Journal of Advanced Data Science and Machine Learning
Vol. X, No. X, March 2024, pp. 1-10
ISSN xxxx-xxxxx

Forecasting Tourist Arrivals Using Seasonal Autoregressive Integrated Moving Average Optimization Based on Genetic Algorithm in Kuningan Regency

Luthfi Arie Zulfikri [a,1] , Nurul Hidayat [a,2, *]

[a] Department of Informatics, Engineering Faculty, Universitas Jenderal Soedirman, Purwokerto, 53122, Indonesia
[1] luthfiaz49@gmail.com; [2] nurul@unsoed.ac.id*
* corresponding author

(Received ..., ... Revised ..., ... Accepted ..., ..., Available online ..., ...)

Abstract

Tourism plays a strategic role in supporting global economic growth, particularly in developing countries. This sector significantly contributes to the increase of Gross Domestic Product (GDP) and the reduction of poverty rates. However, predicting the number of tourist arrivals remains a challenge due to seasonal patterns. The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is commonly used to address this issue, yet selecting optimal parameters within the SARIMA model remains complex. This study aims to optimize parameter selection and improve the accuracy of tourist arrival forecasts in Kuningan Regency through a hybrid approach that integrates SARIMA with a Genetic Algorithm (GA). The Genetic Algorithm is employed to automate and optimize the parameter selection process in the SARIMA model. Experiments were conducted using various combinations of population sizes (50 and 100) and generations (10, 20, and 50) to determine the best configuration. The results indicate that the integration of GA and SARIMA effectively overcomes the limitations of conventional SARIMA in parameter optimization and recognizing complex data patterns. Increasing the number of generations tends to enhance model accuracy, albeit at the cost of increased computational time. The best model was obtained with the SARIMA (9, 0, 5) × (0, 1, 2, 12) configuration, yielding a Mean Absolute Error (MAE) of 15,507.07. These findings demonstrate that the GA-SARIMA hybrid approach has strong potential to enhance seasonal data forecasting performance, particularly in the tourism sector.

Keywords: *Forecasting, Genetic Algorithm, Parameter Optimization, Seasonal Autoregressive Integrated Moving Average (SARIMA), Time Series, Tourism*

1. Introduction

Tourism has long been recognized as a key driver of economic growth and development [1]. Globally, tourism is one of the primary sectors that generate substantial employment opportunities, contribute significantly to national income, and promote economic prosperity [2]. A 1% increase in the tourism sector in developing countries can significantly boost Gross Domestic Product (GDP) by 0.051%, Foreign Direct Investment (FDI) by 2.647%, energy development by 0.134%, and agricultural development by 0.26%, while also reducing poverty by 0.51% in the long term [3].

Forecasting is a crucial initial step in investment decision-making and planning [4]. Accurate forecasting is essential for tourism destinations, where decision-makers and business managers strive to optimize sectoral growth while maintaining local environmental sustainability and economic performance [5].

One of the most widely used and effective forecasting algorithms is the Seasonal and Non-Seasonal Autoregressive Integrated Moving Average (SARIMA and ARIMA) models [6], [7], [8]. The ARIMA model, based on the Box-Jenkins approach, can be used to predict future trends by transforming data into a stationary series and eliminating seasonal patterns [9]. The SARIMA model is particularly useful when the data exhibit seasonal periodic fluctuations

that frequently recur within a year [10]. Therefore, SARIMA is a more suitable model for forecasting tourist arrivals due to its ability to capture seasonal trends.

However, SARIMA has certain limitations in optimizing results, particularly in selecting the best parameter values for p, d, q as well as P, D, Q [10]. Here, p represents the order of Auto-Regression (AR), q is the order of the Moving Average (MA), and d is the differencing order required to make the time series stationary. Similarly, P, D, Q represent the seasonal AR component, seasonal moving average component, and seasonal differencing component, respectively [11]. Finding the optimal combination of these parameter values requires repetitive trial-and-error processes to achieve the best forecasting performance.

The selection of the best parameter combination for SARIMA and ARIMA models can be enhanced using Genetic Algorithms (GA), which help determine the optimal model configuration [10], [12], [13], [14], [15]. The implementation of a Genetic Algorithm facilitates the automatic selection of optimal parameter values and significantly reduces forecasting time. Based on these considerations, this study employs the SARIMA model optimized with Genetic Algorithms as a solution to forecast tourist arrivals in Kuningan Regency.

2. Method

This study employs a hybrid approach combining the Seasonal Autoregressive Integrated Moving Average (SARIMA) model and Genetic Algorithm (GA) for forecasting the number of tourists in Kuningan Regency. The overall research stages are illustrated in Figure 1.

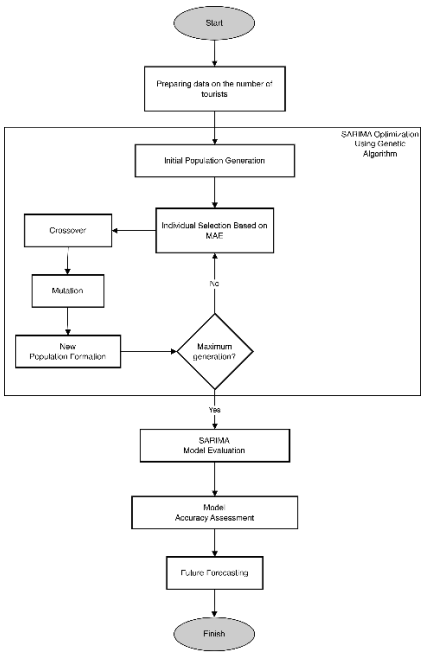

Fig. 1. Research Diagram

2.1. Data Preparation

The data used in this study consists of the number of tourist arrivals in Kuningan Regency from January 2016 to December 2023. The dataset is collected and structured for analysis and forecasting using the SARIMA model.

2.2. SARIMA Optimization Using Genetic Algorithm

The SARIMA model optimization is conducted by utilizing the Genetic Algorithm to determine the optimal model parameters. This optimization process involves several key stages of the Genetic Algorithm, including initial population generation, selection, crossover, mutation, and new population formation.

2.2.1. Initial Population Generation

The initial population is generated by randomly selecting parameter values for the SARIMA model, including (p, d, q, P, D, Q), using a random generator technique. Each individual in the population represents a unique SARIMA model.

2.2.2. Individual Selection

The selection process is performed using the Tournament Selection method. In this method, a subset of individuals is randomly chosen from the population, paired, and compared based on their fitness values [16]. The individual with the lowest Mean Absolute Error (MAE) from each comparison is selected to proceed to the next stage.

2.2.3. Crossover

The crossover stage is carried out using the Two-point Crossover method. Two points on an individual's chromosome are randomly selected, and the genes between these points are exchanged between two individuals to generate new offspring [16].

2.2.4. Mutation

Mutation is performed using the uniform integer mutation method. In this stage, one or more genes in an individual's chromosome are replaced with new values randomly selected from the predefined parameter range [17].

2.2.5. New Population Formation

The crossover and mutation processes generate a new population for the next iteration. This process is repeated until the maximum number of generations is reached. The individual with the lowest MAE is considered the optimal solution for the SARIMA model.

2.3. SARIMA Model Evaluation

The SARIMA model obtained through the optimization process is evaluated by testing the significance of its parameters. This test ensures that the selected parameters contribute significantly to the model's performance. Additionally, a diagnostic test is conducted using the Ljung-Box autocorrelation test on the model's residuals to confirm that the residuals exhibit no significant autocorrelation and follow a white noise pattern.

2.4. Model Accuracy Assessment

The optimized SARIMA model's accuracy is assessed using several evaluation metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics provide insights into the model's forecasting accuracy.

2.5. Future Forecasting

Once the optimal SARIMA model is established and validated, it is used to forecast future tourist arrivals. The forecasting results provide insights into future trends in tourist numbers in Kuningan Regency. This information can serve as a basis for strategic planning and decision-making in the tourism sector.

3. Results and Discussion

3.1. Tourist Arrival Data in Kuningan Regency

The dataset used in this study consists of monthly tourist arrival data in Kuningan Regency from 2016 to 2023. This data was obtained from Open Data Kuningan and produced by the Tourism Office of Kuningan Regency. The dataset comprises 96 observations, with data from 2016 to 2022 used as the training set, while data from 2023 was used to test the model's performance. Table 1 presents a summary of the number of tourists (in thousands) per month over the research period.

Table 1. Total Number of Tourists in Kuningan Regency (thousands)

Year	Jan	Feb	Mar	Apr	Mei	Jun	Jul	Agu	Sep	Okt	Nov	Des
2016	282	83	89	90	115	31	347	98	121	136	144	204
2017	166	98	111	137	135	295	184	126	120	124	219	309
2018	162	90	99	110	114	382	157	146	151	147	127	462
2019	193	113	126	140	57	462	186	160	176	180	177	360
2020	209	116	53	-	-	24	122	180	129	142	147	149
2021	163	87	121	71	259	73	-	75	131	150	115	119
2022	162	99	103	28	302	133	176	119	119	130	98	160
2023	167	99	83	87	128	144	122	110	137	123	114	240

3.2. SARIMA Optimization Using Genetic Algorithm

This study implements a combination of the SARIMA model and Genetic Algorithm (GA) to predict the number of tourists in Kuningan Regency. The optimization process involves evaluating various combinations of population size and the number of generations to determine the best configuration based on Mean Absolute Error (MAE) and computational time. The experiment was conducted on a system with the following specifications:

- Operating System: Windows 11 64-bit
- Processor: Intel Core i5-9300H @ 2.40 GHz
- Memory: 16 GB DDR4
- Storage: 512 GB SSD

3.2.1. Model Implementation and Evaluation

Different combinations of population size (50 and 100) and the number of generations (10, 20, and 50) were tested to optimize the SARIMA parameters. The experimental results indicate that increasing the population size and the number of generations generally leads to longer computational time but also improves prediction accuracy. The computational time for each configuration is visualized in Figure 2.

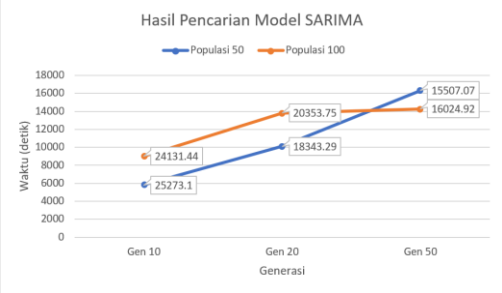

Fig. 2. SARIMA Model Search Results

3.2.2. Hasil Optimasi

Table 2 summarizes the optimization results. The best configuration was obtained with a population size of 50 and 50 generations, yielding a SARIMA model with parameters (9, 0, 5) × (0, 1, 2, 12). This model achieved an MAE of 15,507.07 with a computation time of 16,309.41 seconds.

Table 2. SARIMA Model Optimization Results

Populasi	Gen	Model	MAE	Waktu (detik)
50	10	(8, 0, 3) × (5, 1, 0, 12)	25273.1	5796.56
50	20	(7, 0, 0) × (0, 1, 2, 12)	18343.29	10101.93
50	50	(9, 0, 5) × (0, 1, 2, 12)	15507.07	16309.41
100	10	(1, 0, 1) × (1, 1, 1, 12)	24131.44	9004.17
100	20	(0, 0, 6) × (2, 1, 1, 12)	20353.75	13822.92
100	50	(6, 0, 5) × (0, 1, 2, 12)	16024.92	14275.87

3.3. SARIMA Model Evaluation

The process of developing an optimal Seasonal Autoregressive Integrated Moving Average (SARIMA) model considers the characteristics of the data and several evaluation criteria, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). In this study, MAE was used as the primary indicator for selecting the best SARIMA model to forecast the number of tourists in Kuningan Regency.

Among the various models generated, SARIMA (9, 0, 5) × (0, 1, 2, 12) was selected as the model with the lowest MAE. This model was then tested for parameter significance, as presented in Table 3. The analysis results indicate that only the ma.S.L12 parameter is significant (p-value = 0.038), while the other parameters are not significant (p-value >

0.05). Based on these findings, an alternative model, SARIMA (1, 0, 1) × (1, 1, 1, 12), was explored, with its parameter significance test results shown in Table 4. The alternative model demonstrates that all parameters are statistically significant (p-value < 0.05).

Table 3. Parameter Significance Test Results for SARIMA (9, 0, 5) × (0, 1, 2, 12)

Term	Coefficient	Std Err	z	P>|z|	[0.025	0.975]
ar.L1	0.2122	1.424	0.149	0.882	-2.578	3.002
ar.L2	0.5887	0.885	0.665	0.506	-1.147	2.324
ar.L3	-0.6853	1.078	-0.636	0.525	-2.799	1.428
ar.L4	-0.2445	0.879	-0.278	0.781	-1.966	1.477
ar.L5	-0.1138	0.682	-0.167	0.867	-1.450	1.222
ar.L6	0.4928	0.261	1.888	0.059	-0.019	1.004
ar.L7	0.0169	0.572	0.029	0.976	-1.104	1.137
ar.L8	-0.2573	0.347	-0.742	0.458	-0.937	0.423
ar.L9	0.2624	0.331	0.793	0.428	-0.386	0.911
ma.L1	-0.1732	1.551	-0.112	0.911	-3.214	2.867
ma.L2	-0.4708	1.100	-0.428	0.669	-2.628	1.686
ma.L3	0.7430	1.045	0.711	0.477	-1.305	2.791
ma.L4	0.2665	0.860	0.310	0.757	-1.419	1.952
ma.L5	0.0935	0.680	0.137	0.891	-1.240	1.427
ma.S.L12	-0.5771	0.278	-2.073	0.038	-1.123	-0.032
ma.S.L24	-0.2803	0.298	-0.940	0.347	-0.865	0.304
sigma2	5.141e+09	6.67e-10	7.7e+18	0.000	5.14e+09	5.14e+09

Table 4. Parameter Significance Test Results for SARIMA (1, 0, 1) × (1, 1, 1, 12)

Term	Coefficient	Std Err	z	P>|z|	[0.025	0.975]
ar.L1	0.9796	0.051	19.352	0.000	0.880	1.079
ma.L1	-0.9199	0.078	-11.798	0.000	-1.073	-0.767
ar.S.L12	0.4569	0.250	1.830	0.047	-0.032	0.946
ma.S.L12	-0.8950	0.253	-3.533	0.000	-1.391	-0.398
sigma2	7.885e+09	6.2e-11	1.27e+20	0.000	7.89e+09	7.89e+09

To test the residual assumptions, the Ljung-Box test was conducted, with results presented in Table 5. The analysis shows that the SARIMA (9, 0, 5) × (0, 1, 2, 12) model has a p-value of 0.960, indicating no significant autocorrelation in the residuals, thus satisfying the white noise assumption. In contrast, the SARIMA (1, 0, 1) × (1, 1, 1, 12) model has a p-value of 0.048, indicating significant autocorrelation in the residuals.

Table 5. Ljung-Box Test Results

Model	Ljung-Box Statistic	p-value
SARIMA (9, 0, 5) × (0, 1, 2, 12)	4.925579	0.960419
SARIMA (1, 0, 1) × (1, 1, 1, 12)	21.159175	0.048098

Based on diagnostic results and evaluation criteria, the SARIMA (9, 0, 5) × (0, 1, 2, 12) model was selected as the best model. Although some of its parameters are not significant, this model meets diagnostic criteria and exhibits superior prediction accuracy. Moreover, this model aligns with findings in the literature [10]. The optimization of SARIMA models depends not only on parameter significance but also on meeting diagnostic assumptions and providing accurate forecasts.

3.4. Model Accuracy Assessment

The tourist arrival data for Kuningan Regency was divided into two subsets: training data (January 2016–December 2022) and testing data (January 2023–December 2023). SARIMA (9, 0, 5) × (0, 1, 2, 12) was selected as the best model based on performance evaluation. To assess the forecasting performance, several accuracy metrics were used, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Table 6 presents the forecasting results for 2023 along with accuracy metrics.

Table 6. SARIMA (9, 0, 5) × (0, 1, 2, 12) Model Accuracy Results

Period	Actual Data	Prediction	MSE	RMSE	MAE	MAPE (%)
23-Jan	167042	160925.69	3.74E+07	6116.31	6116.31	3.66
23-Feb	99057	107729.79	7.52E+07	8672.79	8672.79	8.76
23-Mar	82831	89250.13	4.12E+07	6419.13	6419.13	7.75
23-Apr	87185	81745.9	2.96E+07	5439.1	5439.1	6.24
23-May	127568	170342.06	1.83E+09	42774.06	42774.06	33.53
23-Jun	144117	153235	8.31E+07	9118	9118	6.33
23-Jul	122304	137626.4	2.35E+08	15322.4	15322.4	12.53
23-Aug	110208	117409.68	5.19E+07	7201.68	7201.68	6.53
23-Sep	137015	95650.38	1.71E+09	41364.62	41364.62	30.19
23-Oct	123244	138375.15	2.29E+08	15131.15	15131.15	12.28
23-Nov	114286	112719.34	2.45E+06	1566.66	1566.66	1.37
23-Dec	240320	213361.1	7.27E+08	26958.9	26958.9	11.22
Final Result	-	-	4.21E+08	20518.35	15507.07	11.7

The evaluation results indicate that the model performs well, with an average RMSE of 20,518.35 and a MAPE of 11.7%. Based on general classification, a MAPE value in the range of 10%–20% is considered good [18]. This suggests that the model has a reasonably adequate level of accuracy for projecting monthly tourist arrivals in Kuningan Regency.

3.5. Future Forecasting

The SARIMA (9, 0, 5) × (0, 1, 2, 12) model was used to project the number of monthly tourists in Kuningan Regency for the year 2024. Utilizing historical data, this model generated predictions reflecting tourist trends during the period. Table 7 presents the forecasted tourist arrivals for each month of 2024.

Tabel 1. Future Forecasting Results

Period	Predicted Tourists
24-Jan	160,925.69
24-Feb	107,729.79
24-Mar	89,250.13
24-Apr	81,745.90
24-May	170,342.06
24-Jun	153,235.00
24-Jul	137,626.40
24-Aug	117,409.68
24-Sep	95,650.38
24-Oct	138,375.15
24-Nov	112,719.34
24-Dec	213,361.10

4. Conclusion

The results obtained indicate that the integration of the Genetic Algorithm (GA) with the SARIMA model successfully addresses the limitations of traditional SARIMA in parameter optimization and recognizing complex data patterns. Experiments with varying population sizes (50 and 100) and generations (10, 20, and 50) demonstrate that increasing the number of generations tends to enhance model accuracy, although it also increases computational time. The best model, with the lowest Mean Absolute Error (MAE), was obtained using the SARIMA (9, 0, 5) × (0, 1, 2, 12) configuration, achieving an MAE of 15,507.07. These findings suggest that the proposed approach holds significant potential for improving the forecasting performance of seasonal data, such as tourism data.

Further development could involve implementing a parallel genetic algorithm to reduce computational time, thereby enhancing the efficiency of optimal parameter search without sacrificing forecasting accuracy. Additionally, broader variations in parameters, population sizes, and generations could be explored to identify a more optimal configuration.

References

[1] A. León-Gómez, D. Ruiz-Palomo, M. A. Fernández-Gámez, and M. R. García-Revilla, “Sustainable Tourism Development and Economic Growth: Bibliometric Review and Analysis,” *Sustainability*, vol. 13, no. 4, p. 2270, Feb. 2021, doi: 10.3390/su13042270.

[2] R. H. Zadeh Bazargani and H. Kiliç, "Tourism competitiveness and tourism sector performance: Empirical insights from new data," *Journal of Hospitality and Tourism Management*, vol. 46, pp. 73–82, Mar. 2021, doi: 10.1016/j.jhtm.2020.11.011.

[3] A. Khan, S. Bibi, A. Lorenzo, J. Lyu, and Z. U. Babar, "Tourism and Development in Developing Economies: A Policy Implication Perspective," *Sustainability*, vol. 12, no. 4, p. 1618, Feb. 2020, doi: 10.3390/su12041618.

[4] G. Xie, Y. Qian, and S. Wang, "Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach," *Tour Manag*, vol. 82, p. 104208, Feb. 2021, doi: 10.1016/j.tourman.2020.104208.

[5] H. Song, R. T. R. Qiu, and J. Park, "A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting," *Ann Tour Res*, vol. 75, pp. 338–362, Mar. 2019, doi: 10.1016/j.annals.2018.12.001.

[6] G. Tovmasyan, "Forecasting the number of incoming tourists using Arima model: case study from Armenia," *Marketing and Management of Innovations*, vol. 5, no. 3, pp. 139–148, 2021, doi: 10.21272/mmi.2021.3-12.

[7] S. Velos, M. Go, G. Bate, and E. Joyohoy, "A Seasonal Autoregressive Integrated Moving Average (SARIMA) Model to Forecasting Tourist Arrival in the Philippines: A Case Study in Moalboal, Cebu (Philippines)," *Recoletos Multidisciplinary Research Journal*, vol. 8, no. 1, pp. 67–78, Jun. 2020, doi: 10.32871/rmrj2008.01.05.

[8] R. Devi, A. Agrawal, J. Dhar, and A. K. Misra, "Forecasting of Indian tourism industry using modeling approach," *MethodsX*, vol. 12, p. 102723, Jun. 2024, doi: 10.1016/j.mex.2024.102723.

[9] T. Dimri, S. Ahmad, and M. Sharif, "Time series analysis of climate variables using seasonal ARIMA approach," *Journal of Earth System Science*, vol. 129, no. 1, p. 149, Dec. 2020, doi: 10.1007/s12040-020-01408-x.

[10] M. Farsi *et al.*, "Parallel genetic algorithms for optimizing the SARIMA model for better forecasting of the NCDC weather data," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 1299–1316, Feb. 2021, doi: 10.1016/j.aej.2020.10.052.

[11] K. E. ArunKumar, D. V. Kalaga, Ch. M. Sai Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, "Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA)," *Appl Soft Comput*, vol. 103, p. 107161, May 2021, doi: 10.1016/j.asoc.2021.107161.

[12] M. Y. F. Zaelani, "Implementasi Model SARIMA dan Algoritma Genetika pada Prediksi Produksi Minyak Bumi," *Progresif: Jurnal Ilmiah Komputer*, vol. 16, no. 2, p. 01, Sep. 2020, doi: 10.35889/progresif.v16i2.504.

[13] A. Abbasi, K. Khalili, J. Behmanesh, and A. Shirzad, "Estimation of ARIMA model parameters for drought prediction using the genetic algorithm," *Arabian Journal of Geosciences*, vol. 14, no. 10, p. 841, May 2021, doi: 10.1007/s12517-021-07140-0.

[14] R. R. Sharma, M. Kumar, S. Maheshwari, and K. P. Ray, "EVDHM-ARIMA-Based Time Series Forecasting Model and Its Application for COVID-19 Cases," *IEEE Trans Instrum Meas*, vol. 70, pp. 1–10, 2021, doi: 10.1109/TIM.2020.3041833.

[15] M. A. Deif, A. A. A. Solyman, and R. E. Hammam, "ARIMA Model Estimation Based on Genetic Algorithm for COVID-19 Mortality Rates," *Int J Inf Technol Decis Mak*, vol. 20, no. 06, pp. 1775–1798, Nov. 2021, doi: 10.1142/S0219622021500528.

[16] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimed Tools Appl*, vol. 80, no. 5, pp. 8091–8126, Feb. 2021, doi: 10.1007/s11042-020-10139-6.

[17] J. Feng, Q. Wang, and N. Li, "An Intelligent System for Heart Disease Prediction using Adaptive Neuro-Fuzzy Inference Systems and Genetic Algorithm," *J Phys Conf Ser*, vol. 2010, no. 1, p. 012172, Sep. 2021, doi: 10.1088/1742-6596/2010/1/012172.

[18] R. Novita, I. Yani, and G. Ali, "Sistem Prediksi untuk Penentuan Jumlah Pemesanan Obat Menggunakan Regresi Linier," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, no. 1, pp. 62–70, May 2022, doi: 10.57152/malcom.v2i1.198.