The 4th International Seminar of Science and Technology
ISST 2024 Vol 4 (2025) 024
Innovations in Science and Technology to Realize Sustainable Development Goals
Faculty of Science and Technology
Universitas Terbuka

Analytical Dashboard Development for Agricultural Commodities Using Data Mining to Support Food Security

Ronny Susetyoko*, Iwan Syarif, Alfi Fadliana, and Abdul Muffid

Politeknik Elektronika Negeri Surabaya, Department of Informatics and Computer Engineering, Surabaya, Indonesia, 60111

Abstract - Sustainable Development Goals (SDGs) include no poverty, zero hunger, and the achievement of food security. Requirements for food security include 1) adequate availability, 2) stability of availability, and 3) accessibility. A region's food availability can be represented by the potential of its agricultural commodities. The stability of food availability can be seen from time series data on food crop production. Analytical dashboards have become an urgent application platform available to agricultural stakeholders to monitor, predict, map and position commodities as a basis for decision making and establishing policies/programs to support national food security. The aim of this research is to develop an analytical dashboard using data mining. Several data mining techniques were used to build this dashboard. Agricultural commodity predictions use the Autoregressive Integrated Moving Average (ARIMA) and Neural Network (NN) methods. Commodity mapping uses the K – Means and Ordering Points to Identify Clustering Structure (OPTICS) method. The food security model for positioning a region uses Factor Analysis to select significant factors. For food security classification using ordinal Logistic Regression and Random Forest. The best performing methods are implemented into analytical dashboards as needed.

Keywords: Analytics, dashboard, data mining, agricultural, food security

1 Introduction

Food security is a global challenge amidst population growth, climate change and fluctuating economic dynamics. For countries that are highly dependent on the agricultural sector, food security is highly dependent on food availability, accessibility and stability of the supply of agricultural commodities. To overcome these challenges and support food security, the use of data-based analytical technology has been developed, one of which is an analytical dashboard. The aim of this research is to develop an analytical dashboard that uses several data mining techniques, including data exploration, prediction, clustering, and food security classification.

Many studies related to forecasting use the ARIMA method. The ARIMA method is compared with other methods to predict profits for the next 5 years [1]. A combination of ARIMA and NN methods is also used to predict water temperature to estimate fish catches [2]. The combination of ARIMA and LSTM models is stated to be effective in accommodating linear and nonlinear components in epidemic data [3]. In the field of clustering, in [4] It is stated that the development of an unsupervised K-Means algorithm can find the optimal number of clusters without initialization

*Corresponding author: rony@pens.ac.id

and parameter selection. To group data with high density, DBSCAN can be used [5]. Meanwhile, to group overlapping and low-dimensional data, you can use the Automatic Fuzzy-DBSCAN (AFD) method.[6].

Currently, dashboard development is tailored to customer needs to improve user experience [7][8]. Dashboards should support both analytical and communicative objectives [8]. The novelty in this research is that the dashboard was developed using data mining according to customer needs. The forecasting methods used to predict land area and food crop production are ARIMA, Exponential Smoothing, and Neural Network (NN). These three methods were selected due to their complementary strengths in time series forecasting. ARIMA is effective in modeling linear trends and autocorrelations in stationary data. Exponential Smoothing is suitable for capturing level, trend, and seasonal components with relatively simple computations. Neural Network, on the other hand, is capable of modeling complex nonlinear relationships and uncovering hidden patterns within the data. Together, these approaches accommodate both linear and nonlinear dynamics in agricultural time series data to map the potential of food and horticultural crops, the K–Means and OPTICS (Ordering Points to Identify Clustering Structure) methods are used. Meanwhile, Ordinal Logistic Regression and Random Forest methods are used to classify the level of food security. Blitar Regency was used as a research object because it has great potential in the agricultural sector [9].

2 Materials and methods

2.1 Dataset

The dataset used in this research is sourced from the Blitar Regency Central Statistics Agency and the National Food Agency, which includes:

a. data on land area and production of food and horticultural crops;
b. data related to food security factors, namely: number of minimarkets, number of restaurants, number of stalls, number of food stalls, food and energy security credits, availability of village public transport with fixed routes, availability of village public transport without fixed routes, there are no village public transportation, average expenditure, rice production, corn production, sweet potato production, cassava production, Human Development Index (HDI), average worker wages, open unemployment rate, poverty percentage, and Gross regional domestic product (GRDP).

2.2 Methodology

The methodology in this research is shown in Figure 1. Data was collected from the websites of the Central Statistics Agency (BPS) and the National Food Agency. Before analysis, the data was cleaned first to ensure there was no missing data. Next, data exploration is carried out to find insights. There are several analyzes carried out in this research. Exponential Smoothing, ARIMA, and NN (Neural Network) methods to predict land area and production.

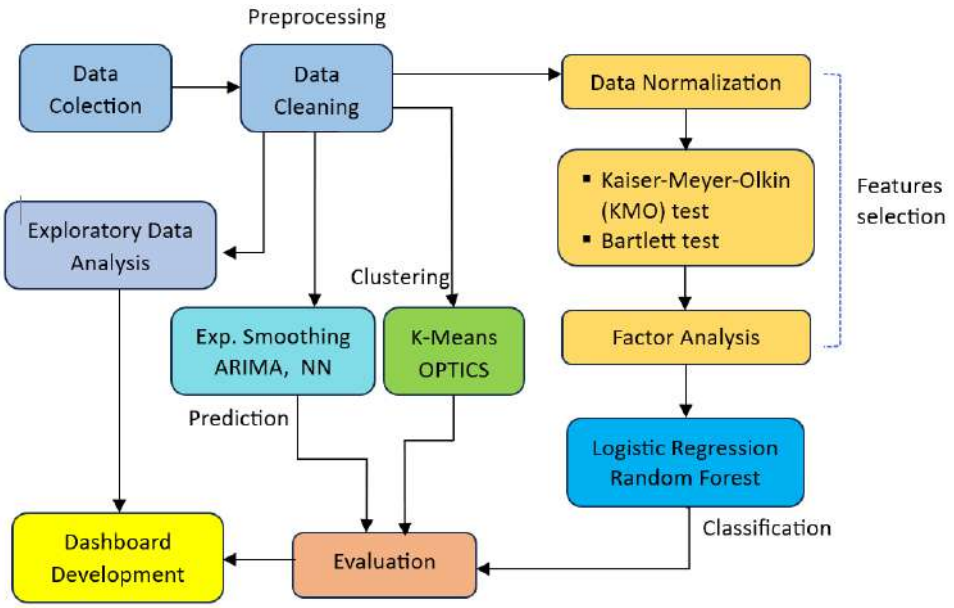


Fig. 1. System Design

The prediction method is evaluated using MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and RMSE (Root Mean Square Error) to provide a more comprehensive evaluation of model performance. K-Means and OPTICS methods were used for grouping land area and production. These clustering methods were compared using the Silhouette score, Davies-Bouldin Index, and Calinski-Harabasz Index to provide a more robust evaluation of clustering performance. The classification process commenced with data normalization to ensure a uniform scale, followed by an assessment of data adequacy using the Kaiser-Meyer-Olkin (KMO) test and the evaluation of inter-feature significance through the Bartlett test. The KMO test, yielding a value above 0.6, was selected to measure sampling adequacy by assessing the proportion of variance explained by the underlying factors, thereby confirming that the data exhibited sufficient correlation structure for factor analysis (Kaiser, 1974). Conversely, the Bartlett test was employed to test the hypothesis of significant correlations among variables or features, supporting the validity of factor analysis (Bartlett, 1950). Subsequently, factor analysis was conducted to extract significant features relevant to food security classification modeling. The classification methods, namely Logistic Regression and Random Forest, were evaluated based on the highest accuracy achieved during the evaluation phase to determine the optimal approach.

3 Results and discussion

3.1 Exploratory Data Analysis

Several methods the distribution of rice harvested land area from year to year tend to fluctuate. The median value of rice harvested land area in 2018 was higher than in subsequent years. There was a large decrease in 2019 and 2020 when compared to 2018, but the variability tends to be constant, shown in Figure 2(a). Likewise, the distribution of land area harvested for corn commodities is shown in Figure 2(b). When compared, the area of harvested land between rice and corn commodities is relatively balanced.

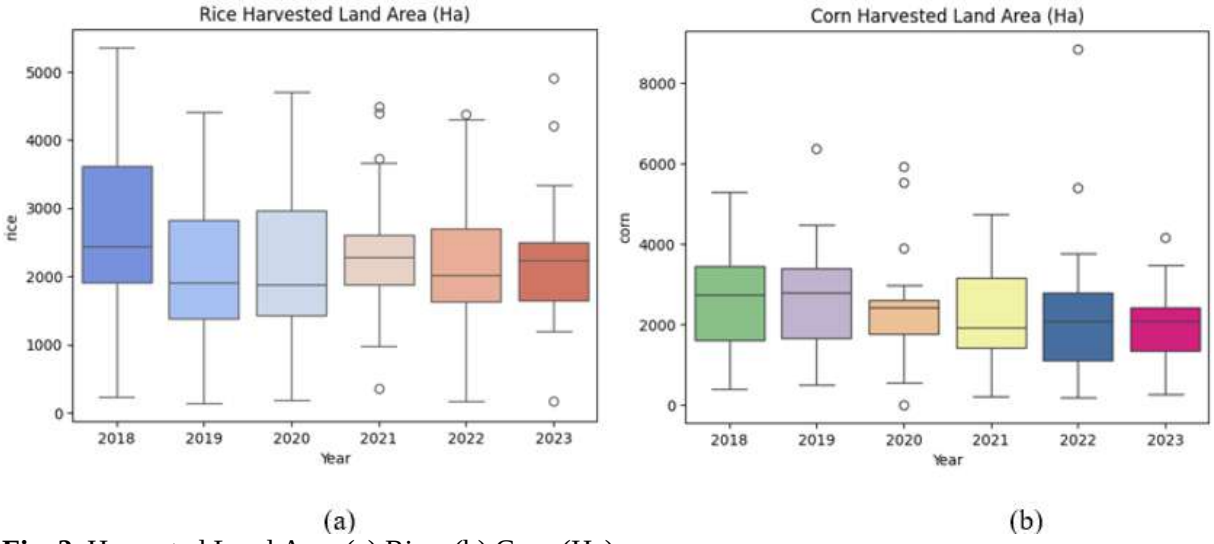


Fig. 2. Harvested Land Area (a) Rice, (b) Corn (Ha)

3.2 Prediction

Several prediction methods to predict harvest area and production of food and horticultural crop commodities use several methods, namely Exponential Smoothing, ARIMA, and Neural Network (NN). Based on the comparative analysis of the three methods, the Double Exponential Smoothing method shows suitable results due to limited time series data. The autocorrelation function (ACF) plot and forecasting results for corn commodities are shown in Figure 3(a). ACF for lag 1 to lag 5 is not significant so it is not possible to model using ARIMA. The forecasting results are shown in the green line in Figure 3(b) which shows the trend trend.

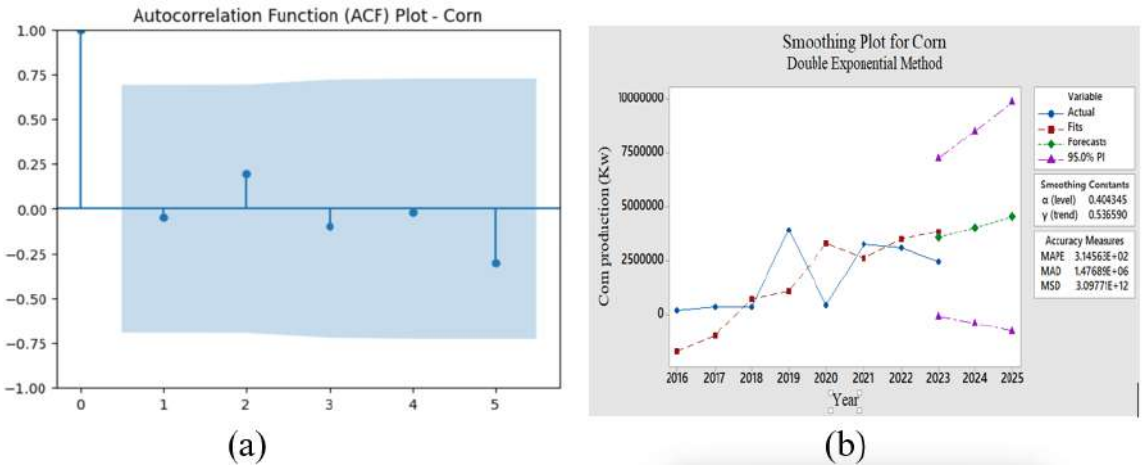
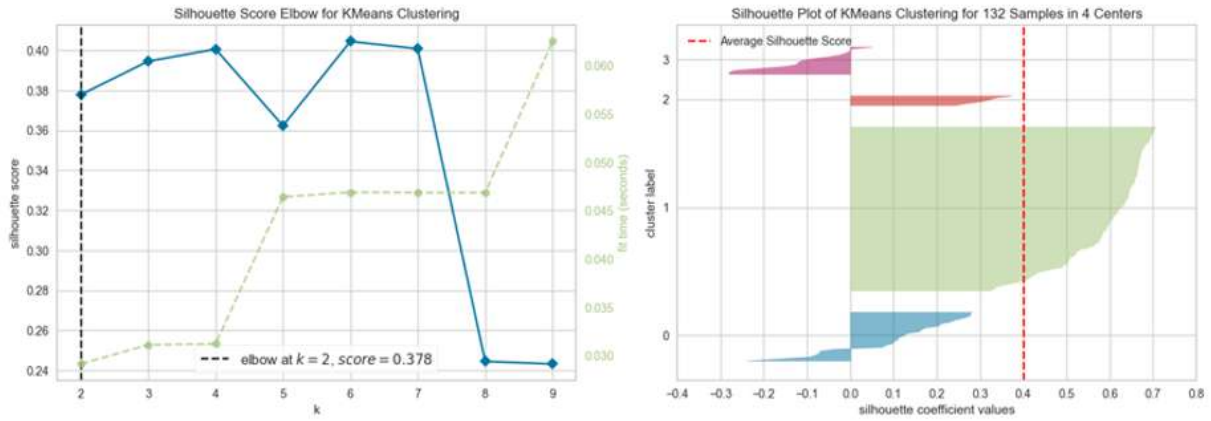


Fig. 3. ACF Plot of corn production (a) and forecasting (b)

3.3 Clustering

Silhouette scores of land area clusters for k = 1 to 9 are shown in Figure 4(a). For this reason, k = 4 was chosen considering that the Silhouette score is quite large and makes interpretation easier too. Meanwhile, Figure 4(b) is a Silhouette plot for 132 samples. Next, the K-Means method is compared with OPTICS. From data processing using Jupyter Notebook, a Silhouette score of 0.40052 was obtained for the K-Means method and -0.2442 for the OPTICS method. Meanwhile, the Davies-

Bouldin Index for the K-Means and OPTICS methods is 1.5912 and 1.7257 respectively. Thus, the K-Means method was chosen to be implemented in dashboard visualization.



(a) (b)

Fig. 4. Silhouette Score Elbow (a), and Silhouette Plot for 132 samples

3.4 Classification

Classification was applied to determine the level of food security across districts and cities in East Java through a data-driven approach. The Kaiser-Meyer-Olkin (KMO) test yielded a value of 0.7913, falling within the 0.7–0.8 range, indicating a strong sampling adequacy for factor analysis and confirming the data's suitability for feature extraction. Meanwhile, the Bartlett test produced a Chi-Square value of 2013.6297 with a p-value of 0.0 ($p < 0.05$), significantly exceeding the critical Chi-Square value at a 0.05 significance level, thereby indicating a significant correlation among variables and rejecting the null hypothesis that the correlation matrix is an identity matrix. Based on these results, factor analysis was conducted to identify the number of underlying factors shaping the food security structure, which were subsequently utilized for classification modeling.

The results of factor analysis are shown in Table 1. By taking variables with absolute factor coefficients > 0.5, there are 3 factors that form the structure. Factor 1 is represented as economic growth, namely the variables: 1) average expenditure, 2) HDI, 3) average worker wages, 4) open unemployment rate, 5) poverty rate, and 6) GRDP. Factor 2 is represented as food accessibility, including: 1) number of minimarkets, 2) number of restaurants, and 3) availability of village public transportation with fixed routes. Meanwhile, factor 3 is represented as food availability, including: 1) rice production, 2) corn production, and 3) number of food stalls.

Table 1. Loading Factor

No.	Features	Factor 1	Factor 2	Factor 3
1	rice_prod	-0.165945	0.063623	*0.799740*
2	corn_prod	-0.374558	0.208156	*0.607125*
3	swpot_prod	-0.022663	0.065565	0.213628
4	cass_prod	-0.462949	0.373611	-0.023851
5	no_minmar	0.203372	*0.818168*	0.325038
6	no_restau	0.485426	*0.715117*	0.177308
7	no_fstall	-0.108004	0.426708	*0.802014*
8	avpt_with	-0.033046	*0.837858*	0.351402
9	ave_exp	*0.876203*	0.041936	-0.322327
10	hdi	*0.842092*	0.044975	-0.344103

No.	Features	Factor 1	Factor 2	Factor 3
11	ave_wages	*0.799871*	0.162217	-0.132838
12	op_unempl	*0.669961*	0.072585	-0.002887
13	poverty	*-0.617307*	-0.088098	0.378461
14	grdp	*0.556019*	0.401647	-0.012890

The evaluation was carried out by comparing the accuracy values of the 6 models studied, the results are shown in Figure 5. In general, the accuracy of the 3 Random Forest models is higher than Logistic Regression. At this stage, the Random Forest - SMOTE model was selected with the highest accuracy value of 0.93.

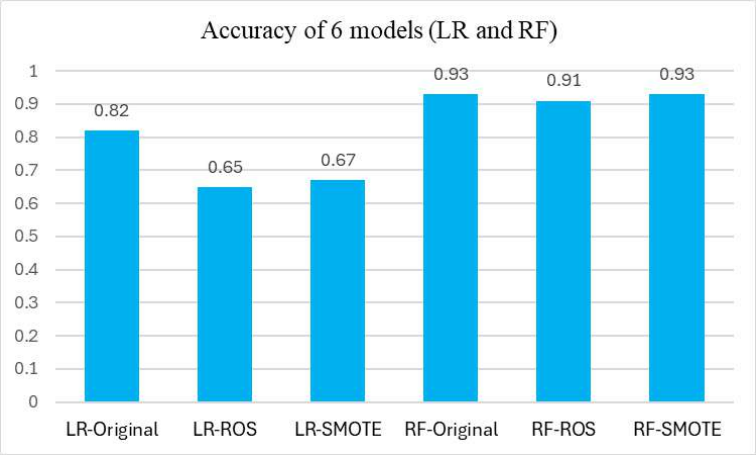


Fig. 5. Accuracy of 6 models (LR and RF)

3.5 Analytics Dashboard

The results of the analytical dashboard are shown in Figure 6 and Figure 7. Page Home is a dashboard visualization that focuses on exploring the harvest area and production of food and horticultural crop commodities shown in Figure 6. The top row shows the total production and harvest area of food crop commodities for 2022.

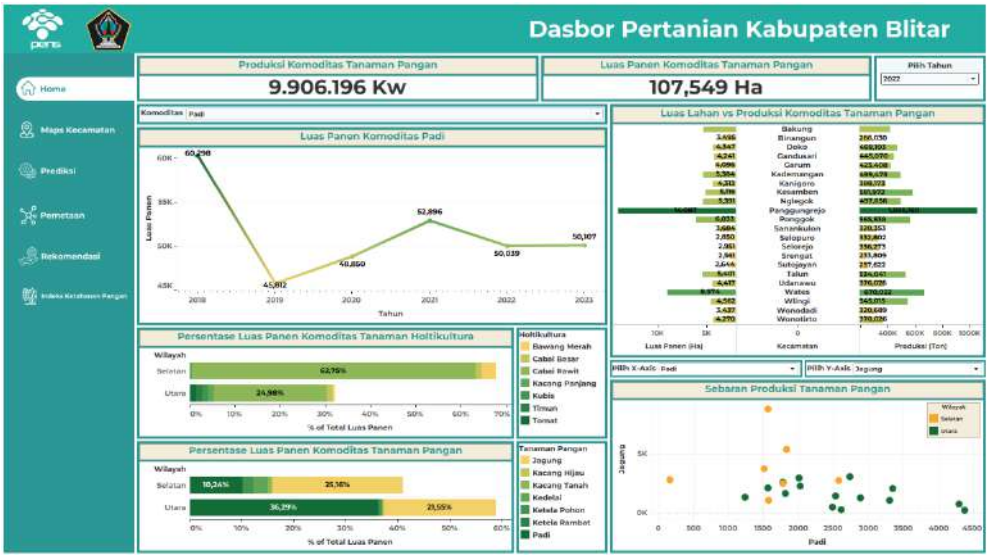


Fig. 6. Analytical Dashboard: Home Page

The time series plot for each commodity can be seen in the top left image, which illustrates food stability. In the bottom left image, there is a stacked bar plot showing the percentage of harvested area for horticultural commodities and food crops which are differentiated based on areas north and south of the Brantas river. This will make it easier for users to find out the dominant potential in the sub-districts north and south of the Brantas river. The top right image is the land area and production in each sub-district. From this picture, it can be seen which regions are superior in certain commodities. Meanwhile, the bottom right image is the distribution of production of 2 selected commodities based on the areas north and south of the Brantas river.

The visualization effectively presents regional agricultural potential by implementing the results of K-Means clustering to describe the potential map of each sub-district in Blitar Regency, as shown in Figure 7, thereby facilitating users in interpreting food stability and dominant commodities. The left image is a map based on the results of clustering of harvest area, which is divided into 4 clusters. Cluster-1 is dominated by land for harvesting sweet potatoes and vegetables, cluster-2 is dominated by land for harvesting food crops, cluster-3 is dominated by land for harvesting corn, soybeans, cassava and cayenne pepper, and cluster-4 is dominated by peanuts and large chilies. The centroids of the 4 clusters are presented in the table below.

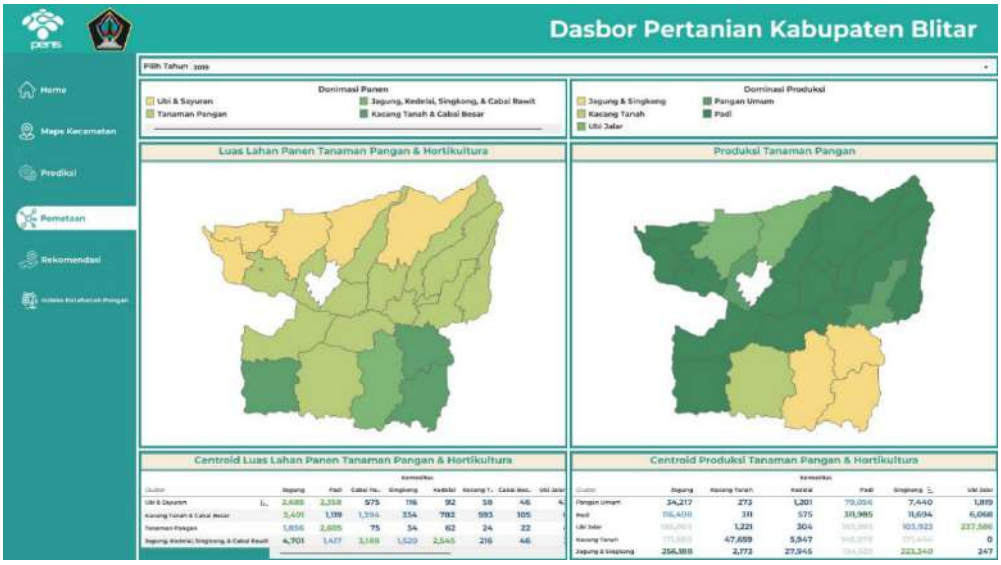


Fig. 7. Analytical Dashboard: Mapping Page (Clustering)

Likewise, the clustering results for food crop production are shown in the right image. There are 5 clusters formed, namely cluster-1 is dominated by corn and cassava production, cluster-2 is dominated by peanut production, cluster-3 is dominated by sweet potato production. Meanwhile, cluster-4 is general food crop production, and cluster-5 is dominated by rice production. The centroids of the 5 clusters are presented in the table below.

4 Conclusion

Based on the results of the analysis, there are 3 food security factors formed from factor analysis, namely: economic growth, food accessibility, and food availability. The study provides valuable insights into food security factors by conducting several explorations to obtain data insights, which were then visualized into an analytical dashboard using well-chosen models, thereby enhancing decision-making for agricultural stakeholders. The most appropriate prediction model from the

available data is Single Exponential Smoothing. Among the two clustering methods evaluated, the K-Means method was selected as the preferred approach due to its superior performance in capturing data patterns. For classifying food security levels, the RF-SMOTE model, which applies the Random Forest algorithm with dataset resampling using the SMOTE method, was identified as the best model, demonstrating high accuracy and robustness.

Data exploration and visualization becomes an interesting performance that can be easily understood by stakeholders. Time series data on land area and production can be used to see the stability of food availability. The analytical dashboard platform has become an urgent application available for agricultural stakeholders to monitor, predict, map and position commodities as a basis for decision making and establishing policies/programs to support national food security.

5 Acknowledgements

The author would like to thank the Surabaya State Electronics Polytechnic, especially the Center for Research and Community Service, which has provided facilities and guidance in this research. Thank you also to the Academic Directorate of Vocational Higher Education, Directorate General of Vocational Education for providing research funding.

References

[1] U. M. Sirisha, M. C. Belavagi, and G. Attigeri, "Profit Prediction Using ARIMA, SARIMA and LSTM Models in Time Series Forecasting: A Comparison," *IEEE Access*, vol. 10, no. November, pp. 124715–124727, 2022, doi: 10.1109/ACCESS.2022.3224938.

[2] Y. Zhang, M. Yamamoto, G. Suzuki, and H. Shioya, "Collaborative Forecasting and Analysis of Fish Catch in Hokkaido From Multiple Scales by Using Neural Network and ARIMA Model," *IEEE Access*, vol. 10, pp. 7823–7833, 2022, doi: 10.1109/ACCESS.2022.3141767.

[3] Y. C. Jin, Q. Cao, K. N. Wang, Y. Zhou, Y. P. Cao, and X. Y. Wang, "Prediction of COVID-19 Data Using Improved ARIMA-LSTM Hybrid Forecast Models," *IEEE Access*, vol. 11, no. June, pp. 67956–67967, 2023, doi: 10.1109/ACCESS.2023.3291999.

[4] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[5] Z. Li, Y. Li, W. Lu, and J. Huang, "Crowdsourcing Logistics Pricing Optimization Model Based on DBSCAN Clustering Algorithm," *IEEE Access*, vol. 8, pp. 92615–92626, 2020, doi: 10.1109/ACCESS.2020.2995063.

[6] Y. Aref, K. S. Cemal, Y. Asef, and S. Amir, "Automatic fuzzy-DBSCAN algorithm for morphological and overlapping datasets," *J. Syst. Eng. Electron.*, vol. 31, no. 6, pp. 1245–1253, 2020, doi: 10.23919/JSEE.2020.000095.

[7] A. Vazquez-Ingelmo, F. J. Garcia-Penalvo, and R. Theron, "Information Dashboards and Tailoring Capabilities-A Systematic Literature Review," *IEEE Access*, vol. 7, pp. 109673–109688, 2019, doi: 10.1109/ACCESS.2019.2933472.

[8] V. Setlur, M. Correll, A. Satyanarayan, and M. Tory, "Heuristics for Supporting Cooperative Dashboard Design," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 1, pp. 1–11, 2023, doi: 10.1109/tvcg.2023.3327158.

[9] Badan Pusat Statistik Kabupaten Blitar, "Statistik Kesejahteraan Rakyat Kabupaten Blitar 2023."