iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46580

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

Future Price Prediction of Pulses Using Machine Learning

Shashank Sangale[1], Gaurav Chauhan[2], Varad Luktuke[3], Nilay Pande[4], Prof. P. S. Gaikwad[5]
[1, 2, 3, 4]*Department of Computer Science, AISSMS's Institute of Information Technology, Pune, Maharashtra, India 411001*
[5]*Computer Department, AISSMS's Institute of Information Technology, Pune, Maharashtra, India 411001*

Abstract: Price is the key factor in financial activities. Insta- bility in the market is indicated by unexpected price fluctuations. Machine learning offers a wide range of strategies for predicting product prices in order to deal with market volatility.In this study, we investigate the use of a machine learning approach to forecast moong dal prices.A system which makes use of various machine learning algorithms like SVR, Random Forest, XGBoost and ARIMA is proposed. Based on a comparative study of the results given by these algorithms, the most optimal algorithmcan be chosen for further predictions.
Index Terms: Machine Learning, Price prediction, Pulses, Support Vector Regression, XGBoost, Random Forest, Neural Networks, ARIMA.

I. INTRODUCTION

One of the needs of human life as an economic being is the need for food. Food is a basic human need that must be met to survive. Food has an important role in the economy and the health of a nation. When the amount of food availability is lower than the needs, it can cause economic instability. In realizing a country's food security, the government needs to pay attention to 4 aspects of food, namely availability, access, utilization, and stability. One of them that needs attention is the stability of food prices. Food price stability is an important aspect to be considered, because volatile commodity prices can cause various negative impacts when significant changes occur. Furthermore, price fluctuations in food commodities will be the most significant contributor to a country's inflation rate. The high fluctuation of staple food prices in Asia, can make everyone anxious and threaten government stability. Stable price of staple food brings various benefits for developing countries, for instance, this can make poor farmers and con- sumers improve their economic life and avoid poverty traps. As a result, the government will implement several strategies in order to reduce market fluctuations in staple food prices. In order to help the government to decide the market price of staple food materials, the prediction of staple food price can be considered. In this paper, A system is proposed for the price prediction of moong dal, taking into consideration several influence factors such as location, variety, date etc. Here, 4 algorithms viz. SVR, Random Forest, XGBoost and ARIMA are proposed. These algorithms are known as good methods for predicting using multi variables. Enormous studies have been done to predict future prices of various merchandise goods. Some of these studies have applied Machine Learning to solve this problem. But, most of the studies use different algorithms as per the datasets available. In this paper ,we try to proposea generic system which can be used to predict the future price of moong dal.

II. PROBLEM DEFINITION

Various studies have been performed to predict future prices using different Machine Learning models and algorithms. The aim of this paper is to propose a system which can be usedto predict future prices of pulses using previous data.

III. LITERATURE SURVEY

Md. Mehedi Hasan et al. [1] have looked into the appli- cation of machine learning approach to forecast the price of onion. For making prediction various machine learning algorithms e.g. K- Nearest Neighbour (KNN), Naıve Bayes, Decision Tree, Neural Network (NN), Support Vector Machine (SVM) were used. It was observed that KNN being a easyto implement algorithm, gave sub-optimal results. SVM and Naıve Bayes algorithms need more data as compared to other algorithms to give an equivalent accuracy. Decision tree gave the second highest accuracy rate. Neural Network achieved the highest forecasting rate. Also, it can be concluded that for numerical time series data, machine learning approach works fruitfully.
Lobna Nassar et al. [2] have highlighted the importance of imputation for having highly performing prediction models in this paper. Three imputation techniques were tested against a non-imputation approach that discards records with any missing values; the complete-case analysis (CCA).

The deep learning linear memory vector recurrent neural network-RNN (LIME) imputation model was tested along with two other non-deep learning models - namely the linear function and Last Observation Carried Forward. The best performing imputation model, the one with the lowest price and yield prediction errors, was chosen using a basic LSTM deep learning (DL) prediction model. Based on the aggregated measures, the LIME imputation model proved to significantly improve the prediction results of a simple LSTM DL model.

Wahyu Hidayat et al. [3] The data points analysed to determine whether a student is active or inactive in school are the type of school, Grade Points(GP),Grade Points Average(GPA) and the parents' jobs. The SVM system can accurately predict 311 active students and 53 non-active students, as can be observed. After the overall accuracy calculation is performed, it is found that SVM has the best classification accuracy of 95

Slamet Wiyono et al. [4] have used Multiple Linear Regression (MLR) to predict food prices, especially in the modern market, based on the predicted prices, then a decision support system is made to make an alternative ranking of food selection accumulation. They have used a Simple Additive Weighting (SAW) method to rank alternative food staples that have nutritional weight and price. Based on testing of the application of SAW, the same results are obtained between manual calculations and calculations provided by the system for checking the accuracy. In comparison to other food items, the forecast of the price of food "Rice" has the least mistake results, according to the error level testing.

Haviluddin et al. [5] have compared two algorithms, Backpropagation Neural Networks (BPNN) and Single Moving Average (SMA). They predict the price of chili which is one of the food commodities that can affect the inflation rate. Its uncertain price and even increasing at certain times will negatively impact the society. BPNN is a supervised learning method that has a target to be wanted. Its characteristic is to minimize errors in the output generated by the network. Meanwhile, SMA is a forecasting method that uses several actual new demand data to generate a forecast value for future demand. After testing, the BPNN method has a better level of accuracy than SMA. It is because the data used is a long term and fluctuating time-series type. The SMA method is not suitable for long-term and volatile forecasting. Also, SMA is unable to cope well with trends or seasonality. It will have a better forecasting accuracy ifused on constant or stable data, and the data used is more complicated.

Said Fadlan Asnhar et al. [6] have forecasted staple food price in multivariable factors using combination of ARIMA and regression model. They have used two type of regression models, multiple linear and Fourier regression model. The significant different results between linear regression with ARIMA and Fourier regression with ARIMA is only for fluctuating commodities like green cayene pepper and garlic. Therefore, it can be concluded that Fourier regression can produce good result of accuracy if the data in high fluctuation.

Triyanna Widiyaningtyas et al. [7] have applied the Extreme Learning Machine(ELM) method to predict the price of staple food commodities in East Java Province and also measure the performance of the ELM in predicting staple food commodities price. Although the ARIMA approach has been successful in predicting time-series data, it is not without flaws.These flaws include the necessity for a significant amount of data, the lack of mechanisms to update the model in the event of new data, and the time it takes to build a competent model. In overcoming this problem, authors have chosen another predictive method known as Extreme Learning Machine(ELM). This method has advantages in terms of learning speed. This method also produces a good level of accuracy and is very promising for problems with time-series data. After testing, it was shown that ELM had an accuracy of 98.79

Pramod K. Mishra [8] found out that the time series data is not enough to generate the prediction of unorganized supply chains like agri-foods. The supply chain data (primarily collected data) are needed to be integrated with secondary (time series) data for better clarity. Moreover, it was also discovered that seasonality was not a major concern for price fluctuations, rather market dynamics had major role in controlling the price. Furthermore, author believes that the arrival and the wholesale prices are good enough to predictthe retail prices with about 90

IV. METHODOLOGY

The proposed system makes use of machine learning to predict the prices of moong dal. Time-series data containing features such as location, variety, date etc. is used to train the models. The results of all 4 models are compared to select an optimal model. The trained model is then be used to predict future prices. Figure 1 shows the overall architecture of the system.

There are 3 aspects to the proposed system:

- Training
- Comparison
- Prediction

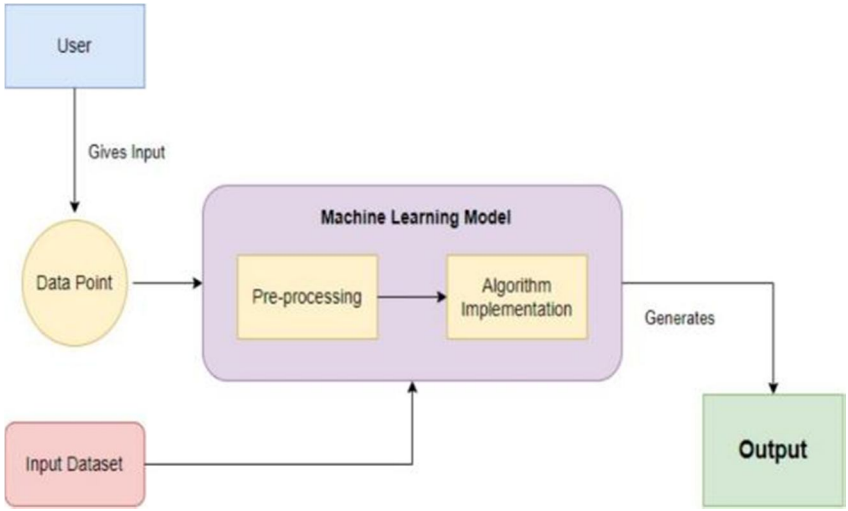


Fig. 1. System Architecture

A. Training

Model training in machine learning is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved.There are various types of machine learning models, with supervised and unsupervised learning being the most common. Supervised learning is possible when he training data contains both the input and output values. In the context of this system, price is the output value.

Training phase can be broadly classified in 3 steps:

1) *Preprocessing:* Data preprocessing is the procedure for preparing raw data for use in a machine learning model. It involves processes such as data collection, finding missing data, encoding categorical data, splitting dataset into training and test set, etc.
2) *Processing:* Once the data is processed, the next step is to feed that data to the machine learning models. The proposed system makes use of four algorithms andthus 4 different models are trained.
3) *PostProcessing:* In this phase, the trained models are evaluated using various measures such as R-squared, mean absolute error, etc.

B. Comparison

This aspect of the proposed system involves comparing the results of all 4 models to find the best model. The model selected in this phase is used to predict futureprices.

C. Prediction

The model which was trained in the earlier phases and gave the best results is used in this phase to predict the price of moong data for a given point in time.

V. ALGORITHMS

A. Support Vector Regression (SVR)

SVR is based on the same premise as SVM, however it is used to solve regression problems. It is particularly suited to handle multiple inputs. It provides high prediction accuracy and also has the ability to tackle overfitting problem. However, it is sensitive to user's defined free parameters.

B. Random Forest

Random Forest is a robust method for forecasting since its design that is filled with various decision trees, and the feature space is modelled randomly. It works with both discrete and continuous variables and handles missing values automatically. It demands more processing power and resources, as well as longer time to train, because it builds more trees.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

C. XGboost

Extreme Gradient Boosting (XGBoost) is a technique de- veloped by University of Washington academics. It is a C++ library that optimises the training process for Gradient Boosting. Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error.

In this algorithm, decision trees are created in sequential form.All of the independent variables are given weights, which are subsequently fed into the decision tree, which predicts outcomes. The weight of factors that the tree predicted in- correctly is increased, and these variables are fed into the second decision tree. Individual classifiers/predictors are then combined to form a more powerful and precise model. It can be used to solve problems including regression, classification, ranking, and user-defined prediction.

D. Autoregressive Integrated Moving Average(ARIMA)

Auto-regressive integrated moving average (ARIMA) also known as Box-Jenkins method is highly sophisticated but addresses both trend and seasonal properties of time-series. It works well for linear time series data. For short-run fore- casting, it provides more robust and efficient results than the relative models with more complex structures. However, this model does not work well for non-linear time series and also requires comparatively more data.

E. BackPropogation Neural Network

The BPNN is a multilayer feed-forward neural network that is trained according to an error back-propagation algorithm. It has flexible nonlinear modelling capacity. Because of its capacity to massively use parallel computing, it gives fast responses. Even though it has high learning accuracy, the model is sensitive to noise. Since its actual performance is based on initial values, it sometimes might have slow convergent speed.

VI. IMPLEMENTATION

A. Data Collection

Data collection is the systematic gathering and measurement of information from various sources in order to obtain a complete and accurate picture of a subject area. To train the machine learning model for the proposed system, we used datasets from data.gov.in and kaggle.

B. Data Preparation

Data preparation is a process of getting data ready to use by cleaning and transforming raw data prior to processing and analysis. It's a crucial stage before processing that often include reformatting data, making data changes, and integrating data sets to improve data.

C. Exploratory Data Analysis

Exploratory Data Analysis is a process of understanding the data by visually representing it in the form of graphs, pie charts, histograms, etc.

D. Data mining:

Data mining is a process used by companies to turn raw data into useful information. Businesses can learn more about their customers by employing software to seek for trends in massive batches of data. This allows them to design more successful marketing campaigns, improve sales, and cut costs.

E. Training and Prediction:

Training consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the inputdata through the algorithm to correlate the processed output against the sample output.

VII. RESULTS

Applying the chosen algorithms to the gathered dataset, we calculated the performance of the algorithms and inserted them into the accuracy table for further comparison and following are graphical representations of results:

VIII. CONCLUSION

After extensive survey of the research literature, it was found that various algorithms like KNN, Naive Bayes, De- cision tree, SVM, ARIMA, BPNN, Simple Moving Average,

Algorithm	R2 Score		
	Minimum	Maximum	Modal
SVR	92%	91.14%	91.7%
Random Forest	99.03%	98.20%	99.26%
XGBoost	98.61%	98.21%	99.07%
ARIMA	-	-	-

MAPE

Algorithm			
	Minimum	Maximum	Modal
SVR	6.63%	6.16%	6.25%
Random Forest	1.44%	1.33%	1.28%
XGBoost	2.24%	2.06%	1.97%
ARIMA	6.35%	6.09%	6.04%

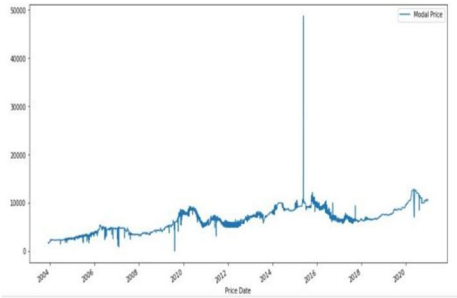


Fig. 2. Data plot graph

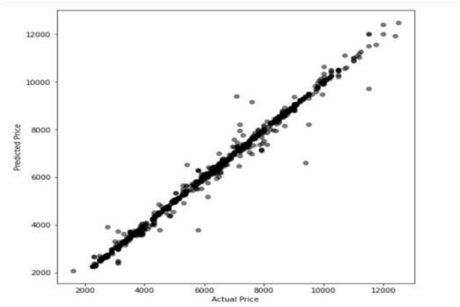


Fig. 3. Random Forest

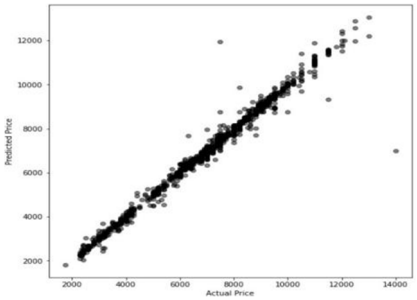


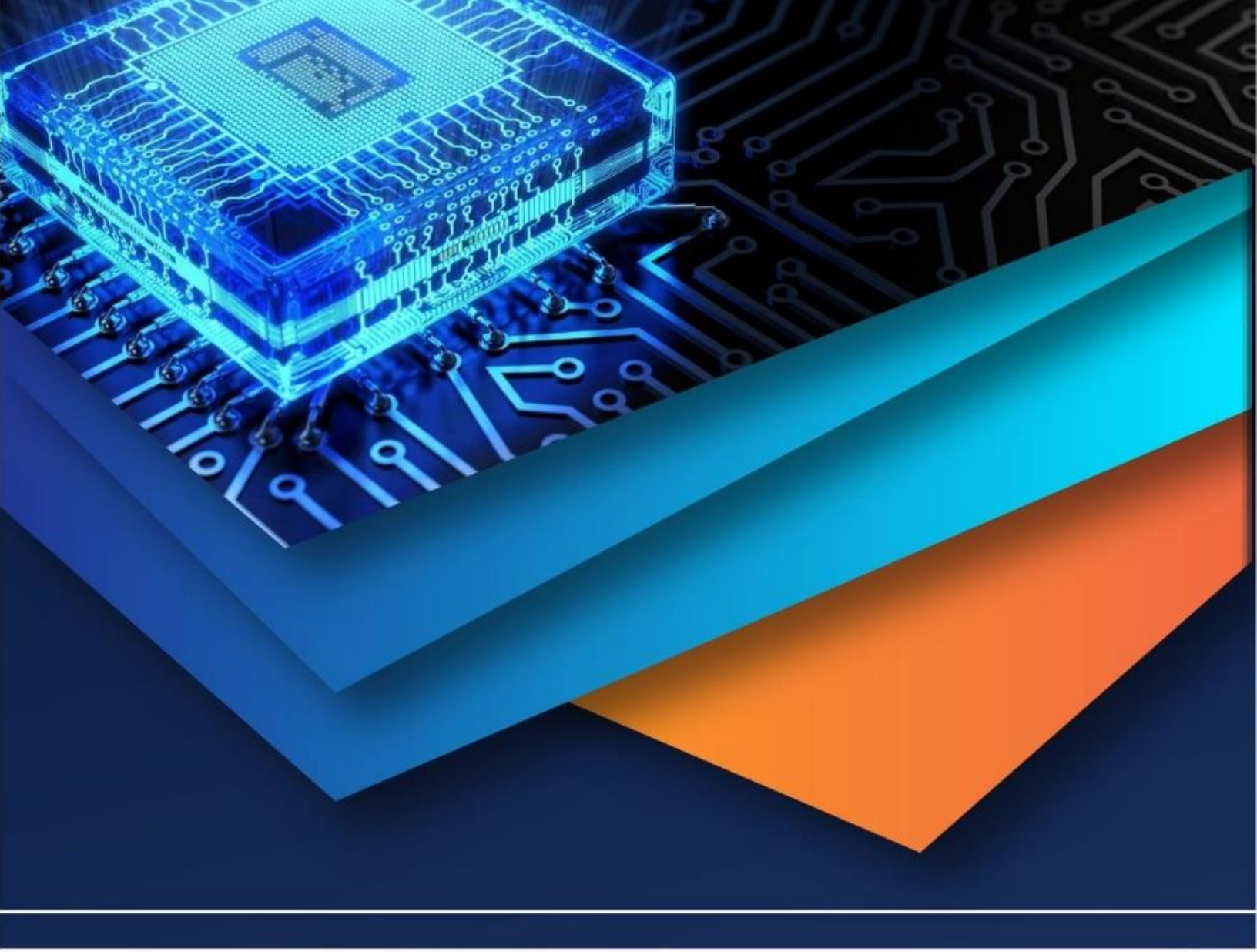
Fig. 4. XGboost

ELM etc. were used for time-series analysis till date. Each algorithm performs admirably in our tests. We investigated four algorithms performance and found the best algorithm for a better pulse price prediction. It is difficult to choose a better algorithm between XGBoost and Random Forest based on just R2 score, but taking MAPE in consideration, Random Forest proves to be the better algorithm between the two. Depending on this forecasting price we can calculate the demand and supply of pulses, as we know demand-supply plays the main role in market equilibrium state.

Now if we can calculate the future demand-supply of pulses based on this prediction we can maintain equilibrium state in pulse market which will help us to remove pulse market instability. The main limitation of our work is unusual behavior of data and low number of records. Due to which,the performance of algorithms like ARIMA and BPNN were comparatively lower.

REFERENCES

[1] Md. Mehedi Hasan, Rubaiya Hafiz, Muslima Tuz Zahara, Mohd. Saifuzzaman, Md. Mahamudunnobi Sykot, ”Solving Onion Market Instability by Forecasting Onion Price Using Machine Learning Approach,” 2020 International Conference on Computational Performance Evaluation (ComPE) North-Eastern Hill University, Shillong, Meghalaya, India. Jul 2-4, 2020.
[2] Lobna Nassar, Muhammad Saad, Ifeanyi Emmanuel Okwuchi, Mohita Chaudhary, Fakhri Karray, Kumaraswamy Ponnambalam, ”Imputation Impact on Strawberry Yield and Farm Price Prediction Using Deep Learning,” 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) October 11- 14, 2020. Toronto, Canada.
[3] Wahyu Hidayat, Mursyid Ardiansyah, Kusrini Kusrini, ”Decision Support System for Selection of Staples Food and Food Commodity Price Prediction PostCOVID-19 Using Simple Additive Weighting and Multiple Linear Regression Methods,” 2020 3rd International Conference on Informations and Communications Technology (ICOIACT).
[4] Slamet Wiyono, Taufiq Abidin, ”COMPARATIVE STUDY OF MACHINE LEARNING KNN, SVM, AND DECISION TREE ALGORITHM TO PREDICT STUDENT’S PERFORMANCE”,[Wiyono et. al., Vol.7 (Iss.1): January 2019]
[5] Haviluddin, Taufikurrahman Khosyi, Hario Jati Setyadi, Aji Prasetya Wibawa, Felix Andika Dwiyanto, Andri Pranolo, Fachrul Kurniawan, Muslimin, ”A backpropagation neural network algorithm in agricultural product prices prediction”, 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT) April 09-11. 2021, ISTTS Surabaya, Indonesia.
[6] Said Fadlan Asnhari, P. H. Gunawan, Yanti Rusmawati, ”Predicting Staple Food Materials Price Using Multivariables Factors (Regression and Fourier Models with ARIMA)”, 2019 7th International Conference on Information and Communication Technology (ICoICT).
[7] Triyanna Widiyaningtyas, Ilham Ari Elbaith Zaeni, Tyas Ismi Zahrani, ”Food Commodity Price Prediction in East Java Using Extreme Learning Machine (ELM) Method”,2020 International Seminar on Application for Technology of Information and Communication (iSemantic).
[8] Pramod K. Mishra, ”Predicting Agri-Food Prices with Time-Series and DataMining based Methods”, Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS2021) IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381- 1327-2.
[9] https://www.javatpoint.com/data-preprocessing-machine-learning
[10] https://www.researchgate.net/publication/349568120 Forecasting Electricity Consump
[11] https://www.coursehero.com/file/p6ucbk1/Data-collection-is-a-systematic-approach-to-gathering-information-from-a-variety/
[12] https://dl.acm.org/doi/10.1145/2939672.2939785
[13] https://dmlc.cs.washington.edu/xgboost.html
[14] https://ieeexplore.ieee.org/document/8734193
[15] https://www.researchgate.net/publication/336876643 Short Term Prediction on Bitcoin
[16] https://www.researchgate.net/publication/349568120 Forecasting Electricity Consump



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY