Development of Prediction System for Crude Palm Oil (CPO) Production with Time Series Data Mining Approach

Achmad Solichin
Faculty of Information Technology
Universitas Budi Luhur
Jakarta, Indonesia
achmad.solichin@budiluhur.ac.id

Uswatun Hasanah
Faculty of Information Technology
Universitas Budi Luhur
Jakarta, Indonesia
uswah.987@gmail.com

Jayanta
Faculty of Computer Science
Universitas Pembangunan Nasional
Veteran Jakarta
Jakarta, Indonesia
jayanta@upnvj.ac.id

***Abstract*—Palm oil is a plantation commodity that snowballs when compared to other plantation crops such as coffee or cocoa. The Indonesian palm oil industry has a comparative advantage in the form of a large area of land and the lowest production cost of Crude Palm Oil (CPO) in the world. Indonesia's palm oil production in August 2019 recorded an increase of 14% over the same period in 2018. However, the amount of Indonesia's CPO production can still be optimized and increased. The amount of CPO production is very dependent on several factors, such as weather conditions, land area, and the number of Fresh Fruit Bunches (FFB). To help the Palm Oil Mill (POM), this study compares three data mining algorithms to predict the amount of CPO production based on the number of FFBs. The algorithms being compared are multilayer perceptron (MLP), support vector regression (SVR), and linear regression (LR). Based on test results using test data from a palm oil company in Indonesia, the SVR algorithm can provide higher accuracy than the other two algorithms. The SVR gets a PTA value of 0.694, MSE of 955.002, MAPE of 55.169, and MAD of 22.227. Then, we developed a prototype that applied the SVR algorithm to predict the amount of CPO production. The SQA test results on the prototype resulted in 80.225 software quality in the good category.**

Keywords—crude palm oil, data mining, support vector regression, multilayer perceptron, linear regression

I. INTRODUCTION

Indonesia is one of the leading palm oil-producing countries in the world. Oil palm plantations increase compared to other types of plantation crops. The Indonesian palm oil industry has a comparative advantage in the form of a large area of land and the lowest production cost of Crude Palm Oil (CPO) in the world. Indonesia's palm oil production in August 2019 recorded an increase of 14% over the same period in the previous year [1]. Based on 2017 data, the amount of palm oil production in the form of the Crude Palm Oil (CPO) in several years continues to increase, especially in the case of people's plantation (PR) by 33.88% and the Private Large Plantation (PLP) of 58.56 % [2]. Meanwhile, production from the State Plantation (SP) is relatively slow, at 7.55% [2].

The number of Palm Oil Mills (POM) also increases every year. In 2012 there were 695 POM units with a capacity of 37,213 tons of FFB per hour and increased in 2013 to 713 units with a capacity of 34,628 tons of FFB per hour. In other words, an increase in the number of POM by 2.59% [3]. Also, during the 2011-2015 period, world CPO production and consumption continued to grow, increasing by an average of 4.81% and 5.54% per year [4]. In 2016, the share of world CPO production reached 40% of the total world's main vegetable, while soybean oil had a share of 33.18% [5].

This research was conducted at a Palm Oil Mill (POM), which has an installed capacity of 45 tons of FFB / hour with working hours as much as 20 hours/day. In the period of 1 January 2016 to 31 December 2016, POM operates for 195 days, so ideally, POM can provide as much as 3,900 hours of processing time/year with a total of FFB though around 175,500 tons/year. If the factory stagnation hour caused by a process failure due to engine damage that is permitted is 5% of the available processing hours, the number of processed FFBs that can be prepared is 166,725 tons of FFB / year. A decrease in production caused by a 5% stagnation hour in the POM indicates that in the production process, there is a 5% decrease in POM productivity so that the total POM productivity can be achieved by 95%. This condition is, of course, still tolerated considering that it is still within the limits of reasonableness, which ideally POM productivity ranges from 95-98%. The productivity problem is dominated by low CPO production [6].

Based on the explanation above, we offer an alternative solution in the form of CPO production predictions that will be achieved based on input data of tons of FFB predictions to be harvested at each age of the plant and the extent of the garden blocks using a data mining algorithm. This research can assist Palm Oil Mill (POM) in developing strategies to increase palm oil production capacity and assist management in decision making if CPO production is not on target. This study compares three data mining algorithms to predict the amount of CPO production based on the number of FFBs. Based on the literature study that has been done, in this study, the algorithm is compared to is the multilayer perceptron (MLP), support vector regression (SVR), and Linear Regression (LG). Some researchers have compared several prediction methods, but applied to different cases or data [7], [8].

Several studies that utilize data mining algorithms to predict the amount of production of goods in a company have been carried out. Adhikari et al. explained in detail and in full the application of machine learning in-demand predictions [9]. Next, Rozikin and Solichin predicted food supplies at fast-food restaurants [10]. İşlek and Öğüdücü research to anticipate the needs of various products from a primary distribution warehouse [11]. The study addresses the problem of forecasting different product demands of central

978-1-7281-9167-6/20/$31.00 ©2020 IEEE

distribution warehouses. Demand forecasting is the activity of building forecasting models to estimate the quantity of a product that customers will purchase. The proposed methodology clusters similar warehouses according to their sale behavior using bipartite graph clustering. After that, the hybrid forecasting phase, which combines the moving average model and Bayesian Network machine learning algorithm is applied.

Other studies by Fradinata et al. propose a Support Vector Regression (SVR) method to predict demand for instant noodles in a company [12]. The SVR method provides positive results in forecasting demand for instant noodles. Meanwhile, Rohana and Arifuddin, in their research, examined the classification of tax compliance [13]. They find the problem will be challenging to find and present taxpayer data on large amounts of data, so it is challenging to distinguish taxpayers who obey and disobey to pay taxes. The proposed method is an artificial neural network to classify the taxpayer. The study also compared two artificial neural network algorithms, namely the Multilayer Perceptron and Support Vector Regression. Comparisons are made based on the level of accuracy that can be provided by the two algorithms.

Many recent studies use time-series data to predict things. The prediction of time series can be realized through the mining of time series data so that we can obtain the development process and regularity of social-economic phenomena reflected by time series, and extrapolate to predict its development trend. And in the era of big data, time series became one of the essential data sources that built big data itself. The study of Wang et al. introduce various time series autoregressive (AR) model, moving average (MA) model, and ARIMA model for predicting the risk of the National SME Stock Trading [14]. The case studies show that the results of the analysis are consistent with the actual situation, which has dramatically helped the prediction of financial risks.

Several previous studies have made comparisons of prediction methods for time series data. Choubin et al. compared multi-linear regression, multi-layer perceptron and ANFIS methods to predict climate processes [15]. As a result, the MLP method is superior to the other two methods. In line with this, an old study [16] comparing multi-layer perceptron and linear regression methods to predict epidemiological data also stated that the MLP method was superior. Recent studies that attempt to predict wind speed have concluded that the application of the neural network method is preferable to the linear regression method [17], , in line with other research predicting lake surface temperatures [18].

In this study, we also use time-series data to predict the amount of CPO production. To get the best prediction algorithm, compared to three data mining algorithms, namely multilayer perceptron (MLP), support vector regression (SVR), and linear regression (LR). The comparative measures used are Predicted Trend Accuracy (PTA), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and mean absolute deviation (MAD). After the best algorithm is obtained to make predictions, then it is implemented on a prototype application to predict the amount of CPO production. The results of this study are beneficial for POM to optimize and increase CPO production.

II. Research Method

In this study, we adopt the CRISP-DM data mining stage [19]. The CRISP-DM stage consists of five steps, namely business understanding, data understanding, data preparation / preprocessing, modeling, evaluation, and deployment. We also add the evaluation stage to evaluate the prediction system performance. The research method is illustrated in Figure 1.

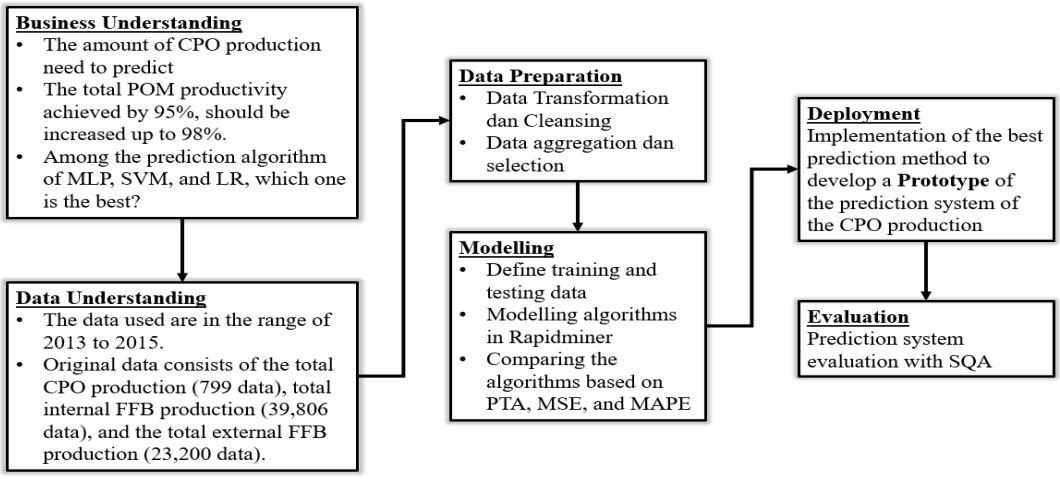

Fig. 1. The proposed method stages

id_material	mat_desc	PstngDate	DocDate	qty	sat
AG00100010	Crude Palm Oil (CPO)	03-01-2013	03-01-2013	6,95	MT
AG00100010	Crude Palm Oil (CPO)	04-01-2013	04-01-2013	19,58	MT
AG00100010	Crude Palm Oil (CPO)	05-01-2013	05-01-2013	60,75	MT
AG00100010	Crude Palm Oil (CPO)	07-01-2013	07-01-2013	42,96	MT

(a)

id_mat	mat_desc	Plnt	SLoc	MvT	MatDoc	Item	PstngDate	DocDate	Qty	EUn	EntryDate	Time	Customer	AmountLC	Crcy	User	Reference	HeaderText
WI10100010	TBS Internal	50J1	FB01	Z31	4900053833	1	31-12-2013	31-12-2013	6,16	MT	31.12.2013	10:58:33		0	IDR	SUMUZ	II	50J1R020018
WI10100010	TBS Internal	50J1	FB01	Z31	4900053832	1	31-12-2013	31-12-2013	2,08	MT	31.12.2013	10:58:33		0	IDR	SUMUZ	III	50J1R020011
WI10100010	TBS Internal	50J1	FB01	Z31	4900053834	1	31-12-2013	31-12-2013	4,72	MT	31.12.2013	10:58:33		0	IDR	SUMUZ	II	50J1R020020
WI10100010	TBS Internal	50J1	FB01	Z31	4900053835	1	31-12-2013	31-12-2013	2,46	MT	31.12.2013	10:58:33		0	IDR	SUMUZ	III	50J1R020021
WI10100010	TBS Internal	50J1	FB01	Z31	4900053836	1	31-12-2013	31-12-2013	7,26	MT	31.12.2013	10:58:34		0	IDR	SUMUZ	II	50J1R020022

(b)

id_mat	mat_desc	Plnt	SLoc	MvT	MatDoc	Item	PstngDate	DocDate	qty	EUn	Entry Date	Time	AmountLC	Crcy	User	Reference	HeaderText	Vendor
WI10100030	TBS Eksternal (NV)	50J1	FB01	Z21	4900053912	1	31-12-2013	31-12-2013	4,52	MT	31.12.2013	15:39:14	0	IDR	SUMUZ	PT. KHARISMA ALA	50J1R020080	3000588
WI10100030	TBS Eksternal (NV)	50J1	FB01	Z21	4900053911	1	31-12-2013	31-12-2013	5,89	MT	31.12.2013	15:28:33	0	IDR	SUMUZ	PT. KHARISMA INT	50J1R020081	3000590
WI10100030	TBS Eksternal (NV)	50J1	FB01	Z21	4900053910	1	31-12-2013	31-12-2013	6,71	MT	31.12.2013	15:28:32	0	IDR	SUMUZ	PT. KHARISMA INT	50J1R020082	3000590
WI10100030	TBS Eksternal (NV)	50J1	FB01	Z21	4900053909	1	31-12-2013	31-12-2013	5,48	MT	31.12.2013	15:12:12	0	IDR	SUMUZ	KOPERASI SERBA U	50J1R020076	3000642
WI10100030	TBS Eksternal (NV)	50J1	FB01	Z21	4900053907	1	31-12-2013	31-12-2013	5,67	MT	31.12.2013	15:00:59	0	IDR	SUMUZ	PT. KHARISMA INT	50J1R020075	3000590

(c)

Fig. 2. Example of original data obtained from the company: (a) total CPO production, (b) total internal FFB production, and (c) total external FFB production

In the first stage, we try to understand the problems and business needs in the research object. Based on the analysis results, a prediction system for CPO production is needed to increase company productivity. Also, based on literature studies, it was obtained three best prediction methods, namely multilayer perceptron, support vector regression, and linear regression. We compared the three methods to obtain the method that best matches the CPO production data.

TABLE I. THE ORIGINAL DATA SET OF THIS STUDY

Table name	Number of data	Description
tb_cpo	799	Total CPO production
tb_tbs_int	39806	Total internal FFB production
tb_tbs_ext	23200	Total external FFB production

The second stage involves understanding the data. Table 1 presents the original data obtained from the research object. The data used are in the range of 2013 to 2015. The original data consists of the total CPO production (799 data), total internal FFB production (39,806 data), and the total external FFB production (23,200 data). Meanwhile, in Figure 2 the data structure of the three types of data is presented.

The data preparation stages are divided into two parts. First, data transformation and cleansing. CPO production data for 2013-2015 as presented in Table 1 is entered into the database by applying data transformation, including changing the '*PstngDate*' field type from string to Date, and rounding the '*quantity*' field. Second, we do data aggregation and selection. Data aggregation is carried out on CPO production data, internal FFB production data, and external FFB production data. In the three tables, the data on the amount of production is aggregated for each day. The final result of the data aggregation and selection process is in the form of a table consisting of 734 data.

Data selection is done by using correlation analysis in which the attributes obtained are correlated with the research target, namely the amount of CPO production, so that the attributes obtained are good for further use. Figure 3 shows a graph of Internal TBS production data results of the aggregation and cleansing that has been done. The number of Internal FFB production has a correlation of 0.428550996 with CPO production. Next, Figure 4 presents data on the output of External FFB that has been cleansed. The amount of External FFB production has a correlation of 0.655901613 with CPO production. Finally, Figure 5 presents a graph of the processed CPO production plot. This data is the prediction target of this study. Based on the chart, it appears that CPO production tends to be unstable. The lowest CPO production is 6.95 MT, and the highest CPO production is 223 MT.

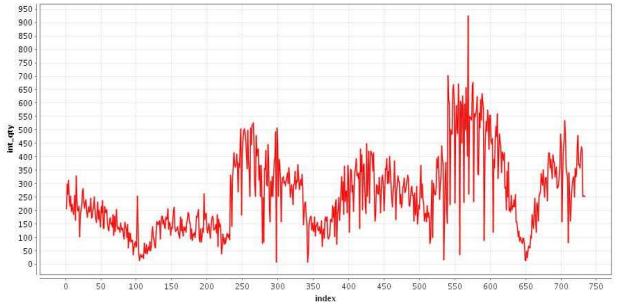

Fig. 3. Total internal FFB production

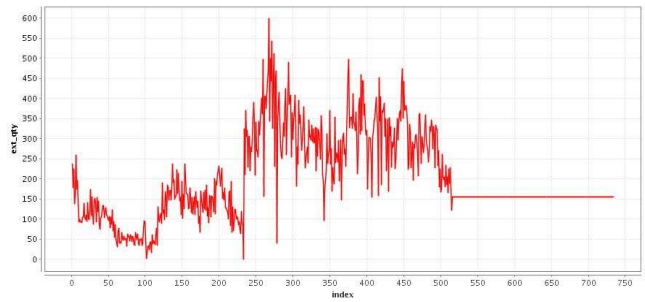

Fig. 4. Total external FFB production

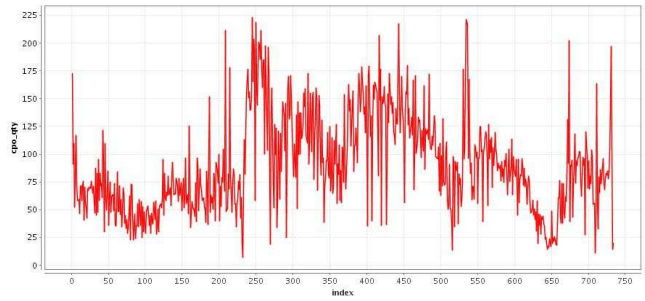

Fig. 5. Total CPO production

The modeling stage is divided into three sections, namely defining the training and testing data, modelling algorithms with Rapidminer, and comparing the algorithms based on PTA, MSE, and MAPE measurements. We describe the results of algorithm comparisons in the next chapter. Then, we evaluate them to conclude the best prediction algorithm. Based on these conclusions, in the development stage, a prototype prediction system for CPO production was developed using the best data mining algorithm. Finally, we evaluate the prototype system by the software quality assurance (SQA) method.

III. Result and Discussions

A. Modelling and Comparison Results

In this study, three algorithms were compared for predicting time series, namely Multilayer Perceptron, Support Vector Regression, and Linear Regression. The three algorithms are tested to predict CPO production one day in the future by studying patterns in one week. The measurements of the comparison algorithm used are predicted trend accuracy (PTA), mean squared error (MSE), mean absolute percentage error (MAPE), and mean absolute deviation (MAD). The prediction of time-series data mining by using the three algorithms is carried out with several variations of a combination of the training set and testing set, as presented in Table 2.

Based on the tests that have been done using the scenarios in Table 2, the results obtained by measuring the values of PTA, MSE, MAPE, and MAD from each algorithm tested. Table 3 presents the results of testing the Multilayer Perceptron algorithm, Table 4 shows the results of the Support Vector Regression algorithm, and Table 5 displays the results of testing the Linear Regression algorithm.

TABLE II. Training and Testing Dataset

Experiment	Number of data training	Number of data testing
#1	551 (75%)	184 (25%)
#2	587 (80%)	147 (20%)
#3	624 (85%)	110 (15%)
#4	661 (90%)	73 (10%)

TABLE III. The Experiment Result of Multilayer Perceptron Algorithm

Experiment	Multilayer Perceptron (MLP)			
	PTA	***MSE***	***MAPE***	***MAD***
#1	0.569	3063.464	88.993	39.010
#2	0.569	2568.983	113.211	41.582
#3	0.565	1404.981	77.325	29.192
#4	0.569	1803.228	64.720	31.645

TABLE IV. The Experiment Result of the Support Vector Regression Algorithm

Experiment	Support Vector Regression (SVR)			
	PTA	***MSE***	***MAPE***	***MAD***
#1	0.608	1223.581	65.918	28.688
#2	0.611	1152.587	68.065	26.709
#3	0.611	1170.782	73.660	25.780
#4	0.694	1146.054	47.485	22.333

TABLE V. The Experiment Result of Linear Regression Algorithm

Experiment	Linear Regression (LR)			
	PTA	***MSE***	***MAPE***	***MAD***
#1	0.575	1170.327	62.076	28.270
#2	0.632	1121.654	61.105	26.709
#3	0.593	1161.497	66.868	27.001
#4	0.653	1212.974	48.126	22.875

Based on the test results presented in Table 3, Table 4, and Table 5, it can be concluded that the Support Vector Regression (SVR) algorithm obtained a higher PTA value than the other two algorithms. Meanwhile, seen from the MSE, MAPE, and MAD values, the SVR algorithm gets lower costs than the Multilayer Perceptron algorithm and Linear Regression. Thus, it can be concluded that the Support Vector Regression algorithm has higher accuracy than the Multilayer

Perceptron and Linear Regression algorithms in predicting time series for palm oil CPO production.

B. *Development of the Prediction System*

Based on the evaluation that has been done, it is concluded that the Support Vector Regression Algorithm has a better level of accuracy compared to the Multilayer Perceptron and Liner Regression algorithms in predicting time series data mining for CPO production. Support Vector Regression can provide MSE of 1146.054, MAPE of 47.485, MAD of 22.333, and PTA of 0.694. Furthermore, the SVR algorithm is applied in the development of a prototype prediction system for CPO production. Figure 6 is a prototype main page display that was built. To run the system, users are asked to specify a dataset table, where to store predictions, and several options. The prediction results will be saved in a Microsoft Excel file.

To test the quality of the prototypes produced, we use the Software Quality Assurance method. In SQA, there are four factors tested, namely functionality, portability, usability, and efficiency. The four factors are weighted with the composition of functionality = 30%, portability = 20%, usability = 20%, and efficiency = 30%. Software Quality Assurance is carried out by distributing questionnaires to the head of the plantation division, plantation administrators, data analysts and business analysts at the oil palm company.

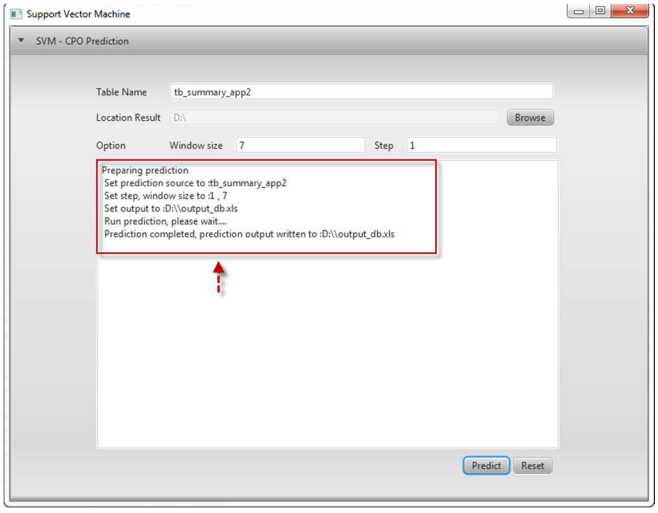

Fig. 6. Prototype Prediction System

The SQA instrument processing results showed that the quality value of the prototype produced was 80.225 which was in a good category. Thus, although it is still elementary, prototypes can be used to predict CPO production in oil palm companies.

IV. Conclusion

Based on the results of the study and a series of tests that have been carried out in solving the prediction problem of palm oil CPO production, it can be concluded that the time-series data mining approach can be used to predict CPO production in oil palm companies with a PTA value of 0.694. In contrast to previous studies [15]–[17], in this study it turned out that the Support Vector Regression algorithm provides higher accuracy compared to Multilayer Perceptron and Linear Regression algorithms in predicting CPO production. This study also developed a prototype of a CPO production prediction system using the SVR algorithm. Based on evaluation of the prototype using the Software Quality Assurance standards, the resulting quality value is 80.225, which means it is in the good category.

References

[1] I. F. Timorria, "Produksi Minyak Sawit Indonesia Tumbuh 14 Persen," *bisnis.com*, 2019. [Daring]. Tersedia pada: https://ekonomi.bisnis.com/read/20191017/99/1160433/produksi-minyak-sawit-indonesia-tumbuh-14-persen. [Diakses: 02-Mar-2020].

[2] Bambang, "Statistik Perkebunan Indonesia 2015 -2017 Kelapa Sawit," *Sawit*, hal. 81, 2017.

[3] A. Rifin, "Efisiensi Perusahaan Crude Palm Oil (CPO) di Indonesia," *J. Manaj. dan Agribisnis*, vol. 14, no. 2, hal. 103–108, 2017.

[4] W. W. Pamungkas, M. S. Maarif, T. T. Irawadi, dan Y. Arkeman, "Pemodelan Statistical Control Detection Adaptive (SCDA) Untuk Monitoring Dan Prediksi Volume Produksi Crude Palm Oil (CPO) Nasional," *J. Teknol. Ind. Pertan.*, vol. 27, no. 1, hal. 1–8, 2018.

[5] J. H. V Purba dan T. Sipayung, "Perkebunan Kelapa Sawit Indonesia dalam Perspektif Pembangunan Berkelanjutan," *Masy. Indones.*, vol. 43, hal. 81–94, 2017.

[6] S. Djohar, H. Tanjung, dan E. R. Cahyadi, "Membangun Keunggulan Kompetitif CPO Melalui Supply Chain Management: Studi Kasus di PT. Eka Dura Indonesia, Astro Agro Lestari, Riau," *J. Manaj. Agribisnis*, vol. 1, hal. 20–32, 2014.

[7] Saigal S dan Mehrotra D, "Performance Comparison Of Time Series Data Using Predictive Data Mining Techniques Advances in Information Mining," vol. 4, no. 1, hal. 57–66, 2012.

[8] J. K. Mantri, "Comparison between SVM and MLP in Predicting Stock Index Trends," no. 9, hal. 81–82, 2013.

[9] N. C. Das Adhikari *et al.*, "An Intelligent Approach to Demand Forecasting," in *Proceedings of the 2nd International Conference on Inventive Computation Technologies (ICICT 2017)*, 2017, hal. 1105–1111.

[10] C. Rozikin dan A. Solichin, "Implementasi Algoritma Genetika dan Regresi Linier Berganda Untuk Prediksi Persediaan Bahan Makanan Pada Restoran Cepat Saji Implementation Of Genetic Algorithm and Multi-Linear Regression For Predicting Food Supplies At Fast Food Restaurants," in *Seminar Nasional Multidisiplin Ilmu (SENMI) 2017*, 2017, no. April, hal. 10–17.

[11] İ. İşlek dan Ş. G. Öğüdücü, "A retail demand forecasting model based on data mining techniques," in *2015 IEEE 24th International Symposium on Industrial Electronics (ISIE)*, 2015, hal. 55–60.

[12] E. Fradinata, Z. M. Kesuma, S. Rusdiana, dan N. Zaman, "Forecast Analysis of Instant Noodle Demand using Support Vector Regression (SVR)," in *1st South Aceh International Conference on Engineering and Technology*, 2019, hal. 1–9.

[13] T. Rohana dan M. Arifuddin, "Kajian Algoritma Jaringan Syaraf Tiruan untuk Mendeteksi Secara Dini Kepatuhan Wajib Pajak Orang Pribadi," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2013, hal. 1–6.

[14] F. Wang, M. Li, Y. Mei, dan W. Li, "Time Series Data Mining: A Case Study with Big Data Analytics Approach," *IEEE Access*, vol. 8, hal. 14322–14328, 2020.

[15] B. Choubin, S. Khalighi-Sigaroodi, A. Malekian, dan Ö. Kişi, "Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals," *Hydrol. Sci. J.*, vol. 61, no. 6, hal. 1001–1009, 2016.

[16] J. Gaudart, B. Giusiano, dan L. Huiart, "Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data," *Comput. Stat. Data Anal.*, vol. 44, no. January, hal. 547–570, 2004.

[17] S. Barhmi, O. Elfatni, dan I. Belhaj, "Forecasting of wind speed using multiple linear regression and artificial neural networks," *Energy*

Syst., vol. 11, no. 4, hal. 935–946, 2020.

[18] S. Zhu, M. Ptak, Z. M. Yaseen, J. Dai, dan B. Sivakumar, "Forecasting surface water temperature in lakes: A comparison of approaches," *J. Hydrol.*, vol. 585, hal. 124809, 2020.

[19] J. Han dan M. Kamber, "Data Mining : Concepts and Techniques (2nd edition)," *SIGKDD Explor.*, hal. 1–7, 2006.