," *Jurnal Online Informatika*, vol. 5, no. 1, pp. 53-60, 2020. doi: 10.15575/join.v5i1.537

[32] C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," *2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India*, 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.

[33] Zhagparov, Z. Buribayev, S. Joldasbayev, A. Yerkosova and M. Zhassuzak, "Building a System for Predicting the Yield of Grain Crops Based On Machine Learning Using the XGBRegressor Algorithm," *2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), Nur-Sultan, Kazakhstan*, 2021, pp. 1-5, doi: 10.1109/SIST50301.2021.9465938.

[34] A. Thaniserikaran, B. Sriphani Vardhan, A. Rahman Mateen Syed, M. Abdul Muqeet, A. Khot and B. K. Tejas, "The prediction of cern electron mass collision by using CATBoosting and LGBMR," *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India*, 2022, pp. 1-5, doi: 10.1109/ICCCNT54827.2022.9984588.

[35] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, pp. 045-076, 2019. doi: 10.28945/4184

[36] Chicco D, Warrens MJ, Jurman G. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." *PeerJ Computer Science*, vol. 7, p. e623, 2021.doi: 10.7717/peerj-cs.623

[37] Q. Sun, W. Zhou, and J. Fan, "Adaptive Huber Regression," *in Journal of the American Statistical Association*, vol. 115, no. 529, pp. 254-265, 2020. doi: 10.1080/01621459.2018.1543124

[38] M. Eppert, P. Fent, and T. Neumann, "A Tailored Regression for Learned Indexes: Logarithmic Error Regression," *in Fourth Workshop in Exploiting AI Techniques for Data Management (aiDM '21), Virtual Event, China*, 2021, pp. 9-15, doi: 10.1145/3464509.3464891

[39] L. F. Tratar and E. Strmčnik, "The comparison of Holt–Winters method and Multiple regression method: A case study," *Energy*, vol. 109, pp. 266-276, 2016. [Online]. Available: https://doi.org/10.1016/j.energy.2016.04.115

[40] S. Mirzaei, G.M. Borzadaran, M. Amini, and H. Jabbari, "A comparative study of the Gini coefficient estimators based on the regression approach," *Communications for Statistical Applications and Methods*, vol. 24, no. 4. *The Korean Statistical Society*, pp. 339–351, 31-Jul-2017. doi:10.5351/csam.2017.24.4.339.

[41] P. Rodríguez, M. A. Bautista, J. Gonzalez, and S. Escalera, "Beyond one-hot encoding: Lower dimensional target embedding," *Image and Vision Computing*, vol. 75, pp. 21-31, Apr. 2018. Doi: 10.1016/j.imavis.2018.04.00

[42] E. Bisong and E. Bisong,"Introduction to Scikit-learn," *in Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, 2019, pp. 215-229. Doi: 10.1007/978-1-4842-4470-8_18

[43] P. Biswas, "Tomato prices are on fire — and will not come down soon. Here is why," *The Indian Express*, Online, June 29, 2023. [Accessed: July 20, 2023]. Available: https://indianexpress.com/article/explained/explained-economics/why-tomato-prices-high-8689168/

[44] Agricultural Produce Market Committee, Pune, "Annual Report of 2023 Agricultural Produce Market Committee, Pune", [Online] http://www.puneapmc.org/rates.aspx [Last Accessed: 30/04/2024].

Dr. Nilesh P. Sable

Dr. Nilesh P. Sable is a senior member IEEE and working as Associate Professor, Head Department of Computer Science & Engineering (Artificial Intelligence) at Vishwakarma Institute of Information Technology, Pune, India. He has completed his Ph.D. in Computer Science & Engineering from Kalinga University, Raipur. He has 16 + years of teaching and research experience. He is guiding 4 Ph.D. students in the area of Machine Learning, Federated Learning and IoT under his supervision from SPPU. He is working as Research Advisory Committee (RAC) Member for various Research Centres. He is a reviewer for various journals and conferences of the repute. He has published 75+ papers in National, International conferences and Journals. He had Filed and Published 16 Patents and 18 Copyrights. He has authored books published by National/International publishers. He is also the recipient of the "Distinguished Performance Award" by Vishwakarma Institute of Information Technology. He has delivered 50+ lectures at national and international level.

Rajkumar V. Patil

Rajkumar received his Bachelor of Engineering (Computer Engineering) in 2018 from Savitribai Phule Pune University, Pune, India, and his Master of Engineering in Computer Engineering in 2020 from Smt. Kashibai Navale College of Engineering, Pune, India. Cyber Physical systems for healthcare, Trust management, Machine learning, Web technologies, Algorithms and Explainable AI are his areas of interest. He has around 3 years of teaching and research experience. He has published research articles and book chapters in international journals. He is reviewer for several international journals. He is presently working as assistant professor at MIT Art, Design and Technology University, Pune and research scholar at Vishwakarma Institute of Information Technology Pune.

M. Deore

M. Deore is working as an Asst. Professor in Computer Engineering Department at MKSSS's Cummins College of Engineering for Women, Pune 411051, India. He was awarded his Master of Technology Degree from Bharati Vidyapeeth Deemed University College of Engineering, Dhankawadi, Pune. He received doctoral degree from Swami Ramanand Teertha Marathwada University, Nanded, India in 2020. He has 16 + years of teaching and research experience. He is a reviewer for various journals and conferences of the repute. He has published 25+ papers in National, International conferences and Journals. His areas of interest are big data, Security, Computer Networks and Machine learning. He has Fourteen years' experience in teaching.

Ratnmala Bhimanpallewar

Ratnmala Bhimanpallewar holds a Doctor of Philosophy degree in Computer Science and Engineering from K L University, Vijaywada, India. She has received her master's (M.E. Computer Science and Engineering) degree from PICT, Savitribai Phule Pune University, Pune, India. She is working as an Assistant Professor in the Information Technology Department of Vishwakarma Institute of Information Technology, Kondhwa (Bk.), Pune. She has 14 years of working experience. Her area of interest is Databases, Machine Learning and IoT She is a lifetime member of ISTE. She has completed the funded research project under SPPU ASPIRE scheme.

Parikshit N. Mahalle

Dr Parikshit is a senior member IEEE and is Professor, Dean Research and Development and Head - Department of Artificial Intelligence and Data Science at Vishwakarma Institute of Information Technology, Pune, India. He completed his Ph. D from Aalborg University, Denmark and continued as Post Doc Researcher at CMI, Copenhagen, Denmark. He has 23 + years of teaching and research experience. He is an ex- Board of Studies, Ex-Chairman at various Universities and autonomous colleges across India. He has 15 patents, 200+ research publications and authored/edited 60+ books with Springer, CRC Press, Cambridge University Press, etc. He is editor in chief for IGI Global –International Journal of Rough Sets and Data Analysis, Inter-science International Journal of Grid and Utility Computing, member-Editorial Review Board for IGI Global – International Journal of Ambient Computing and Intelligence and reviewer for various journals and conferences of the repute. His research interests are Machine Learning, Data Science, Cognitive Computing, Algorithms, Internet of Things, Identity Management and Security. He is guiding 8 PhD students in the area of IoT and machine learning and 6 students have successfully defended their PhD under his supervision from SPPU. He is also the recipient of the "Best Faculty Award" by Sinhgad Institutes and Cognizant Technologies Solutions and State Level Meritorius Teacher Award. He has delivered 200 + lectures at national and international level.