IJASST
International Journal of
Applied Sciences and Smart Technologies

p-ISSN 2655-8564, e-ISSN 2685-943, pp. 169-182, 2024

Volume 06
Issue 01

Clustering and Trend Analysis of Priority Commodities in The Archipelago Capital Region (IKN) Using A Data Mining Approach

Pandu Pangestu[1], Syamsul Maarip [1], Yuldan Nur Addinsyah[1],
Vega Purwayoga[1*],

[1]*Department of Informatics, Faculty of Engineering, Siliwangi University,*
Tasikmalaya, West Java, Indonesia
[*]*Corresponding Author: vega.purwayoga@unsil.ac.id*

(Received 20-12-2023; Revised 19-05-2024; Accepted 21-05-2024)

Abstract

The process of moving the capital requires careful preparation. One thing that needs to be considered is food security in IKN. This research provides recommendations for the main food commodities in IKN by applying data mining. We collect food productivity data available on the official website for East Kalimantan province. These data are processed and grouped into two groups, namely horticulture and livestock products using the K-Means method. After grouping, we predict the increase in productivity of each group using the ARIMA method. This research produces output in the form of grouping commodities into horticulture and livestock products. Productivity results for each type of commodity are displayed from 2016 to 2020 based on data on the official East Kalimantan Province website. Based on this data, predictions are made using the ARIMA method to predict productivity results from 2021 to 2025. Commodities with total productivity are grouped into high-priority commodities. Grouping the amount of productivity is carried out using the clustering method by comparing the amount of productivity for each commodity and producing commodities that are low priority, middle priority, priority and top priority based on the highest to lowest productivity numbers. The cluster quality for grouping horticultural commodities is 99.1%, while the cluster quality for grouping livestock commodities is 87.5%. Hasil prediksi terbaik yaitu ketika memprediksi produksi salak dan slaughter cattle dengan model ARIMA (0, 1, 0) dan ARIMA (2, 2, 2).

Keywords: ARIMA, Clustering, Holticultural Commodities, K-Means, Livestock Commodities

1 Introduction

Food security is a key issue in fulfilling community welfare because it will determine economic, social and political stability in a country [1]. Food security is a challenge for Indonesia considering that Indonesia is an archipelagic and agricultural country. Food security must include factors of availability, distribution and consumption. One way of sustainable development is paying attention to aspects of food security [2]-[3]. Indonesia, as the fourth largest country in the world and a country with a tropical season, certainly has the advantage of diversity in agricultural and livestock products. On the other hand, this can also be a challenge because the limited availability of agricultural and livestock products must be able to meet the increasing needs of society [4].

The migration process and especially the relocation of the capital from Jakarta to East Kalimantan can affect the food needs of the affected areas [5]. There are many factors that will change a region's food needs. Land that was previously agricultural land was replaced with factories or buildings [6]. This transfer of agricultural land can reduce the productivity of agricultural products. Because agricultural output decreases, people's food needs will change. Therefore, food management is needed to regulate agricultural and livestock products so that they are sufficient for the community, especially the new capital area. If it is not managed well, the worst impact will be a food crisis. The food crisis occurs because food commodities are not managed well [7]. Several studies related to grouping priority commodities have been carried out by several researchers. The research [8] has grouped plantation crops using K-Means which has produced several regional groups according to their productivity levels. Other research was conducted by [4] using K-Means to identify commodities with high to low production. According to [9] research, the performance of K-Means for grouping agricultural commodities produces good cluster quality.

Several studies assess that K-Means has good performance, especially in grouping priority commodities in a region. K-Means can be used to group data, but K-Means is not an algorithm that can predict the productivity of a commodity. With the current development of data technology, prediction systems have become very important to prevent problems that will occur in the future, such as food crises. The prediction system can be used as the right solution to overcome the food crisis problem [10]. The results of the predictions can be used as a recommendation for selecting which commodities are considered important [11]. One prediction algorithm with good performance is Autoregressive Integrated Moving Average (ARIMA). The ARIMA algorithm is suitable for determining patterns and trends in data over a certain period of time [12]. This research not only carries out groupings as in previous research, but also carries out a prediction process using ARIMA to determine the need for livestock commodities and horticultural commodities in the future.

2 Researth Methods

Study Area and Research Data

The study area in this research is data on livestock commodities and horticultural commodities in East Kalimantan [9]. East Kalimantan is a province that is planned to become (IKN), thus allowing the need for livestock and horticultural commodities to increase [13]. The data obtained contains horticultural and livestock production data from 2016 to 2020. This research has several stages which are presented in Fig. 1.

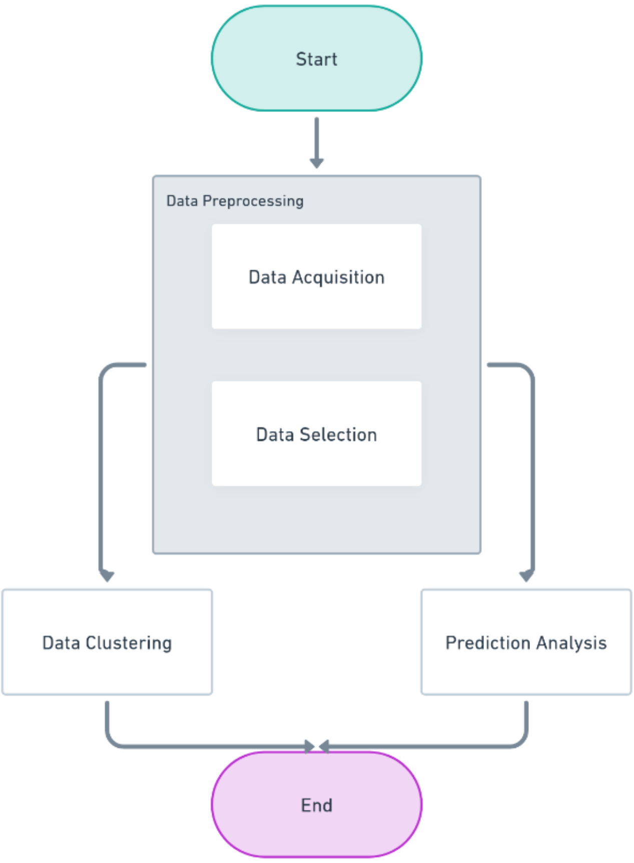

Figure 1. Research Stages

Data Preprocessing

Data preprocessing is carried out in the data analysis process to improve the quality of the data set [14]. Good data analysis results influence the quality of the data [15]. Data preprocessing is one of the main processes in data analysis so it needs to be done. Data preprocessing in this research, namely data collection and data selection [16].

Clustering of Priority Commodities for IKN Regions

Clustering is the process of grouping data to find out whether data belongs to a certain group based on the closeness of the value of an object to other objects [17][18]. This research uses the K-Means algorithm to calsterize priority commodities in the IKN area. K-Means is a popular clustering algorithm which performs well for grouping data [19]. K-Means uses Euclidean distance to measure the closeness of values between objects. The Euclidean distance method has been presented in formula 1.

$$dist(i,j) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \cdots + \left(x_{in} - x_{jn}\right)^2} \qquad (1)$$

Where, x = object, i = (x_{i1}, x_{i2},…, x_{in}), j= (x_{j1}, x_{j2},…, x_{jn}) is two-dimensional object data.

Trend Analysis of Priority Commodity Predictions for IKN Regions

The prediction process is carried out using ARIMA. ARIMA is an algorithm that is reliable in predicting time series data in short periods of time [20]. The prediction process is carried out by utilizing past data to calculate values that will occur in the future.

3 Results and Discussions

Data Preprocessing

In data preprocessing, what is carried out is the data acquisition process obtained from data.kaltimprov.go.id. After the data acquisition process, the data obtained is selected based on the attributes of livestock commodities, horticultural commodities and the amount of production each year. The selected data will be analyzed using clustering methods and trend analysis. This research uses the R programming language. The results of acquisition and pre-processing of livestock and horticultural commodities data are presented in Table 1 and Table 2.

Table 1. Livestock data preprocessing results

Number	Livestock Commodities	Production Amount (Tons)
1	Slaughter Cattle	7310
2	Dairy Cattle	140,76
3	Goat	520,38
4	Sheep	0,15
5	Pig	1835,96
…	…	…
14	Duck Egg Manila	237,82
15	Manila Duck	28,59
16	Quail Egg	68,47
17	Quail	6,65
18	Pigeons	1,38

Table 2. Horticultural data preprocessing results

Number	Horticultural Commodities	Production Amount (Tons)
1	Mango	4310
2	Orange	12692
3	Papaya	15113
4	Banana	95149
5	Pineapple	21948
…	…	…
22	Onion	267
23	Chili	9079
24	Mustard	7694
25	Spring onion	319
26	Cauliflower	207

Clustering of Priority Commodities for IKN Regions

The clustering process is carried out using the K-Means algorithm with the help of the cluster library in R programming language [21]. Before the clustering process is carried out, the data is converted into a standard scale using StandardScaler [22]. The use of StandardScaler aims to increase grouping accuracy. The clustering and categorization of livestock commodities can be seen in Table 3. Meanwhile, the clustering and categorization of horticultural commodities can be seen in Table 4. Quality testing of grouping results was carried out using the total within-cluster sum of square measures method. The cluster quality for grouping horticultural commodities is 99.1%, while the cluster quality for grouping livestock commodities is 87.5%.

Based on cluster quality calculations, the clustering results for livestock commodity data are classified as good. Table 3 shows that the data is grouped into 4 clusters. Cluster 0 consists of eleven commodities, cluster 1 consists of one commodity, cluster 2 consists of three commodities, and cluster 3 also consists of three commodities. Based on the average commodity value of each cluster, cluster members are converted into 4 categories, namely low priority, middle priority, priority, and top priority. A visualization of the livestock commodities category is presented in Fig. 2.

Table 3. Result Clustering Livestock Commodities

Cluster	Livestock Commodities	Category
0	Dairy cattle, Goat, Sheep, Buffalo, Horse, Duck, Duck egg manila, Manila duck, Quail egg, Quail, Pigeons	Low priority
1	Broiler	Top priority
2	Slaughter cattle, Rabbit, Laying hans	Middle priority
3	Pig, Organic chicken, Duck egg	Priority

Based on Fig. 2, many commodities are included in the low priority category. Low priority means that the amount of production in that category is still low. The cluster quality results for horticultural commodities show good results, namely 87.5%, but are lower than the cluster results for livestock commodities. Cluster results for horticultural commodities can be seen in Table 4.

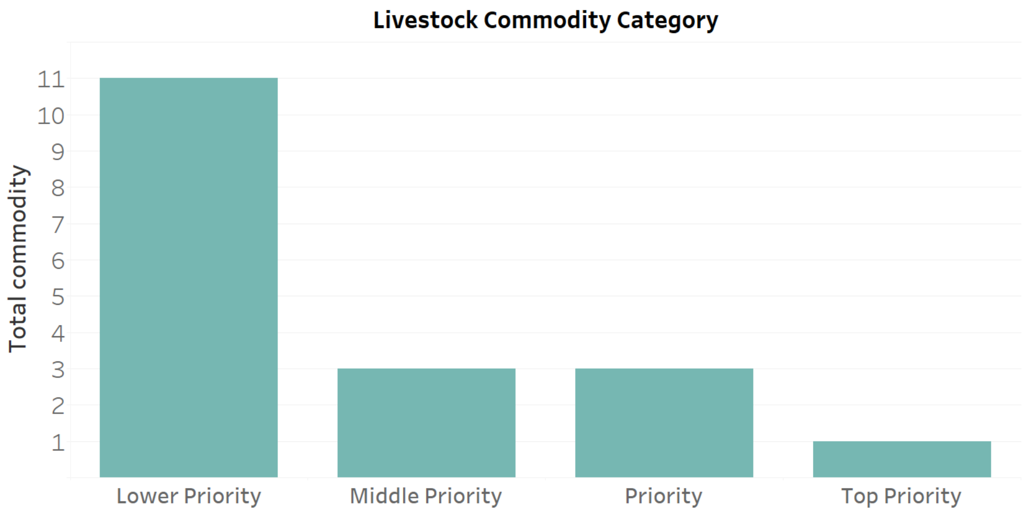

Figure 2. Distrbution of livestock data categorization results

Table 4. Result Clustering Horticultural Commodities

Cluster	Horticultural Commodities	Category
0	Mango, Mangosteen, Melon, Avocado, Starfruit, Guava, Rose Apple, Sapodillah, Soursop, Passion fruit, breadfruit, Melinjo, Onion, Spring onion, Cauliflower	Low priority
1	Banana	Top priority
2	Pineapple, Sneakfruit	Priority
3	Orange, Papaya, Durian, Duku, Jackfruit, Rambutan, Chili, Mustard	Middle riority

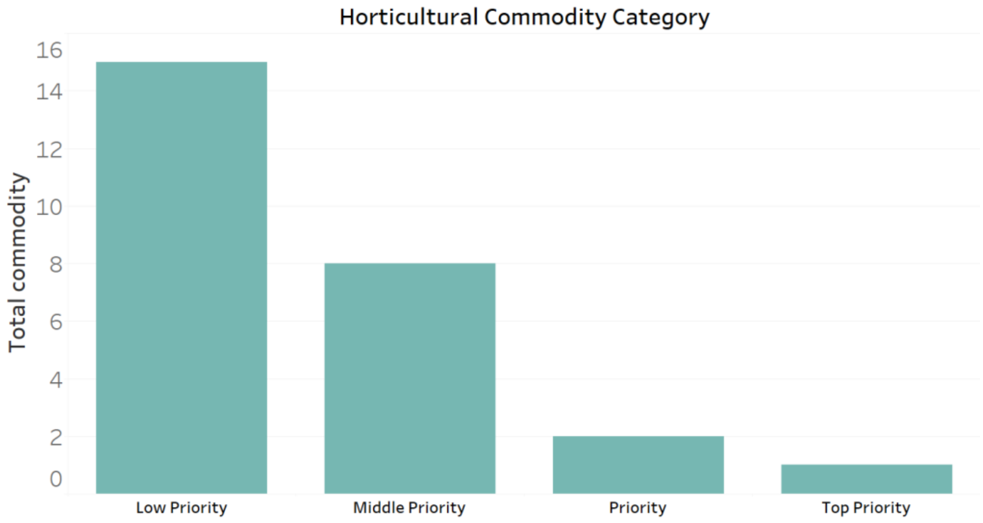

Figure 3. Distrbution of horticultural data categorization results

As presented in Table 4, cluster 0 consists of fifteen commodities, cluster 1 consists of one commodity, cluster 2 consists of two commodities, and cluster 3 consists of eight commodities. A visualization of commodities in a category is presented in Figure 3. Based on Fig. 3, the number of commodities classified as low priority dominates compared to other categories. In top priority there is only one commodity, namely bananas. This shows that only bananas have very high productivity.

Trend Analysis of Priority Commodity Predictions for IKN Regions

The ARIMA method is applied to livestock and horticultural data using the tseries and forcaste library. The steps involve training a model using historical data to make data predictions for subsequent years. The prediction results for livestock and horticultural commodities using them are presented in Tables 5 and 6. Predicted trends in production quantities for each commodity are presented in Fig. 4 and Fig. 5.

A visualization of livestock commodity trends can be seen in Fig. 4. The data used to make predictions is from 2016 to 2020. Prediction accuracy assessment is carried out by looking at the Mean Absolute Percentage Error (MAPE) value. The best MAPE is produced when predicting Slaughter Cattle production, namely 21% with the ARIMA model (0, 1, 0).

Table 5. Result Clustering Livestock Commodities

Number	Livestock Commodities	Total production in 2016	Total production in 2025
1	Slaughter Cattle	65369,34	291752,1
2	Dairy Cattle	7310	8000,88
3	Goat	6911,08	29077,89
4	Sheep	5202,76	5270,04
5	Pig	1468,12	871,35
…	…	…	…
14	Duck Egg Manila	237,82	146,23
15	Manila Duck	28,59	24,52
16	Quail Egg	68,47	42,17
17	Quail	64,15	58,21
18	Pigeons	1,38	1,27

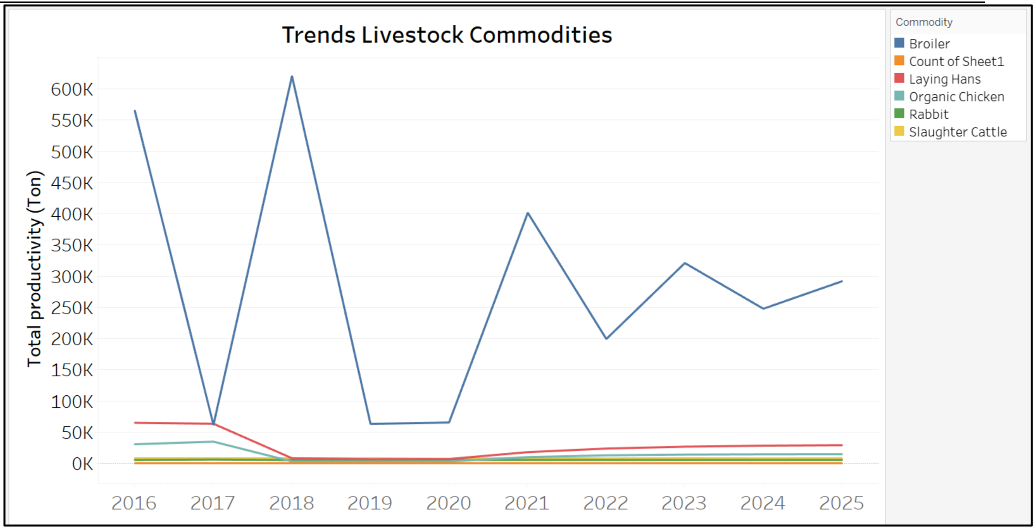

Figure 4. Livestock data productivity trends

Table 6. Horticultural commodities predictions

Number	Livestock Commodities	Total production in 2020	Total production in 2025
1	Mango	95149	78776,45
2	Orange	21948	22592,55

3	Papaya	19850	88508,54
4	Banana	12692	16195,48
5	Pineapple	15113	17836,33
…	…	…	…
22	Onion	1158	12965,3
23	Chili	519	487,98
24	Mustard	463	389,27
25	Spring onion	3683	36473,13
26	Cauliflower	207	136,12

A visualization of horticultural commodity trends can be seen in Fig. 5. The data used to make predictions is from 2016 to 2020. Each horticultural commodity is predicted using ARIMA to determine the production level of each commodity. The best ARIMA model was produced when predicting snake fruit production, namely 20%. The best ARIMA model is (2, 2, 2).

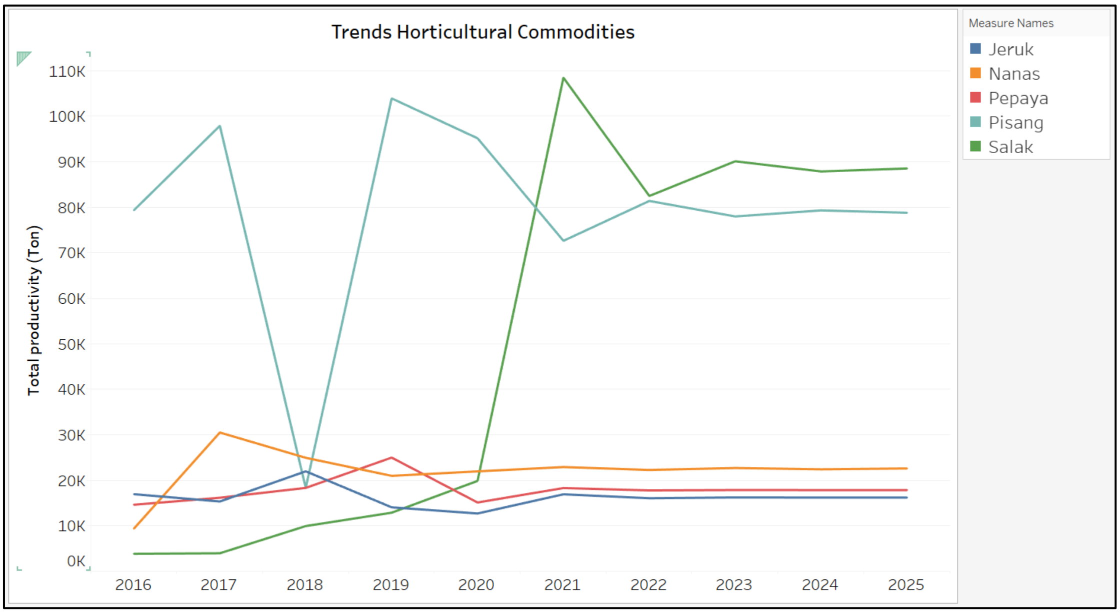

Figure 5. Horticulture data productivity trends

4 Conclusions

This research has succeeded in grouping horticultural commodities and livestock products using the K-Means method with good quality. Not only does this research carry out groupings, it also predicts the productivity of each commodity by applying the ARIMA method. The best ARIMA model quality for predicting horticultural commodities and livestock commodities is ARIMA (2, 2, 2) and ARIMA (0, 1, 0). The MAPE value of the ARIMA model cannot be said to be ideal, because the MAPE value is still ≥ 20%. The ideal MAPE value is ≤ 20% [19].

Thus, this research makes an important contribution to decision making regarding food policy in the IKN Region. The resulting priority commodity recommendations can become a basis for optimizing food production and achieving food independence in the region. The use of the K-Means and ARIMA methods in predictions also provides reliability in projecting productivity results for the future. It is hoped that future research will pay attention to aspects of public consumption in the IKN area for certain commodities. When the production level of a commodity is low, but demand is high, the commodity must be developed. The production development process can be carried out by expanding the land for a horticultural commodity using a land suitability analysis approach.

References

[1] R. Harini, I. Sukri, R. D. Ariani, E. P. I. Faroh, H. Nadia, and U. Kafafa, "The Study of Food Security in the Special Region of Yogyakarta, Indonesia," *Forum Geografi*, vol. 35, no. 2, Feb. 2022, doi: 10.23917/forgeo.v35i2.15855.

[2] A. Garbero and L. Jäckering, "The potential of agricultural programs for improving food security: A multi-country perspective," *Glob Food Sec*, vol. 29, p. 100529, Jun. 2021, doi: 10.1016/j.gfs.2021.100529.

[3] T. T. Tora, D. T. Degaga, and A. U. Utallo, "Drought vulnerability perceptions and food security status of rural lowland communities: An insight from Southwest Ethiopia," *Current Research in Environmental Sustainability*, vol. 3, p. 100073, 2021, doi: 10.1016/j.crsust.2021.100073.

[4] N. Endey, I. Kadek, S. Arsana, A. Y. Katili, A. Sahabi, and M. A. Talalu, “Analisis Daya Saing Komoditi Unggulan Gorontalo Dalam Mendukung Ibu Kota Negara Baru Republik Indonesia,” vol. 3, [Online]. Available: http://journal.unismuh.ac.id/index.php/equilibrium

[5] A. Zezza, C. Carletto, B. Davis, and P. Winters, “Assessing the impact of migration on food and nutrition security,” *Food Policy*, vol. 36, no. 1, pp. 1–6, Feb. 2011, doi: 10.1016/j.foodpol.2010.11.005.

[6] S. Min, L. Hou, W. Hermann, J. Huang, and Y. Mu, “The impact of migration on the food consumption and nutrition of left-behind family members: Evidence from a minority mountainous region of southwestern China,” *J Integr Agric*, vol. 18, no. 8, pp. 1780–1792, Aug. 2019, doi: 10.1016/S2095-3119(19)62588-8.

[7] A. Parven *et al.*, “Impacts of disaster and land-use change on food security and adaptation: Evidence from the delta community in Bangladesh,” *International Journal of Disaster Risk Reduction*, vol. 78, p. 103119, Aug. 2022, doi: 10.1016/j.ijdrr.2022.103119.

[8] A. A. Simangunsong, I. Gunawan, Z. M. Nasution, and G. Artikel, “Pengelompokkan Hasil Produksi Tanaman Perkebunan Berdasarkan Provinsi Menggunakan Metode K-Means Clustering Production of Plantation Crops by Province Using the K-Means Method Article Info ABSTRAK,” *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 4, pp. 2828–9099, 2022, doi: 10.55123/jomlai.v1i4.1661.

[9] L. M. Harahap, W. Fuadi, L. Rosnita, E. Darnila, and R. Meiyanti, “Klastering Sayuran Unggulan Menggunakan Algoritma K-Means,” *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 3, Dec. 2022, doi: 10.28932/jutisi.v8i3.5277.

[10] S. Nosratabadi, S. Ardabili, Z. Lakner, C. Mako, and A. Mosavi, “Prediction of Food Production Using Machine Learning Algorithms of Multilayer Perceptron and ANFIS,” *Agriculture*, vol. 11, no. 5, p. 408, May 2021, doi: 10.3390/agriculture11050408.

[11] R. Nugroho, A. Polina, and Y. Mahendra, “Tourism Site Recommender System Using Item-Based Collaborative Filtering Approach,” *International Journal of*

Applied Sciences and Smart Technologies, vol. 2, no. 2, pp. 119–126, Dec. 2020, doi: 10.24071/ijasst.v2i2.2987.

[12] J. Fattah, L. Ezzine, Z. Aman, H. El Moussami, and A. Lachhab, “Forecasting of demand using ARIMA model,” *International Journal of Engineering Business Management*, vol. 10, p. 184797901880867, Jan. 2018, doi: 10.1177/1847979018808673.

[13] A. H. Salsabila and N. Nurwati, “Deforestasi dan Migrasi Penduduk ke Ibu Kota Baru Kalimantan Timur: Peran Sinergis Pemerintah Dan Masyarakat,” *Prosiding Penelitian dan Pengabdian kepada Masyarakat*, vol. 7, no. 1, p. 27, Jul. 2020, doi: 10.24198/jppm.v7i1.28259.

[14] K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, “Data preprocessing,” in *Computational Learning Approaches to Data Analytics in Biomedical Applications*, Elsevier, 2020, pp. 7–27. doi: 10.1016/B978-0-12-814482-4.00002-4.

[15] V. Purwayoga and I. S. Sitanggang, “Clustering Potential Area of Fusarium Oxysporum As A Disease of Garlic,” in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics Publishing, Jul. 2020. doi: 10.1088/1755-1315/528/1/012040.

[16] V. Purwayoga, “Modified skyline query to measure priority region for personal protective equipment recipient of COVID-19 health workers,” *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 3, pp. 167–173, Jul. 2021, doi: 10.14710/jtsiskom.2021.14003.

[17] V. Purwayoga, “Optimasi Jumlah Cluster pada Algoritme K-Means untuk Evaluasi Kinerja Dosen,” *Jurnal Informatika Universitas Pamulang*, vol. 6, no. 1, p. 118, Mar. 2021, doi: 10.32493/informatika.v6i1.9522.

[18] E. H. S. Atmaja, “Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta,” *International Journal of Applied Sciences and Smart Technologies*, vol. 1, no. 1, pp. 33–44, Jun. 2019, doi: 10.24071/ijasst.v1i1.1859.

[19] H. Jurnal, S. Budi, and H. Sakur, "Jurnal Informatika dan Tekonologi Komputer Perbandingan Distance Measures Pada K-Means Cluster Dan Topsis Dengan Korelasi Pearson Dan Spearman," *Maret*, vol. 3, no. 1, pp. 74–81, 2023.

[20] L. N. Kasanah, "Aplikasi Autoregressive Integrated Moving Average (ARIMA) untuk Meramalkan Jumlah Demam Berdarah Dengue (DBD) di Puskesmas Mulyorejo," *Jurnal Biometrika dan Kependudukan*, vol. 5, no. 2, p. 177, Sep. 2017, doi: 10.20473/jbk.v5i2.2016.177-189.

[21] S. D. Sudrazat, H. Purba, E. Wijaksono, W. Pranowo, and M. I. Hibatullah, "Prediksi Kecepatan Gelombang S Dengan Machine Learning Pada Sumur 'S-1', Cekungan Sumatera Tengah, Indonesia," *Lembaran publikasi minyak dan gas bumi*, vol. 54, no. 1, pp. 29–35, Apr. 2020, doi: 10.29017/LPMGB.54.1.502.

[22] N. D. Arianti, E. Saputra, and A. Sitorus, "An automatic generation of pre-processing strategy combined with machine learning multivariate analysis for NIR spectral data," *J Agric Food Res*, vol. 13, p. 100625, Sep. 2023, doi: 10.1016/j.jafr.2023.100625.