Journal of Mines, Metals and Fuels, 71(6): 804-810; 2023. DOI: 10.18311/jmmf/2023/34497

Print ISSN : 0022-2755

Journal of Mines, Metals and Fuels

Contents available at: www.informaticsjournals.com/index.php/jmmf

Predicting Prices of Cash Crop using Machine Learning

Vaidehi Bhaskara[1*], Ramesh K T[1] and Sayan Chakraborty[2]

[1]Department of Industrial Engineering and Management, B.M.S College of Engineering, Bengaluru, India.
E-mail: vaidehi.im18@bmsce.ac.in
[2]Department of Operations and IT, ICFAI Business School, Hyderabad, India

Abstract

More than half of the Indian population depends on agriculture as a source of livelihood. But, India's marginal farmers especially, earn meagre amounts from their harvested yields. This may be attributed partly due to relatively smaller land holdings, and partly due to minimal access to resources that aid with informative price forecasts. In order to alleviate the stress caused by the lack of sound financial planning, this research proposes the utilization of machine learning to predict commodity prices. The solution obtained through such a model would assist farmers in predicting the price and associated estimates can be made with respect to yield, sowing patterns and suitable recommendations for sales. The solution developed in this research is a result of a thorough exploration of the literature in this domain, identification of verified secondary sources for data collection, and proposes a methodology to design a machine learning model that predicts prices for seasonal cash crops specific to the markets of Karnataka. Cotton has been used as the crop of focus in this study. ARIMA and Bayesian ridge regression have been used for predictive analytics, and the results obtained indicate a high correlation between the predicted and actual.

Keywords: *Agriculture, Cash crop, Price prediction, Machine learning, Regression*

1.0 Introduction

Indian agriculture faces many challenges in the overall development towards improving welfare of its rural poor like increasing agriculture productivity per unit of land, reducing rural poverty through social strategy comprising both agriculture and non-farming employment, and making sure that agricultural growth responds to food security needs[1]. This chapter provides an overview of the context of agriculture in India.

1.1 Agriculture in India

India is the second most populous country in the world with a population of around 1.38 billion. Agriculture has been the source of livelihood. Around 70% of rural households depend on agriculture. India is the largest producer of pulses, milk, and spices. Along with these, rice, cotton, tea, sugarcane, wheat and various fruits, vegetables are also majorly produced. It is seen that around 195mha is used for cultivation of which 63% is rain fed and the remaining 37% is irrigated[2].

Over time, as the Indian economy has diversified, the country's overall GDP has been declining indicating growing concerns. Even though India houses leading food producers, many people in India suffer from hunger: around 190 million people are undernourished. As per the Global Hunger Index- 2021, it was seen that India ranked 101 among 116 countries[3]

[*]Author for correspondence

1.2 Lack of Sound Financial Resources: A major problem faced by Marginal Farmers in India

Marginal farmers are those farmers who own up to one hectare (~2.5 acres) of cultivable land[4]. The state of small and marginal farmers in India is very shocking. While 82% of Indian farmers fall into either one of these two categories, their land holdings account for only 44% of the total. In fact, one-third of India's farming community owns 0.4 hectares of cultivable land or less[5]. It is unfortunately ironic that while these farmers account for a larger proportion of working Indians, their incomes are meager and their contribution to the GDP is only around 18%[6].

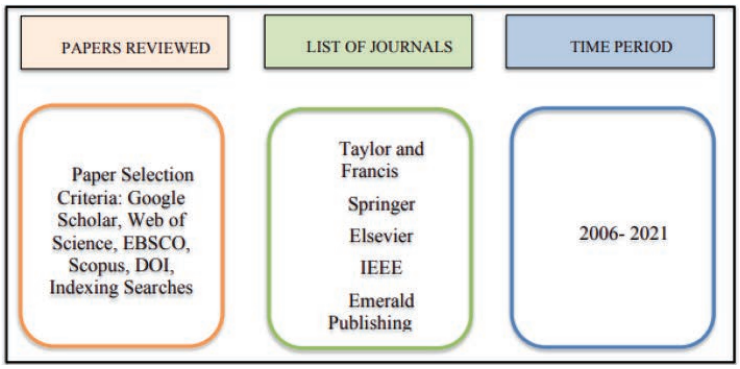

Figure 1: Literature review overview

The most obvious adversity is that of debt and financial standing. Marginal farmers tend to lack the services that aid in marketability of their produce, eventually leading to lower margins earned per yield. This contributes to an accumulation of debt and inability to repay it, driving them to lower standards of living and agonizingly worse off though common- ending their lives. Lack of sufficient income also prohibits farmers from growing in the agricultural space: it prevents them from getting access to quality products like seeds, farming equipment, fertilizers, adequate market forecasts, etc.[7]

Middlemen play a major role in exploiting farmers. Due to misery, when farmers are left with no choice but to approach middlemen, it seems like even the slightest of their earnings are robbed away. In a similar extension, the government policies and farmer – benefit programmes do not reach all farmers, particularly the marginal farmers who require the most assistance. Various schemes, loan and credit card policies introduced by the Indian government have unfortunately failed to reach the grass root levels.

Crop price prediction thus plays a critical role towards justifying minimum support prices and pricing strategies based on the crop yield, to provide farmers with an estimate of the worth of their crops. By factoring the potential yield and demand for any given crop, farmers can manage their finances and plan suitable growing recommendations to maximize returns.

2.0 Literature Review

A systematic literature review of various publications obtained from search engines, extending from 2006-2021, was adopted to learn the applications of decision sciences, data analytics, forecasting and machine learning in the specific context of agriculture. An overview of this research's literature review is shown in Figure 1.

Key findings observed from the extensive review have been summarized in this section. A report published by the NITI Aayog, Government of India[8] emphasizes the impact of policy decisions, specifically, providing the Minimum Support Price (MSP) for crops on three critical societal goals: (a) ensuring that the producers are protected against price fluctuations, (b) influencing farmers to grow the produce of customers' choice and (c) protecting end customers against surge prices. With respect to price prediction specifically, Cenas, 2017[9] proposes a combination of ARIMA and Kalman Filter algorithms. In order to bring out a holistic view of the historic and current prices, ARIMA was used where the trend and seasonal indices were factored. Following this, the price prediction was carried out using the ARIMA and Kalman Filter ML models. Applications of machine learning algorithms for crop price prediction especially addressed towards the small and marginal farmers in India is discussed extensively[10]. Decision tree-based classifiers and ANN were employed to carry out this analysis based on the modal prices of commodities in Odisha from India's Agmarknet portal. ANN was selected because of its ability to provide evidence for the given forecast, which in this case corresponds to predictors justifying the change in direction in a forecast. Basso & Liu[11] focuses on crop yield forecasting. Remote sensing data from IOT and sensors was extracted for primary data drivers being temperature, precipitation, humidity and soil maps. This data was used to carry out a linear regression analysis. Vohra et al[12.] primarily proposes a decision-making assistance model for agricultural commodity price prediction. This article also offers agricultural data mining approaches that will assist farmers in forecasting agricultural commodity prices. The goal of this work is to create a system with enhanced efficiency of price prediction characteristics. Taking into account the historical crop prices, historical weather data and data quality related features, Jain et al.[13], demonstrates the efficiency of Tomato and Maize in fourteen market places in India. The proposed models not only produce accurate results but end up with a more robust model. Sabu & Kumar[14] acknowledges the psychological trauma that farmers face due to fluctuating prices, both monetarily and mentally. By addressing risk mitigation in

times of such price fluctuation, along with climatic uncertainty, this paper highlights the significance of predictive analytics for price prediction of Areca Nuts in Kerala, India. The RMSE is used as the evaluation metric for a variety of ML models: SARIMA, Holt-Winter's Seasonal Method and LSTM neural networks, for a ten-year period, of which the LSTM model displayed the best results. Prassanna et al.[15], carries out a comparative study between random forest and decision tree regressor, and the performance of each has been evaluated. In Mgale et al.[16], ARIMA and Holt-Winters Exponential Smoothing models were used to forecast the price of rice in Mbeya, Tanzania. Time series data was used for both the trend and seasonal variation, and a comparative study was made between them. An ARIMA model was designed[17] to predict annual prices of onions in the market of Kayamkulam, Kerala. Modal prices of large varieties of onions in this region were collected from data.gov.in. ARIMA eliminates the non-stationarity of data used to give a prediction of an existing time series based on past values. Also, it reveals the hidden trend of the moving average pattern. It was found that there exists a close relationship between actual and obtained values. In Kalichkin[18], predictive decision tree model was developed to predict the yield of spring wheat. Some qualitative factors like level of intensification, tillage system and quantitative factors like temperature and precipitation were taken into account. The prices for wheat crops were predicted using two regression algorithm techniques-decision trees and random forests[19]. With the successful prediction, it was seen that the obtained result was 92% accurate and this prediction could be further enhanced to around 97% by considering more features in addition to rainfall like temperature, humidity and soil fertility. Chen et al., discusses the added benefit of a price prediction model for planning the plantations[20]. The configuration of each of these models is described. The non-linearity of these datasets is considered while using these ML techniques.

As it can be observed from this thorough literature review, there has been a pressing focus on integrating smart technology with agriculture for more than a decade. Based on the findings from this literature review and the extent of data available from verified sources, this research has been carried out and a model has been developed.

3.0 Problem Definition

Agriculture contributes 17 per cent of India's GDP. It is seen that almost around 80 per cent of the farmers are small and marginal farmers, owning less than 2 hectares of land[2]. Income of such farmers is usually from the profit they make out of their produce, because of which they generally tend to sell their commodities at high prices. However, as these prices reach unreasonably high levels, it is seen that there is an increase in the risk like leftovers of commodities and unsold produce which are unfit for consumption. This brings farmers back to square one, of selling their commodities at meager prices to avoid cost of facilities like cold storage, market price forecasting, etc. In order to assist farmers with feasible price prediction solutions, this research proposes a machine learning model to predict crop prices of seasonal cash crops, specifically cotton, in the markets of Karnataka.

3.1 Objectives

The overarching aim of this study is to contribute towards research intended to promote sustainable agriculture which would aid in rural development and help the farmer communities, especially in India, by designing and developing an intelligent model that would aid in the price prediction of seasonal crops in order to assist the agriculture supply chain in decision making towards minimizing risk caused by crop price fluctuations. Such a model would also aid in financial planning, and make crop prices transparent to the farmer well in advance before the season, thereby assisting them with suitable crop planning and recommendations.

4.0 Methodology

The methodology followed in this research is explained as follows:

Step 1: Identify the area of focus: It is a challenge for farmers to estimate the price of their commodities well in advance. Due to the plight faced by marginal farmers with respect to financial instability, the chosen area of focus was brainstormed to be central towards providing intelligent models for sound financial and crop planning.

Step 2: Exhaustive Literature Review: Research articles, papers and policy studies from reputed journals and authentic government portals, were extensively studied and analyzed to aid with defining the problem and its scope for this project.

Step 3: Problem Definition: The problem was specifically defined in this step with respect to the four main pointers of predicting prices of seasonal cash crops in the markets of Karnataka using a suitable ML algorithm.

Step 4: Data Collection: For any comprehensive study, data has to be gathered from authentic sources. For this research, government approved sources like Krishi Marata Vahini and Agmarknet portals to name a few, were used to collect data based on which the type of crop and ML technique may be selected.

Step 5: Data Preprocessing: Data obtained from aforementioned data sources is raw and includes many null values and few outliers. Hence, pre-processing techniques are

required to filter empty and irrelevant observations to fit the existing data. In this step, data is cleaned, normalized and checked if all the data types are valid.

Step 6: Designing the ML Model: An ML model is to be designed based on which crop prices prediction is done for the selected crop. Based on the kind of data extracted on completing Step 4, an appropriate ML algorithm was selected, specifically ARIMA and Bayesian Ridge Regression.

Step 7: Data Analysis: Data exploration also known as E.D.A (Exploratory Data Analysis) is to be carried out, where based on the data model designed for a given crop type, trend analysis of the designed model is analyzed.

Step 8: Model Training and Testing: Data is split into fractions for training and testing. The data is fed to the designed model to train the model. Accuracy of the model is tested using various performance metrics like. RMSE, MSE, MAPE, etc. If the error obtained is greater than the predetermined threshold value, the model is made to train again. Else, the obtained model is used for predictions.

Step 9: Prediction and Validation: Results are to be tested and predictions are to be made. Validation is done by comparing the predicted crop price result with the actual market price data.

Step 10: Conclusion and Inferential Analysis: Conclusions were drawn from the obtained results. An inferential analysis was carried out on the findings from this model. As future work, other ways to enhance this model may include solving problems with respect to yield prediction, recommendations, and sowing patterns.

Following an exhaustive literature review and defining the problem, data for cotton in the markets of Haveri, Karnataka was collected from a verified and Indian- government approved secondary source: the Indian Agmarknet portal. Data obtained from this source is raw and includes many null values and few outliers. Hence, preprocessing techniques are required to filter empty and irrelevant observations to fit the existing data. In this step, data was cleaned, normalized and checked using ARIMA for a two-day period. To check for stationarity, the ADF

test was used along with an analysis of rolling mean as a required prerequisite for time series analysis. Bayesian Ridge Regression has been selected as the technique for prediction, and the ML model was split into a 0.75- 0.25 training- testing ratio. The metrics used to test the quality of output in this study include the mean absolute error (MAE), coefficient of determination (R2) and the mean square error (MSE). The predictions were then made accordingly. Associated conclusions were drawn from the obtained results. An inferential analysis is carried out on the findings from this model. As future work, other ways to enhance this model may include solving problems with respect to yield prediction, recommendations, and sowing patterns.

5.0 Results

Cotton is a seasonal kharif cash crop in India[21]. Specific to Karnataka, Haveri district has the highest area of cotton under cultivation, close to 1,11,550 hectares[22]. This study collected data of cotton prices from the markets of Haveri over a 4-year period from January 2017 to December 2021. A 2- day moving average was used for imputation to account for missing values. An initial forecast analysis was carried out using a 95% confidence, and the results are displayed in

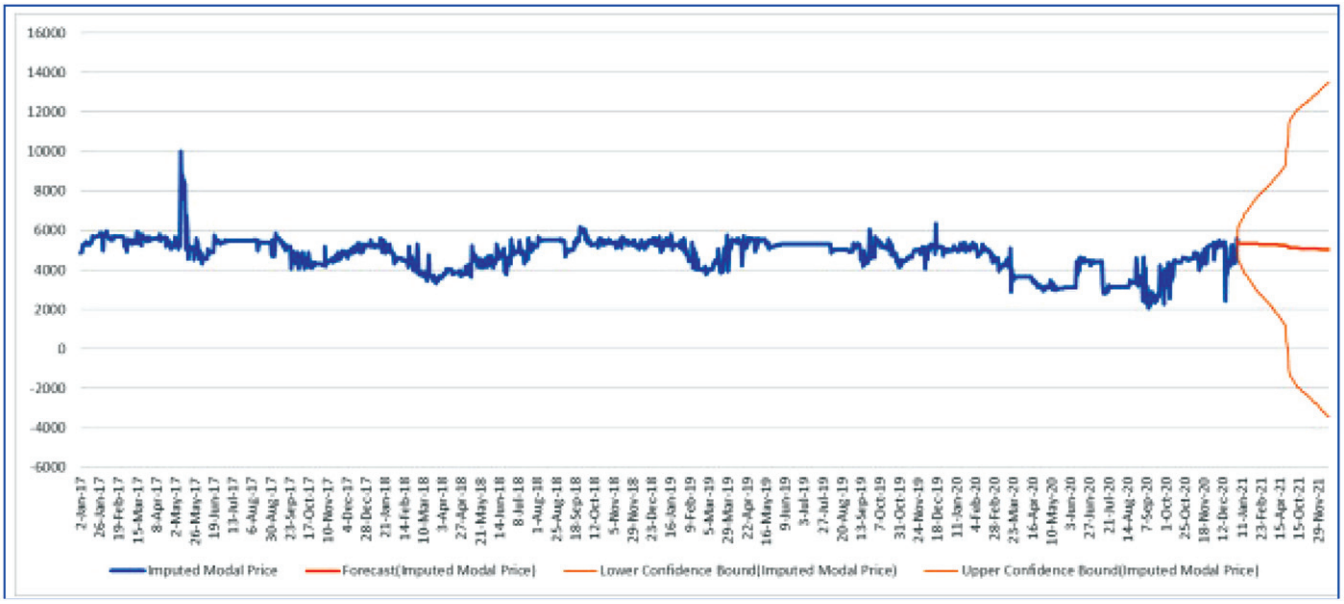

Figure 2: Initial forecast analysis of cotton prices

Figure 2. The MSE for this analysis is 2922310.2. It can be seen that there exists a reasonable stability in the trend due to lower perishability of cotton.

Following this analysis, the model was developed using ML techniques. Beginning with analysis of time-series data, the ADF test was conducted to test for stationarity. This is a commonly used hypothesis test where the null hypothesis H0 represented by a unit root, depicts that the time series is not stationary, while the alternate hypothesis H1 represents that the time series is stationary. It was found that the ADF value was approximately equal to the critical value, and the p-value was approximately equal to 0.05. Thus, the null hypothesis H0 may be rejected representing that the data follows stationarity and is ready for time-series analysis.

Additionally, the rolling mean was also computed, as shown in Figure 3. As there is consistent overlap between the data and the rolling mean, it can be further substantiated that the data is stationary.

The Bayesian Ridge Regression model was developed and parameters were made to reshape and fit the model. A unique feature of this type of a regression model is that linear regression is mapped using probability distributions instead of point estimates. This helps with dealing with insufficient and/or poorly distributed values. Additionally, the output is based on this probability distribution instead of a single value. The coefficients are estimated based on multi-collinearity. A biasing coefficient is added here to prevent overfitting of the model[23]. Lasso regression is another regression model that works along the lines of the ridge regression model. However, since Lasso regression selects variables randomly in cases where high collinearity exists between variables, it has not been discussed within the scope of this research[24].

Several performance metrics have been used in this model to assess the quality of performance and measure the extent of accuracy:

1. Mean Absolute Error (MAE): It is the average of the sum of the difference between the actual and predicted values represented by equation (1), where yi represents the predicted value, xi represents the true value, and n is the total number of samples.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad \text{... (1)}$$

2. Coefficient of Determination (R^2): This represents the proportion of variation in the dependent variable that is predictable from the independent variable.
3. Mean Square Error (MSE): It is the sum of the squares of the differences between the actual and predicted values represented by equation (2), where yi represents the observed values and y represents the predicted value for 'n' number of samples. The root mean square error (RMSE) is an extension of this metric which essentially takes the square root of the MSE value.

$$MSE = \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n} \quad (2)$$

The results obtained from this model indicate an MAE of 547. 3108, R^2 value of 0.0918, and an MSE value of 455844.206 (i.e., RMSE = 675.162). These results are shown below in Figure

4. It can be seen that the results have significantly improved with respect to the initial results with a significant decrease in the MSE as explained in the initial forecast analysis. The positive correlation represented by the ridge regression model indicates a good estimate of the predicted values with the actuals. As strong collinearity forms the basis of a Bayesian Ridge Regression model, it makes predictions based on probability distribution rather than point estimates[25].

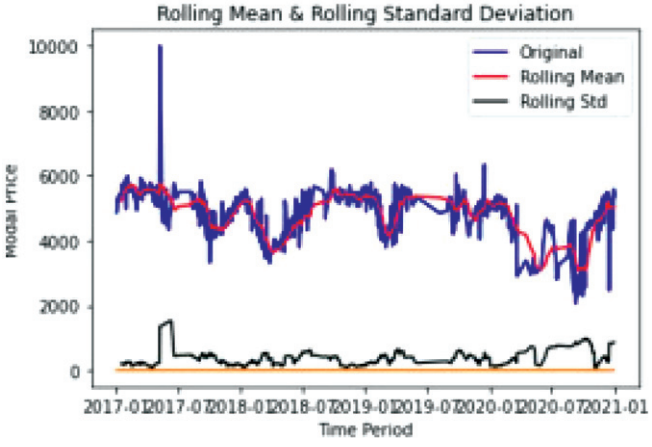

Figure 3: Rolling mean over the dataset

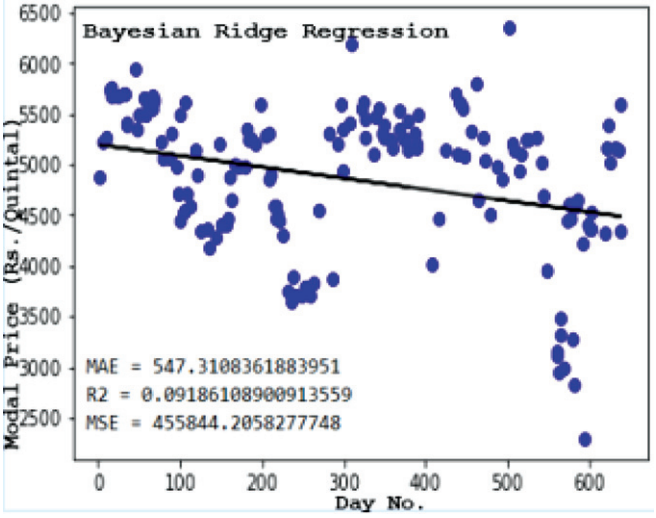

Figure 4: Bayesian ridge regression results

6.0 Conclusion and Future Enhancements

Machine learning has gained traction in a wide spectrum of fields, and its implementation in agriculture is not new. Finding its applications in crop price and yield predictions specifically, machine learning continues to bring innovative solutions in this much-needed space. This research aims to use such machine learning techniques to address farmers, especially the marginal and small farmers of India, with intelligent solutions for price prediction, particularly for seasonal commodities in the markets of Karnataka. The implications of this study are emphasized by the focus on seasonality. Addressing seasonal commodities brings an interesting challenge to the model designers, and being able to predict and deal with such trends sufficiently accurately is the essence of this project. This model used historic price data to create a holistic model with adequate accuracy so that it would not only assist farmers with price prediction for a given seasonal crop, but also aid them with better financial and crop planning over their harvesting seasons. This model can be further enhanced by stitching supplemental influencing data points like environmental parameters (soil moisture, rainfall, weather, etc.) that can bring out closer predictions. While Bayesian Ridge Regression has provided fair results for the given data size, other ML algorithms may be tested to check for consistency and accuracy to suggest the most optimal technique where large datasets would require higher computing power. Further, data of other crops can be input into the model to analyze trends and fluctuations, and make suitable predictions. Associated crop yield predictions may also be made following the similar procedure.

7.0 References

1. Agriculture in India: Industry Overview, Market Size, Role in Development... IBEF. (n.d.). India Brand Equity Foundation. *Retrieved* June 18. 2022.
2. Anubha, Tripathi, K., Kumar, K., & Khandelwal, G. (2021): Onion Price Prediction for the Market of Kayamkulam. Data Analytics and Management, 77–85.
3. Basso, B., & Liu, L. (2019): Seasonal crop yield forecast: Methods, applications, and accuracies. Advances in Agronomy, 154, 201–255.
4. Cenas, P. V. (2017): Forecast of Agricultural Crop Price using Time Series and Kalman Filter Method. *Asia Pacific Journal of Multidisciplinary Research*, 5(4). 2018.
5. Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K. H., & Liew, X. Y. (2021): Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques. *Advances in Science, Technology and Engineering Systems Journal*, 6(4), 376-384.
6. Ghutake, I., Verma, R., Chaudhari, R., & Amarsinh, V. (2021): An intelligent Crop Price Prediction using suitable Machine Learning Algorithm. ITM Web of Conferences. https://doi.org/10.1051/itmconf/20214003040.
7. Israeli Trade and Economic Mission in India. (2020, May 11): Small and Marginal farmers in India – Difficulties and Solutions. India - Israel Trade & Economic Office, Embassy of Israel. *Retrieved* June 18, 2022, https://itrade.gov.il/india/2016/08/29/small-andmarginal-farmers-in-india-difficulties-and-solutions/
8. Jain, A., Marvaniya, S., Godbole, S., & Munigala, V. (2020): A Framework for Crop Price Forecasting in Emerging Economies by Analyzing the Quality of Time-series Data. arXiv. Org. https://doi.org/10.48550/arXiv.2009.04171
9. K., M., Swamy P.S., D., B.R., J., & N.N., N. (2013): Resource Use Efficiency of Bt Cotton and Non-Bt Cotton in Haveri District of Karnataka. *International Journal of Agriculture and Food Science Technology*, 4(3), 253–258.
10. Kalichkin, V. K., Alsova, O. K., & Yu Maksimovich, K. (2021): Application of the decision tree method for predicting the yield of spring wheat. IOP Conference Series: Earth and Environmental Science, 839(3). https://doi.org/10.1088/1755-1315/839/ 3/032042
13. Kanwal, S. (2022, February 14): Agriculture in India - statistics & facts. Statista. Retrieved June 18, 2022, from https://www.statista.com/ topics/4868/agricultural-sector-inindia/#dossierKeyfigures
14. Kumar, G. R. (2017, January 24): Focusing on major problems of marginal farmers. The Hans India. *Retrieved* June 18, 2022, from https://www.thehansindia.com/posts/index/Hans/2017-01-23/Focusing-on-majorproblems-of-marginal-farmers/275349?infinitescroll=1
15. Ma, W., Nowocin, K., Marathe, N., & Chen, G. H. (2019): An interpretable produce price forecasting system for small and marginal farmers in India using collaborative filtering and adaptive nearest neighbours. Proceedings of the Tenth International Conference on Information and Communication Technologies and Development, 6, 1–11. https://doi.org/10.1145/3287098.3287100
16. Mgale, Y. J., Yan, Y., & Timothy, S. (2021): A Comparative Study of ARIMA and Holt-Winters Exponential Smoothing Models for Rice Price Forecasting in Tanzania. OALib, 08(05), 1–9.

17. NITI, Integrated Research and Action for Development. (2007, October): Extension of MSP: Fiscal and Welfare Implications. A Study for the Planning Commission. *Retrieved* June 18, 2022, from https://www.niti.gov.in/planningcommission.gov.in/docs/reports/sereport/ser/ser_msp.pdf
18. Sabu, K.M. and Kumar, T.M. (2020): Predictive analytics in Agriculture: Forecasting prices of Arecanuts in Kerala. Procedia Computer Science, 171, 699–708.
19. Global Hunger Index Scores by 2021 GHI Rank. (n.d.). Global Hunger Index (GHI) – Peer Reviewed Annual Publication Designed to Comprehensively Measure and Track Hunger at the Global, Regional, and Country Levels. *Retrieved* June 18, 2022
20. India at a glance FAO in India Food and Agriculture Organization of the United Nations.
21. Tutorial point (n.d). Scikit Learn - Bayesian Ridge Regression. Scikit Learn. https://www.tutorialspoint.com/scikit_learn/scikit_learn_bayesian_ridge_regression.htm
22. VIT University, Vellore, Tamil Nadu, India. (2020): Crop Value Forecasting using Decision Tree Regressor and Models. *European Journal of Molecular & Clinical Medicine*, 07(02).
23. Vohra, A., Pandey, N., & Khatri, S. (2019): Decision Making Support System for Prediction of Prices in Agricultural Commodity. 2019 Amity International Conference on Artificial Intelligence (AICAI). https://doi.org/10.1109/aicai.2019.8701273
24. What is Marginal Farmer, IGI Global. (n.d.). IGI Global. *Retrieved* June 18, 2022, from https://www.igi-global.com/dictionary/marginal-farmer/68592
25. Scikit learn - Bayesian Ridge regression. Tutorials Point. (n.d.). Retrieved June 30, 2022, https://www.tutorialspoint.com/scikit_learn/scikit_learn_bayesian_ridge_regression.htm
26. Lasso vs Ridge vs elastic net: ML. GeeksforGeeks. (2022, February 11). *Retrieved* June 30, 2022, from https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/
27. Wikipedia contributors. (2006, November 22). Kharif crop. Wikipedia. https://en.wikipedia.org/wiki/Kharif_crop 30, 2022, from https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/
28. Wikipedia contributors. (2006, November 22). Kharif crop. Wikipedia. https://en.wikipedia.org/wiki/Kharif_crop