Moneter : Jurnal Keuangan Dan Perbankan, 11 (1) (2023)18-27
Published by: Universitas Ibn Khaldun
Journal homepage: http://ejournal.uika-bogor.ac.id/index.php/MONETER/index
E-ISSN : 2615-514. P-ISSN : 2302-2213

Implementation Of Extreme Gradient Boosting Algorithm For Predicting The Red Onion Prices

Pungky Nabella Saputri[1], Farrikh Al Zami[2*], Filmada Ocky Saputra[3], Pulung Nurtantio Andono[4], Rama Aria Megantara[5], L Budi Handoko[6], Chaerul Umam[7], Firman Wahyudi[8]

[1234568]Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia,
[7]Ramani, B.V, The Netherlands

E-mail: [1]pungkynabsofficial16@gmail.com, [2]alzami@dsn.ac.id

Article Information

History of the article:
Accepted: December 2022
Corrected: .January 2023
Accepted: January 2023

Keywords :
Red onions, Price prediction, Extreme Gradient Boosting algorithm, Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE)

Abstract

Red Onion or the Latin name Allium Cepa is included in the group of vegetable plants that are needed by the public for food needs. Red Onions are one of the seasonal crops so their availability can change in the market which causes price instability due to a lack of supply of production by several factors: 1) not yet it's harvest time, 2) crop attacked disease pests and fungi, and 3) weather factor. Therefore, a study is needed to predict red onion prices, so that it can be used as information for the government to stabilize red onion prices. The method used in this study is CRISP-DM and the Extreme Gradient Boosting algorithm to predict the price of red onions by taking data samples from Tegal and Pati Cities. The results of this study are that the Extreme Gradient Boosting algorithm is able to produce Tegal District Root Mean Square Error (RMSE) values of 5107.97% and Mean Absolute Percentage Error (MAPE) values of 0.17%. For prediction results with Pati Regency data samples, it produces a Root Mean Square Error (RMSE) value of 6049.74% and a Mean Absolute Percentage Error (MAPE) of 0.17%.

Introduction

Indonesia is an agricultural country that is rich in various types of natural resources, so it is known as a developed country in the agricultural sector. The agricultural sector in Indonesia is very important because it is a resource that helps the existing economy. One of the agricultural sub-sectors that has the most important role in horticulture. Where horticultural plants are one group of plants consisting of vegetables, ornamental plants, etc. One type of horticultural plant is the red onion. Red onions have the Latin name Allium Cepa which is included in the group of vegetable plants that are needed by the community for food needs. Until now, the need for red onions in Indonesia continues to increase from year to year due to several factors(Rizaty, 2021).

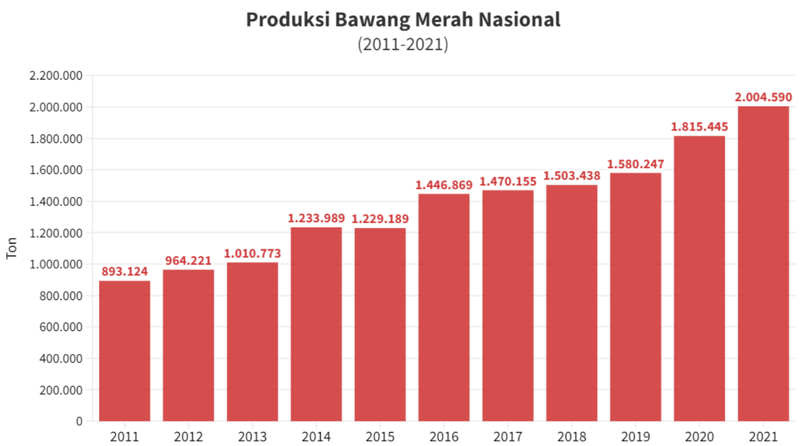


Figure 1 : Production of Red Onion(Rizaty, 2021)

Badan Pusat Statistik (BPS) reports that red onion production in Indonesia in 2021 will increase to 2 million/ton, which is an increase of 10.4% from the previous year of 18.1 million/ton which can be seen in figure 1(Rizaty, 2021). This is due to population growth and the increasing food needs of the people in Indonesia. According to Central Java Province, which is the center of the largest red onion in Indonesia with production reaching 564,255 tons, this amount is equivalent to 28.1% of the total national production(Rizaty, 2021). The red onion-producing areas in Central Java include Brebes, Tegal, Boyolali, Demak, Kendal, Temanggung, and Pati.

Red onions are seasonal crops so their availability can change in the market causing price instability(Saumyamala et al., 2019)(Widiyaningtyas et al., 2020). Then, due to demand is quite large, mades Indonesian Government imposes import on red onions which mades the onion price become unstable during the harvest season(Andono et al., 2022). Moreover, the lack of supply of production results is caused by several factors such as: 1) it is not yet time for harvesting, 2) plants are attacked by pests and fungi, 3)weather factors(Et. al., 2021)(Hasan et al., 2020). This situation greatly affects the reduced supply of red onions in the market. This situation had a significant impact on the decline in the supply of red onions in the market. It is very important to predict the price of red onions to overcome the surge in the selling price of red onions that can occur at any time in the market due to the limited production of red onions at harvest time(Afridar et al., 2023).

The prediction of red onion prices is carried out with the hope that the prediction results can be used as input for relevant agencies in making policies to maintain the stability of red onion prices in the market(Ahmad et al., 2021)(R et al., 2020)(Triswanda et al., 2020)(Wihartiko et al., 2021). In addition, the prediction results can also be used by farmers as an illustration of future prices and determining the timing of planting so as not to experience losses at harvest which helped the SMES in selling the commodities(Iswari et al., 2021)(Madaan et al., 2019).

Research on price predictions has previously been done, such as in research of red onions prices prediction using KKN Regressor where the KNN algorithm succeeded in producing an accuracy of 91.67%(Virdaus & Prasetyaningrum, 2020). The study "Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah" resulted in RMSE and MAPE when modeling was 6.61374% and 3.95485% while at the time of testing the RMSE model was 0.23577 % and MAPE is 0.11643%(Islam et al., 2021). The research "Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process" resulted in an actual data accuracy value of 60.49% and an increase of 93.28% based on the prediction data results(Yun et al., 2021). This proves that the XGBoost modeling process can optimize the accuracy value. Therefore, this study will apply the Extreme gradient boosting (XGBoost) Algorithm for red onions price prediction. From the results of this prediction, later evaluation in planning the production of red onions in the future with a better accuracy value.

Methods

In this study, the data used is onion price data from 2018-2021 which includes districts, dates, and prices. And the method used in this study is the Extreme gradient boosting (XGBoost) algorithm. The stages in this study using CRISP-DM method(Pete Chapman et al, 2000). In the CRISP-DM-based data mining process, there are 6 phases, which can be seen in Figure 2.

Figure 2 : CRISP-DM Phases

Business Understanding is the stage of understanding the data mining activities that will be carried out, usually determining goals, understanding the existing situation, and determining the goals of data mining to be carried out.

Data Understanding is one of the most important stages in the data mining process. At this stage, the data collection stage usually includes understanding the usefulness of the data with existing problems, and detecting an interesting subset of data as hypothesis.

Data Preparation is the stage of preparing data for data mining process. Activities at this stage usually include the selection of attributes used, data construction, and data cleaning.

Modeling is the stage in determining the data mining techniques that will be carried out in processing the data. The data mining technique used in this study is XGBoost. Boosting is a data ensemble technique that is usually used to make predictions and classifications. The ensemble technique itself is a method that is built with several prediction and classification models that will be used to classify new data based on the predicted weights from previous results(Dietterich, 2000). The final model of the boosting technique is a combination of a collection of several models with as many as n iterations to produce the smallest error value from the residual. The finalmodel is defined by the following equation:

$$f(x) = \sum_{m-1}^{M} f_m(x) \quad (1)$$

Or it can be showed as following equation:

$$f(x) = y_0 + \sum_{m=1}^{M} y_m h_m(x) \quad (2)$$

Where $f_x = y_0$ and $f_m(x) = y_m\ h_m(x)$ for m = 1, 2, 3, ….., M with a value of $h_m(x)\varepsilon\{-1,1\}. y_m(x)$ is a weak classification, while y_m is a weight for each classification(Iswaya Maalik S, Wisnu Ananta Kusuma, 2019).

XGBoost is a combination method between boosting and gradient boosting. This method first appeared in Friedman's research on the relationship between boosting and optimization to create a Gradient Boosting Machine (GBM). This model will create a new model to make predictions using errors in the previously created model. The algorithm is called gradient boosting, while reducing errors when creating a new model is called gradient descent. Broadly speaking, the gradient boosting algorithm has the following equation:

$$\{y_m, h_m\} = argmin \sum_{m-1}^{M} L(y_i, f^{(m-1)}(x_i) + y_m h_m(x_i)) \quad (3)$$

XGBoost is a version of the Gradient Boosting Method (GBM) which is more useful and scalable because it can complete various functions such as rangking, classification and regression. XGBoost algorithm can perform optimization 10 times faster than other GBM. XGBoost is a tree ensemble algorithm consisting of classification and Regression trees (CART). The accuracy value of the classification results using the XGBoost method depends on the parameters to be used.

Evaluation is the stage of interpreting the results of data mining that has been processed to obtain a model that is by the objectives that have been set. Model evaluation is done by looking at the results of the calculation of the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). RMSE is a standard method for calculating model error when predicting quantitative results. The larger the RMSE value, the worse the level of accuracy. A lower RMSE value indicates that the prediction results are close to the truth value. In general, RMSE is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} A_t - F_t)^2}{n}} \quad (4)$$

Where:

At : Actual data score

Ft : Score of forecasting result

N : Number of data

∑ : Total Score

MAPE is a measure of the accuracy of the model's predictive score, which is expressed in terms of the average absolute percentage of error. MAPE is the average score of the absolute difference that exists between the predicted score and the realized score stated as a percentage of the realized score. The formula for calculating the Mean Absolute Percentage Error is as follows:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \widehat{Y}_t}{Y_t}\right| x\ 100\% \quad (5)$$

Where:
Y_t : Actual score in period t
$\widehat{Y}_t$:Forecast score in period t
n : Number of forecasting periods involved

Result

The results of the early stage of this study were collecting data on red onion prices for the period January 1 2018 – September 30 2022 with district attributes, dates and values obtained from one of the provincial government websites, namely Central Java prices[20]. The red onion price prediction model includes five phases which are the stages of understanding business and data as well as the stages ofmodeling and evaluation, as follows:

First Business Understanding

Setting Business Goals

The purpose of this study is to predict future onion prices to make it easier for farmer groups to plantat the right time to minimize price spikes.

Conduct a Situation Assessment

The process of selecting the month for planting red onions is appropriate but several other factors hinder production results such as disturbances from pests that attack plants, and unfavorable weatherfactors.

Determining Initial Data Mining Strategy

The initial strategy for data mining is to collect data through interviews with the onion farmer groups. In addition, it also searches for public data by utilizing existing electronics.

Second Data Understanding

The data used is sourced from Hargajateng.org regarding the price of red onions from 2018-2022. The data obtained are 1,241 records, which have 3 parameters, namely district, date, and value (price). The following is a sample of raw red onions prices as be seen at Table 1:

Table 1 : Raw Data Sample Red Onion Prices

Districts	1/12018	1/2/2018	…	9/30/2022
Cilacap Districts	0	16000	…	25000
Banyumas Districts	0	20000	…	30000
Purbalingga Districts	0	0	…	29333
Banjarnegara Districts	0	20200	…	0
…	…	…	…	…
Tegal Districts	0	16000	…	30000

Third Data Preparation

In this process, data selection and data cleaning are carried out to simplify calculations. In the raw onion price data that has been obtained, it can be seen that several date columns are empty. This is because that date is a national holiday so there is no input value on that date. So at this stage, the data on the price of red onions will be cleaned by deleting the missing data in the column. The following is onion price data that has been cleaned from raw data as can be seen at Table 2:

Table 2 : Red Onion Price Net Data Sample

Districts	1/2/2018	1/3/2018	…	9/30/2022
Cilacap Districts	16000	18000	…	25000
Banyumas Districts	20000	20000	…	30000
Purbalingga Districts	<NA>	<NA>	…	29333
Banjarnegara Districts	20200	20200	…	<NA>
Kebumen Districts	20000	20000	…	30000
Purworejo Districts	<NA>	<NA>	…	30000
…	…	…	…	…
Tegal Districts	16000	16000	…	30000

In this study, the data sample to be used was selected, namely Tegal and Pati Regencies as red onion centers in Central Java. In this study, the sample data to be used was selected, namely Tegal and Pati Regencies as red onion centers in Central Java. Table 3a is sample data from Tegal District and table 3b is sample data from Pati District :

Table 3a : Tegal District Data Sample

Districts	1/2/2018	1/3/2018	…	9/30/2022
Tegal Districts	16000	16000	…	28000

Table 3b : Pati District Data Sample

Districts	1/2/2018	1/3/2018	…	9/30/2022
Pati Districts	<NA>	18000	…	30000

After the data is cleaned, a time series is made based on the date index. Where before that, the melting process was carried out to facilitate the data processing process. Melting isa function for sending a DataFrame message into a format where one or more columns are identifying variables, while all other columns, considered to be measured variables, are not pivoted to the row axis, leaving only two columns, a variable, and a non-identifying value. It can be seen in table 4a is the result of melting data in Tegal Districts and table 4b is the result of melting data in Pati Districts.

Table 4a : Melting Data for Tegal Districts

	Districts	Date	Value
0	Tegal District	1/2/2018	16000
1	Tegal District	1/3/2018	16000
2	Tegal District	1/4/2018	16000

	Districts	Date	Value
...	...	...	...
1126	Tegal District	9/30/2018	28000

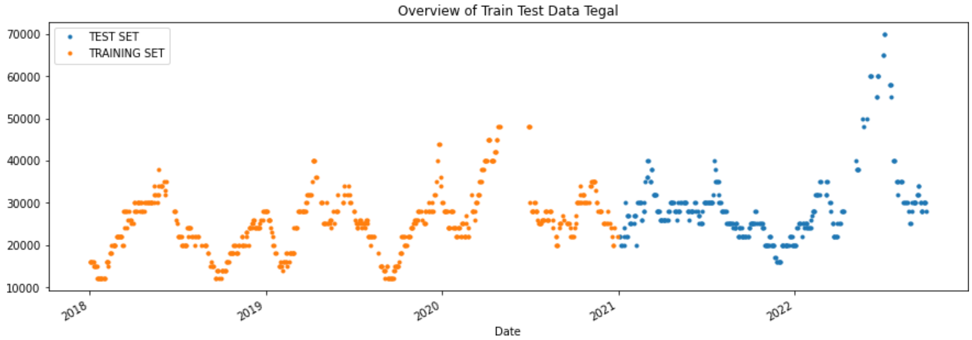
Table 4b : Melting Data for Pati Districts

	Districts	Date	Value
0	Pati District	1/2/2018	<NA>
1	Pati District	1/3/2018	18000
2	Pati District	1/4/2018	<NA>
3	Pati District	1/5/2018	18000
4	Pati District	1/8/2018	18000
5	Pati District	1/9/2018	<NA>
…	…	…	…
1126	Pati District	9/30/2018	30000

After the data is successfully melting, it enters to the modeling stage using the XGBoost Algorithm.

First. Modeling

This study will implement predictions using the XGBoost Algorithm. In this prediction process, we have to make time-series features based on the date index, then proceed with the data split process by dividing the testing and training data. In this study, the data is divided by 60% for testing and 40% for training, after that a visualization process will be carried out so that it is easy to understand. The following is a visualization of the split data.



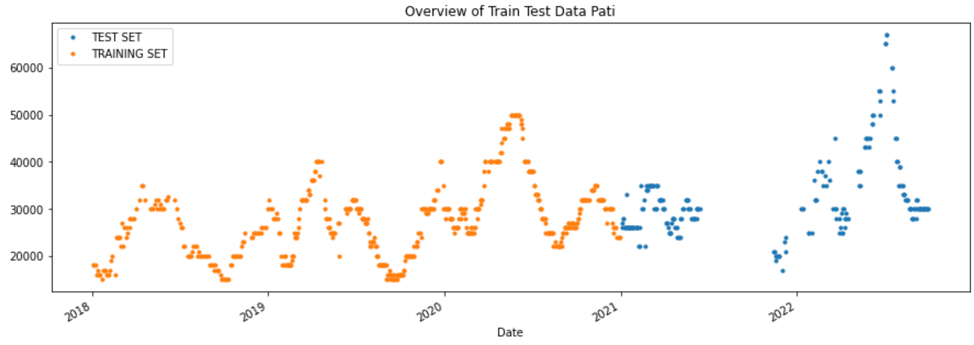


Figure 3 : Overview Train and Test Data

In figure 3 it can be seen that some missing data affect the predicted value. It can be seen in figure 4 which is a visualization between the prediction and the actual that the predicted value is slightly insignificant to the actual value, this is because when plottingdata a lot of data is lost or has a value of 0 so that the prediction results are not optimal.

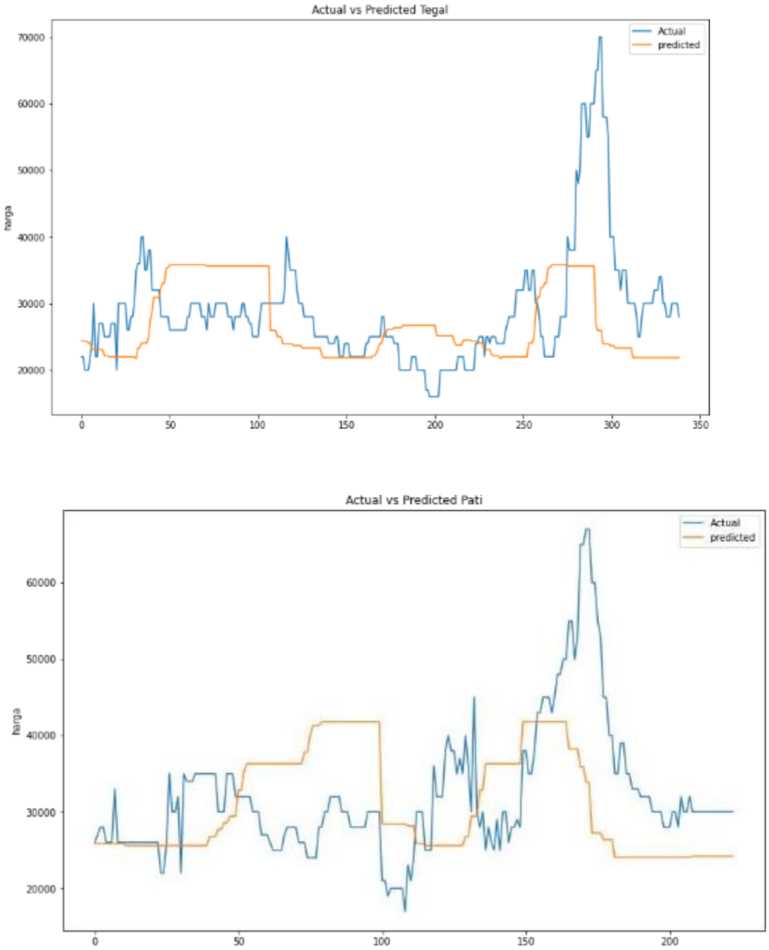


Figure 4 : Actual vs Predicted Visualization

Next in the table 5 to 8 is the result of price predictions that have been carried out by researchers in the next year, here are the results.

Table 4 : Prediction Results of Tegal Districts Onion Prices Based on Data

	Day of week	Quarter	Month	Year	Day of year	Day of month	Week of year	Original price	Price prediction
2018-01-01	0	1	1	2018	1	1	1	0	15192.0
2018-01-02	1	1	1	2018	2	2	1	16000.0	16023.0
2018-01-03	2	1	1	2018	3	3	1	16000.0	16099.0
2018-01-07	1	1	1	2018	9	9	2	0	14826.0
….	….	….	….	….	….	….	….	….	….
2022-09-30	4	3	9	2022	273	30	39	28000.0	28662.0

Table 5 : Our Prediction Results on Tegal Districts Onion Prices in the Future

	Day of week	Quarter	Month	Year	Day of year	Day of month	Week of year	Price prediction
2022-09-30	4	3	9	2022	273	30	39	28662.0
2022-10-01	5	4	10	2022	274	1	39	32244.0
2022-10-02	6	4	10	2022	275	2	39	32239.0
2022-10-03	0	4	10	2022	276	3	40	31291.0
2022-10-04	1	4	10	2022	277	4	40	32174.0
….	….	….	….	….	….	….	….	….
2023-09-30	5	3	9	2023	273	30	39	28662.0

Table 6 : Prediction Results of Pati Districts Onion Prices Based on Data

	Day of week	Quarter	Month	Year	Day of year	Day of month	Week of year	Original price	Price prediction
2018-01-01	0	1	1	2018	1	1	1	0	20125.0
2018-01-02	1	1	1	2018	2	2	1	0	20655.0
2018-01-03	2	1	1	2018	3	3	1	18000.0	17649.0
2018-01-04	3	1	1	2018	4	4	1	0	18218.0
2018-01-05	4	1	1	2018	5	5	1	18000.0	18022.0
….	….	….	….	….	….	….	….	….	….
2022-09-30	4	3	9	2022	273	30	39	30000.0	31661.0

Table 7 : Our Prediction Results on Pati Districts Onion Prices in the Future

	Day of week	Quarter	Month	Year	Day of year	Day of month	Week of year	Price prediction
2022-09-30	4	3	9	2022	273	30	39	30221.0
2022-10-01	5	4	10	2022	274	1	39	30504.0
2022-10-02	6	4	10	2022	275	2	39	30844.0
2022-10-03	0	4	10	2022	276	3	40	31597.0
2022-10-04	1	4	10	2022	277	4	40	31564.0
….	….	….	….	….	….	….	….	….
2023-09-30	5	3	9	2023	273	30	39	30221.0

Second. Evaluation

This study evaluates by performing calculations using the RMSE and MAPE. To see that XGBOOST is optimal in calculating the predicted value, the researcher also uses linear regression as a comparison. The results of the comparison of accuracy in Tegal and Pati Districts can be seen in table 8.

Table 8 : Comparative Results of Accuracy Calculations

	Linier Regression		XGBoost	
	Tegal	Pati	Tegal	Pati
RMSE	5098.39	5561.67	5107.97	6049.74
MAPE	0.16	0.16	0.17	0.17

Conclusions

Based on the calculation process using the XGBoost algorithm on red onion price data where the data is sampled for prediction experiments, namely Tegal and Pati districts, it can be concluded: (1) The XGBoost algorithm has been successfully implemented with the RMSE is 5107.97 and6049.74 for Tegal and Pati which is better than Linear Regression. (2) With the prediction of the price of red onions, farmers and consumers can make preparations toavoid a high price spike due to the limited production of red onions at harvest time. (3) The prediction of red onion prices is carried out as input for the relevant agencies in makingpolicies to maintain the stability of red onion prices in the market. (3) The results of the accuracy calculation have reached the correctness rate with a MAPE valueof 0.17% in Tegal and Pati Districts. The calculation of the onion price prediction can then be done by applying other time series algorithms such as ARIMA/SARIMA. Then for further research, we will add some additional features such as weather, amount of fertilizer supply, amount of pesticide supply, and area of land used.

Acknowledgment

We sincerely thank Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, Republik Indonesia for partially funding this project through the Kedaireka Program. This work is also supported by Universitas Dian Nuswantoro (UDINUS) through the Center of Excellence in Science and Technology, UDINUS and RAMANI B.V. with the grant document contract: Supply Chain and Customer Relationship Management Alignment on Red Onion Commodity using Artificial Inteligent based on Internet of Things and Blockchain, No. 176/E1/KS.06.02/2022.

References

Afridar, H., … G. G.-I. J. of, & 2022, undefined. (2023). Penerapan Metode ARIMA untuk Prediksi Harga Komoditi Bawang Merah di Kota Tegal. *Journal.Peradaban.Ac.Id*, *3*(2), 18–29. http://journal.peradaban.ac.id/index.php/ijir/article/view/1214

Ahmad, S. M., Gaur, A. K., & Kumar, A. (2021). Impact of Machine Learning Applications For Price Prediction Of Onion In India. *Jurnal Ilmu Komputer*, *14*(1), 30. https://doi.org/10.24843/JIK.2021.v14.i01.p04

Andono, P. N., Ocky Saputra, F., Shidik, G. F., & Arifin Hasibuan, Z. (2022). End-to-End Circular Economy in Onion Farming with the Application of Artificial Intelligence and Internet of Things. *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 459–462. https://doi.org/10.1109/iSemantic55962.2022.9920447

Dietterich, T. G. (2000). *Ensemble Methods in Machine Learning* (pp. 1–15). https://doi.org/10.1007/3-540-45014-9_1

Et. al., A. S. (2021). Onion Yield Prediction Based on Machine Learning. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(2). https://doi.org/10.17762/turcomat.v12i2.1972

Hasan, M. M., Tuz Zahara, M., Sykot, M. M., Hafiz, R., & Saifuzzaman, M. (2020). Solving Onion Market Instability by Forecasting Onion Price Using Machine Learning Approach.

2020 International Conference on Computational Performance Evaluation (ComPE), 777–780. https://doi.org/10.1109/ComPE49325.2020.9200033

Islam, S. F. N., Sholahuddin, A., & Abdullah, A. S. (2021). Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah. *Journal of Physics: Conference Series*, *1722*(1), 012016. https://doi.org/10.1088/1742-6596/1722/1/012016

Iswari, N. M. S., Budiardjo, E. K., & Hasibuan, Z. A. (2021). E-business applications recommendation for SMES using advanced user-based collaboration filtering. *ICIC Express Letters*, *15*(5), 517–526. https://doi.org/10.24507/icicel.15.05.517

Iswaya Maalik S, Wisnu Ananta Kusuma, S. W. (2019). ANALISIS PEMBANDINGAN TEKNIK ENSEMBLE SECARA BOOSTING(XGBOOST) DAN BAGGING (RANDOMFOREST) PADA KLASIFIKASI KATEGORI SAMBATAN SEKUENS DNA. *Jurnal Penelitian Pos Dan Informatika*, *9*(1), 27. https://doi.org/10.17933/jppi.2019.090103

Madaan, L., Sharma, A., Khandelwal, P., Goel, S., Singla, P., & Seth, A. (2019). Price forecasting & anomaly detection for agricultural commodities in India. *Proceedings of the Conference on Computing & Sustainable Societies - COMPASS 19*, 52–64. https://doi.org/10.1145/3314344.3332488

Pete Chapman et al. (2000). *CRISP-DM 1.0 Step by Step Data Mining Guide*.

R, N., K, S., R, V. P., & R.V., P. (2020). Onion Price Prediction Based on Artificial Intelligence. *International Research Journal of Multidisciplinary Technovation*, 11–20. https://doi.org/10.34256/irjmt2043

Rizaty, M. A. (2021). *Produksi Bawang Merah Nasional Naik 10,4% pada 2021*. Dataindonesia.Id. https://dataindonesia.id/sektor-riil/detail/produksi-bawang-merah-nasional-naik-104-pada-2021

Saumyamala, M. G. A., Weerasinghe, W., Kumara, J., Sachithra, S. A. L., & Chandrasekara, N. V. (2019). Modelling Open Market Retail Price of Red Onions in Colombo using ARIMA-GARCH MixedModel. *12th International Research Conference, Challenges to Humankind in the Face of New Technologies*, 198.

Triswanda, E., Astutik, S., Nur Hantama, R., Program Studi Sarjana, M., Statistika, J., Universitas Brawijaya, F., & KabMalang, D. (2020). Peramalan harga bawang merah di pasar Kepanjen Kabupaten Malang menggunakan metode Autoregressive Integrated Moving Average (ARIMA). *E-Prosiding Nasional*, *9*(Snso), 2599-2546x. http://prosiding.statistics.unpad.ac.id

Virdaus, D., & Prasetyaningrum, P. T. (2020). Penerapan Data Mining Untuk Memprediksi Harga Bawang Merah Di Yogyakarta Menggunakan Metode K-Nearest Neighbor. *Journal Of …*, *84*, 1–8. http://jisai.mercubuana-yogya.ac.id/index.php/jisai/article/view/15

Widiyaningtyas, T., Ari Elbaith Zaeni, I., & Ismi Zahrani, T. (2020). Food Commodity Price Prediction in East Java Using Extreme Learning Machine (ELM) Method. *Proceedings - 2020 International Seminaron Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, ISemantic 2020*, 93–97. https://doi.org/10.1109/iSemantic50169.2020.9234201

Wihartiko, F. D., Nurdiati, S., Buono, A., & Santosa, E. (2021). Agricultural price prediction models: A systematic literature review. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 2927–2934.

Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Systems with Applications*, *186*,115716. https://doi.org/10.1016/j.eswa.2021.115716