Contemporary Mathematics and Applications
2024.

PARAMETER ESTIMATION USING GENETIC ALGORITHM ON MULTIPLE LINEAR REGRESSION MODEL FOR FORECASTING SALES RESULT

Muhammad Fath Rajihan Nafie

Mathematics Study Program, Department of Mathematics, Faculty of Science and Technology, Airlangga University, Indonesia

Corresponding author Email: muhammad.fath.rajihan-2022@fst.unair.ac.id

Abstract. Forecasting sales results is a critical aspect of business management, particularly for retail grocery stores. Accurate sales forecasts can support efficient stock management, inventory planning, and effective marketing strategies. This research employs a hybrid approach by integrating Multiple Linear Regression with Genetic Algorithm to enhance forecasting accuracy for time-series sales data. Regression is utilized to model relationships between dependent and independent variables, while genetic algorithm optimizes the regression parameters to minimize errors. The dataset consists of 202 days of daily sales data, sourced from Kaggle. The results of this research indicate that the multiple linear regression model, after being optimized using the genetic algorithm, is represented as $Y = 63.5838 + 0.15481X_1 - 0.0156X_2 + 0.11582\mathrm{X}_3 + 0.06003\mathrm{X}_4 + 0.04486X_5 + 0.5572X_6$, this optimized model achieved a MAPE of 0.2197 on the training data and 0.2665 on the validation data, indicating a reasonable level of predictive accuracy. These results highlight the potential of the genetic algorithm in enhancing the performance of regression models.

Keywords: *forecasting; multiple linear regression; genetic algorithm; time-series analysis.*

1 Introduction

The advancement of computing technology has driven humanity to collect and store large amounts of data. This phenomenon can happen in various aspects of life, including the business sector. The growth of industries in the era of globalization has intensified business competition among companies. Undoubtedly, winning this competition has become a primary goal in formulating sales and promotion strategies. However, one of the main challenges in increasing sales is the ability of companies or business actors to predict future sales results before determining an effective strategy to face or maximize future opportunities. To face these challenge, one analytical technique that can be utilized by business actors or companies is sales forecasting, which involves quantitative forecasting of variables related to sales transactions [1].

Sales forecasting (SF) is a critical aspect of business management as it plays a central role in resource allocation, marketing, and financial planning. Additionally, it influences decision-making processes for providing products and services to consumers [2]. On the other hand, errors in sales forecasting can negatively impact strategic decision-making, such as suboptimal product provision or promotional budget waste. So that, accurate techniques or methods are required for sales forecasting based on time series data.

A combined approach using statistical methods and optimization algorithms can be employed by business actors or companies as a tool to forecast or estimate business sales outcomes based on time series data. This approach involves multiple linear regression, which is used to predict the relationship between variables with cause-and-effect associations. This method allows for analyzing patterns in sales results. Subsequently, a genetic algorithm will be integrated to estimate and optimize the parameters of the multiple linear regression model, thereby achieving better forecasting model performance. This study makes a significant contribution to improving the accuracy of sales forecasts, supporting efficient stock management, inventory planning, and marketing strategies based on sales time series data by leveraging a combination of statistical methods and optimization algorithms.

2 Literature Review

2.1 Multiple Linear Regression

Multiple linear regression analysis is an extension of simple linear regression analysis. The primary difference between simple linear regression and multiple linear regression lies in the number of independent variables. In simple linear regression, only one independent variable is used to predict the dependent variable, whereas in multiple linear regression, more than one independent variable is used to predict the dependent variable [3]. The model used for multiple linear regression analysis is represented as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \qquad (1)$$

with:

Y : Prediction result (dependent variable)

β_0 : Intercept

β_n : Coefficient of independent variable

X_n : Predictors (independent variable)

ε : Error

2.2 Genetic Algorithm

The genetic algorithm (GA) is an optimization method in the field of computing based on the principles of genetic evolution and natural selection. As such, terms in the genetic algorithm adopt terminology from living organisms, including the following [4]:

1. **Chromosome:** Also referred to as an individual, a chromosome represents a single solution to a specific optimization problem. A chromosome may consist of several genes (variables).
2. **Parent:** Parents refer to individuals (solutions) selected to be mated or combined with another individual, resulting in a new solution called offspring.
3. **Parent Selection:** The process of selecting individuals from the population to serve as parents.
4. **Crossover:** A process of combining traits or values from two parent individuals. The result is one or two new individuals (offspring) that incorporate characteristics from both parents.
5. **Mutation:** A process of modifying traits or values within an individual to produce a new individual. Mutation serves to explore a broader solution space. A high mutation rate allows for greater exploration but may risk disrupting good solutions. Therefore, the mutation probability must be optimally regulated.
6. **Fitness:** Fitness represents the suitability or viability level of an individual in the population. Individuals with high fitness values have a greater chance of surviving to the next generation. However, in some cases, individuals with low fitness values may also survive due to specific selection mechanisms.

The genetic algorithm was developed by John Holland in 1975 and popularized by David Goldberg in 1989. One of the main advantages of the genetic algorithm compared to conventional methods is that it does not require the evaluation of gradient functions for the objective function [5]. A key feature of the genetic algorithm is its ability to search for multiple potential solutions to find an optimal solution for computational problems.

2.3 Mean Absolute Percentage Error

Mean Absolute Percentage Error (MAPE) is an evaluation metric used to measure the accuracy of forecasted values as a percentage relative to the actual values [6]. This metric is widely applied due to its interpretability, as it provides an error expressed in percentage terms, making it easier to evaluate the deviation of predictions from true values. The formula for MAPE is written as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i' - y_i}{y_i}\right| \times 100\% \quad (2)$$

with:

n : Number of observations

y_i' : Predicted value for the i^{th} data point

y_i : Actual value for the i^{th} data point

3 Research Methodology

This study employs a quantitative research approach. The objective is to forecast the sales result of a retail grocery store by utilizing a hybrid technique combining multiple linear regression as the primary calculation method for forecasting and a genetic algorithm to optimize the parameters within multiple linear regression. The research workflow is illustrated in Figure 1.

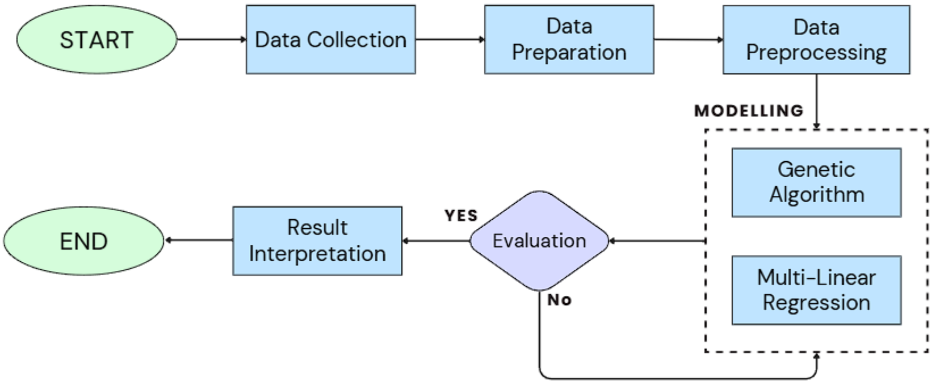

Figure 1 Research Workflow

Based on Figure 1, the research workflow begins with data collection, followed by data processing, which includes data preparation and data preprocessing. After processing the dataset, a hybrid model combining genetic algorithms with multiple linear regression is developed. During the training process, the results are evaluated using the evaluation metric, Mean Absolute Percentage Error. If the evaluation results are unsatisfactory, the dataset will be retraining until satisfactory results are achieved, ensuring the model can accurately forecast sales result.

3.1 Dataset

The type of data used in this research is secondary data, which is readily available in a specific repository without direct data collection. The dataset for this research was obtained from Kaggle, consisting of daily sales records from a retail grocery store over 202 days, spanning March 1, 2024, to September 18, 2024. This time-series data enables temporal analysis and the identification of trends and patterns specific to retail grocery sales. An overview of the dataset is shown in Table 1.

Table 1 Data Overview

No	Date	Day	Sales
1	2024-03-01	Friday	219791

2	2024-03-02	Saturday	381272
3	2024-03-03	Sunday	296996
…	….	….	….
202	2024-09-18	Wednesday	155899

3.2 Data preparation

After data collection, the next step is to ensure the collected data is ready for analysis or model development. The data preparation process in this study involves two primary steps:

1. Removing the column day, as the day information is not useful to the analysis in this research.
2. Transforming the data into a time-series format by grouping the data into seventh-day intervals as the target prediction to identify weekly patterns in sales data.

An overview of the dataset after data preparation is shown in Table 2.

Table 2 Sales data overview

No	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Target
1	219791	381272	296996	306301	408853	290948	293711
2	381272	296996	306301	408853	290948	293711	423443
3	296996	306301	408853	290948	293711	423443	464966
…	….	….	….	….	….	….	….
196	282937	165445	237814	308632	255651	284303	155899

Based on Table 2, the dataset underwent a reduction in size after being transformed into a time-series format with seventh-day intervals, resulting in a total of 196 records.

3.3 Data Preprocessing

Data preprocessing must be conducted to ensure the data format meets the requirements of the algorithm used in this study, namely multiple linear regression. The data preprocessing process in this research involves splitting the data into two parts: training data, used to train the multiple linear regression model, and validation data, used to evaluate the multiple linear regression model. The training data consists of the first 186 records, while the validation data consists of the last 10 records.

3.4 Modelling

Once the dataset has been pre-processed and is ready for use, the next step is modelling. This process aims to identify patterns or relationships within the data that can be used for prediction or classification. The steps involved in this stage include algorithm selection, model training, performance evaluation, and model refinement based on evaluation results.

In this study, a hybrid model combining multiple linear regression with a genetic algorithm is applied to train the processed data. The integration of multiple linear regression and genetic algorithms is depicted in Figure 2.

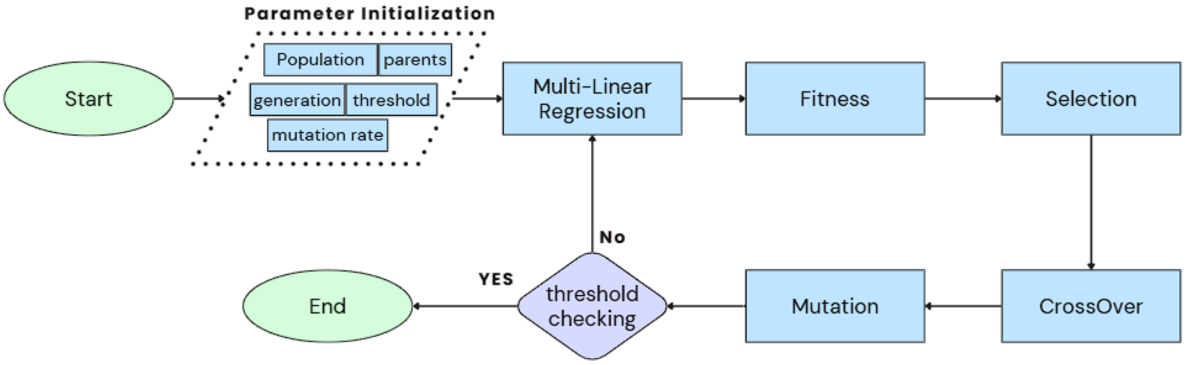

Figure 2 Hybrid algorithm between regression – genetic algorithm

The implementation of the hybrid genetic algorithm and multiple linear regression, based on Figure 2, includes the following steps:

1. **Initialize Parameter Values:** Define key parameters such as population size (population_size), maximum generations (num_generations), mutation probability (mutation_rate), number of parents retained and mated (num_parents), and fitness threshold (threshold_mape).
2. **Model Evaluation:** Evaluate the multiple linear regression model to generate outputs for each individual in the population.
3. **Calculate Fitness:** Compute the fitness value for each individual in the population using MAPE.
4. **Selection:** Select individuals from the population to form a parent subpopulation based on parameter settings.
5. **Crossover:** Perform crossover operations on the parent subpopulation, pairing parents randomly to generate an offspring subpopulation.
6. **Mutation:** Apply random mutations to offspring individuals based on the mutation probability.

7. **Threshold Check:** Determine whether the stopping criteria have been met. If not, return to Step 2 using the combined population of parents and offspring for the next generation. The threshold criteria are satisfied when at least one individual achieves the fitness threshold (threshold_mape).

4 Result and Discussion

4.1 Genetic Algorithm Parameters

In this study, the genetic algorithm is utilized to estimate and optimize the parameters of the multiple linear regression model. The parameters of the genetic algorithm used in this research are presented in Table 3.

Table 3 Genetic algorithm parameters

No	Parameters	Value
1	Population	20
2	Generation	1000
3	Mutation Rate	0.1
4	Threshold Mape	0.05
5	Parent	10

Based on Table 3, the genetic algorithm parameters were selected through initial experiments with various configurations to identify the settings that produce the best results.

4.2 Result

4.2.1 The Best Chromosome

After iterating up to the 1000th generation, the best individual or best chromosome was identified. This chromosome has an optimal regression coefficient, resulting in the best performance for predicting sales data. The best chromosome of each generation is shown in Table 4.

Table 4 Best chromosome over generations

Gen	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	$\boldsymbol{\beta_5}$	$\boldsymbol{\beta_6}$	Fitness
1	0.45462	0.20541	0.20137	0.51403	0.08722	0.48358	0.36217	1.0244
5	0.4562	0.43552	0.13305	0.05773	0.08287	0.10334	0.27848	0.3053
10	0.20724	0.43352	-0.0157	0.05773	0.08287	0.10334	0.27848	0.2452
…	….	….	….	….	….	….		….
1000	63.5838	0.15481	-0.0156	0.11582	0.06003	0.04486	0.55720	0.21971

Based on Table 4, the first generation produced a fitness value of 1.0244, while the fifth generation achieved a fitness value of 0.3053. This iterative process continued until the 1000th generation, where the best chromosome was obtained. This chromosome resulted in a multiple linear regression function with the lowest fitness value among all chromosomes, at 0.21971. Thus, the multiple linear regression model equation for this study is presented in Equation 3.

$$Y = 63.5838 + 0.15481X_1 - 0.0156X_2 + 0.11582\mathrm{X}_3 + 0.06003\mathrm{X}_4 + 0.04486X_5 + 0.5572X_6 \tag{3}$$

4.2.2 Model Performance

The multiple linear regression model estimated using the genetic algorithm demonstrated a significant improvement in performance across generations. The training results of the multiple linear regression model optimized through the genetic algorithm are shown in Figures 3 and 4.

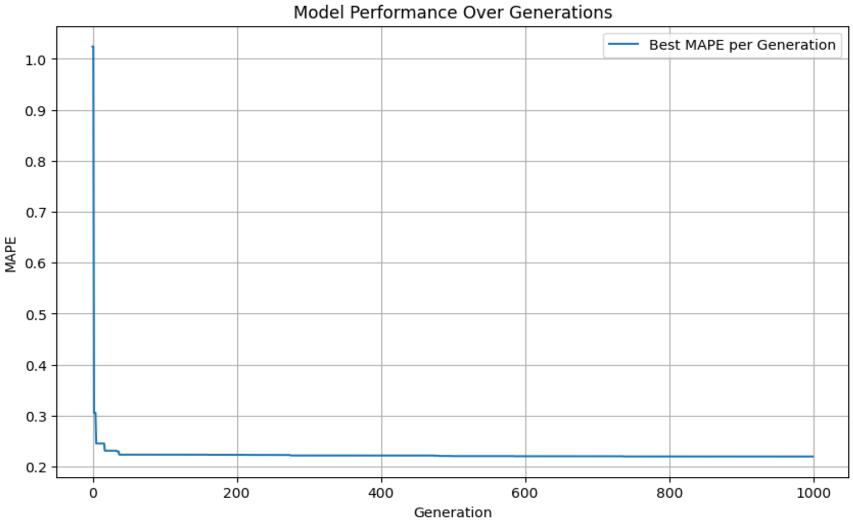

Figure 3 Model performance over generations

Figure 3 illustrates the model's performance during the optimization process using the Genetic Algorithm (GA) on the training data. The graph in figure 3 shows a significant decrease in the MAPE (Mean Absolute Percentage Error) up to the 50th generation, after which it stabilizes around 0.22. Meanwhile, Figure 4 presents a comparison between Actual Sales (X-axis) and Predicted Sales (Y-axis) generated by the linear regression model with optimized coefficients and intercept. The red dashed line indicates the ideal prediction, where predicted values perfectly match actual values.

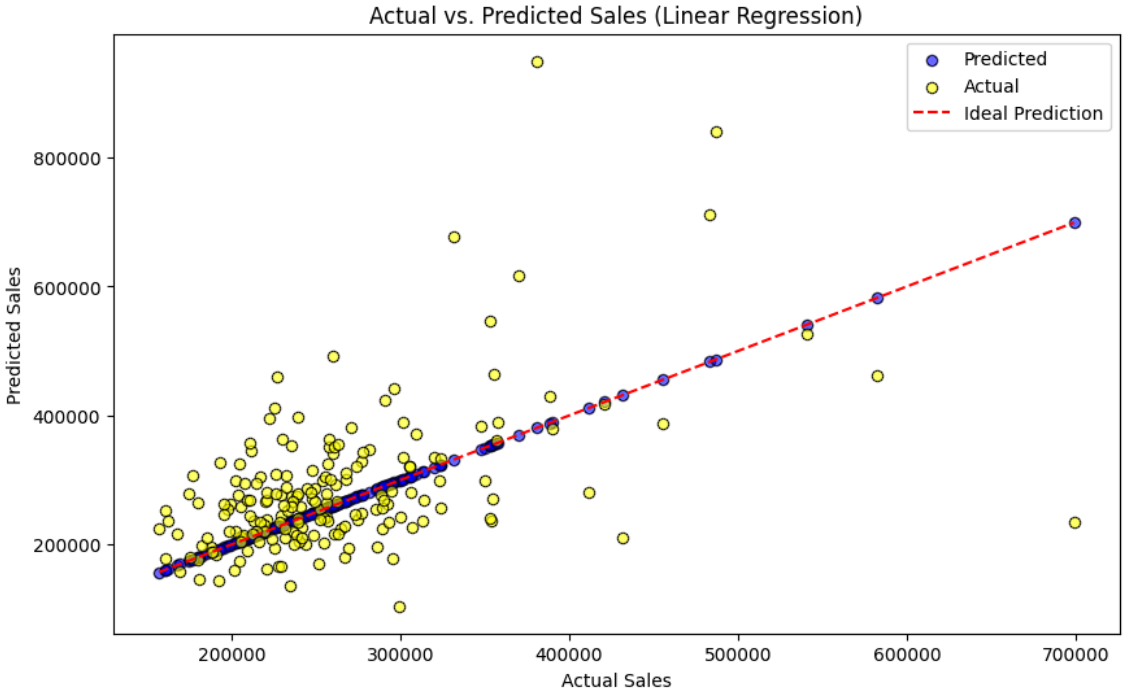

Figure 4 Comparation between actual and predicted sales

After determining the regression coefficients for the multiple linear regression model using the genetic algorithm, the chromosome with the best results during the training process was utilized to test the model using validation data. The results of this testing are shown in Table 5.

Table 5 Model performance result

No	Data	MAPE
1	Training	0.2197
2	Validation	0.2665

Based on Table 5, the findings indicate that the multiple linear regression model with parameters estimated using the genetic algorithm not only performed well on the training data but also maintained relatively high performance on the validation data. This is evaluated by the Mean Absolute Percentage Error (MAPE) values of 0.2197 on the training data and 0.2665 on the validation data.

4.2.3 Sales Forecasting

The next step in the analysis involves forecasting sales results for the upcoming days. The forecasting process utilizes the multiple linear regression model with the best parameters, as estimated by the genetic algorithm and detailed in Section 4.2.2. In this research, the multiple linear regression model was used to predict sales outcomes for the next 21 days. The results of the sales forecasting are presented in Figure 5.

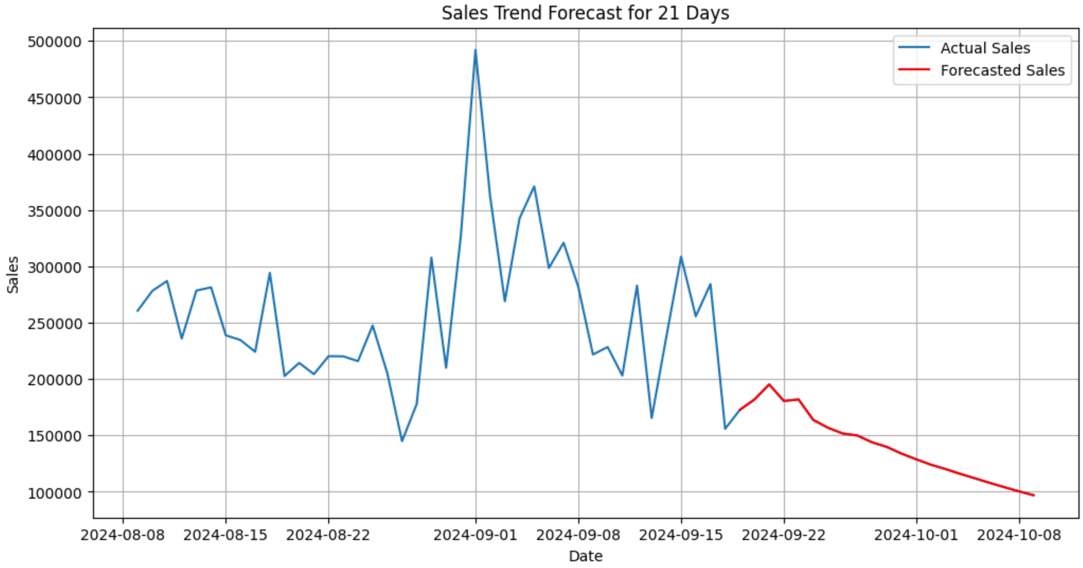

Figure 5 Sales trend forecast for 21 days

Based on figure 5, the sales trend is predicted to decrease significantly over the next 21 days. This decreasing trend might be attributed to seasonal factors, changes in customer demand, or market conditions affecting sales. With a MAPE of 0.2665 on validation data, the model is quite accurate to make the prediction reliable for business decision-making.

5 Conclusion

The implementation of a Genetic Algorithm to optimize parameter estimation in the Multiple Linear Regression model, as expressed in equation (3), has demonstrated its effectiveness in analyzing sales trends and forecasting outcomes based on historical data.

The optimized model achieved a Mean Absolute Percentage Error (MAPE) of 0.2197 on the training data and 0.2665 on the validation data, indicating a reasonable level of predictive accuracy. These results highlight the potential of GA in enhancing regression model performance.

For future research, it is recommended to investigate other advanced machine learning approaches, such as the family of Gradient Boosting algorithms (e.g., XGBoost, LightGBM, CatBoost) or time series forecasting models (e.g., ARIMA, SARIMA, LSTM). Incorporating Genetic Algorithm for hyperparameter optimization in these models is expected to yield more accurate predictions and deeper insights into sales trends. This approach has the potential to develop more sophisticated forecasting models, thus contributing to improved decision-making in sales analysis.

6 References

[1] Bakri, R., Halim, A., Probondani, N., Sekolah, A., Ekonomi, T. I., & Bongaya, M. (2018). SISTEM INFORMASI STRATEGI PEMASARAN PRODUK DENGAN METODE MARKET BASKET ANALYSIS DAN SALES FORECASTING: SWALAYAN KOTA MAKASSAR. In *Jurnal Manajemen Teori dan Terapan Tahun* (Vol. 11, Issue 2).

[2] Guru, P., Sathyapriya, J., Rajandran, K. V. R., Bhuvaneswari, J., & Parimala, C. (2023). Product Sales Forecasting and Prediction Using Machine Learning Algorithm. In *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE* (Vol. 2024, Issue 4s).

[3] Wisudaningsi, B. A., Arofah, I., Konstansius, D., & Belang, A. (2019). PENGARUH KUALITAS PELAYANAN DAN KUALITAS PRODUK TERHADAP KEPUASAN KONSUMEN DENGAN MENGGUNAKAN METODE ANALISIS REGRESI LINEAR BERGANDA. *Jurnal Statistika Dan Matematika), 1*(1).

[4] Windarto. (2015). *Modul Perkuliahan Simulasi Pengantar Algoritma Genetika*.

[5] D.E. Goldberg, 1989, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley Publishing Company, Inc.

[6] Mariani, T., & Rosyida, I. (2023). Implementasi Metode Double Exponential Smoothing untuk Peramalan Luas Panen Padi di Kabupaten Pati dengan Bantuan Software Minitab 16. *PRISMA, Prosiding Seminar Nasional Matematika*, *6*, 707–713. https://journal.unnes.ac.id/sju/index.php/prisma/

[7] Ottaviani, F. M., & Marco, A. de. (2021). Multiple Linear Regression Model for Improved Project Cost Forecasting. *Procedia Computer Science*, *196*, 808–815. https://doi.org/10.1016/j.procs.2021.12.079

[8] Gustriansyah, R., Ermatita, E., & Rini, D. P. (2022). An approach for sales forecasting. *Expert Systems with Applications*, *207*. https://doi.org/10.1016/j.eswa.2022.118043.