Dini Fakta Sari[1]
Informatics
Faculty of Information Technology
Universitas Teknologi Digital Indonesia, Yogyakarta
email: dini@utdi.ac.id

Muhammad Ali Sofian
Master in Information Technology
Faculty of Information Technology
PT Sinergi Informatika Semen Indonesia
email: muh.sofian@sisi.sig.id

Agung Wilis Nurcahyo
Master in Information Technology
Faculty of Information Technology
email: students.agungwilis24@mti.utdi.ac.id

Kelik Wiharyanto
Master in Information Technology
Faculty of Information Technology
email: students.kelikwihar24@mti.utdi.ac.id

Elisabet da Conceição Pereira
Master in Information Technology
Faculty of Information Technology
email: ejhyperreira52@gmail.com

Ali Impron
Informatics
Faculty of Engineering and Agriculture
email: ali.impron@umsa.ac.id

Sales Prediction of Vegetables Seed Products Using Simple Linear Regression

The growth of the modern agricultural sector drives the need for an accurate sales prediction system, especially for vegetable seed products that are highly dependent on the season and market demand. An imbalance between stock and demand can cause losses, either in the form of overstock or undersupply. This condition requires a data-based planning strategy to ensure stock availability according to actual needs in the field. A historical data-based sales prediction approach is a relevant solution to optimize the distribution and procurement process. This study aims to apply a simple linear regression method in predicting vegetable seed sales based on historical data for one year. The prediction model is built using the time variable (month) as the independent variable and the number of seed requests as the dependent variable. This technique was chosen because of its ability to identify linear relationship patterns between time and sales trends in a simple but effective way. The data used comes from internal records of farmers and distributors, which are then classified into two main categories: leafy vegetable seeds (spinach, kale, mustard greens) and fruit vegetable seeds (tomatoes, chilies, eggplants). The results of the study showed that simple linear regression was able to provide fairly accurate predictive results. This model can be used as a basis for decision making in production planning, supply chain management, and seed inventory management, thus supporting the efficiency of farming businesses and reducing potential losses due to mismatches between demand and supply.

***KeyWords**: Sales Prediction, Vegetable Seeds, Simple Linear Regression, Agriculture*

This Article was:
submited: 25-06-25
accepted: 08-07-25
publish on: 20-07-25

How to Cite:
D.F. Sari, et al, "Sales Prediction of Vegetables Seed Products Using Simple Linear Regression", Journal of Intelligent Software Systems, Vol.4, No.1, 2025, pp.1–4, 10.26798/jiss.v4i1.2001

1 Introduction

In modern agricultural systems, managing agricultural product stocks such as vegetable seeds is a crucial aspect to ensure the continuity of production and distribution. Accuracy in planning the amount of seed stock is a challenge, considering the fluctuations in demand due to seasons, weather, and consumer behavior. Historical data-based sales prediction is a relevant approach to support supply planning. One of the statistical methods widely used for prediction is simple linear regression, because it is able to identify trend patterns based on time variables [1]. This study aims to develop a vegetable seed sales prediction model to optimize the distribution and procurement process of agricultural products. This prediction model is built using a simple linear regression approach that utilizes historical data on vegetable seed sales. The data is analyzed to determine the relationship between time (in months) as an independent variable and seed sales volume as a dependent variable [2].

By applying this method, it is expected to be able to provide a more accurate estimate of sales in the following month, especially for December which often experiences an increase in demand due to various factors, such as the start of the planting season or large-scale agricultural activities. The accuracy of this model is evaluated using two main metrics, namely Mean Absolute Deviation (MAD) and the level of prediction accuracy against actual data [3]. MAD is used to calculate the average absolute difference between the predicted value and the actual value, so that it can provide an overview of the level of prediction error in sales units. Meanwhile, the evaluation of specific accuracy in December provides validation of the model's ability to estimate realistic values and can be used by agricultural business actors as a basis for decision making [4]. This analysis process is also supported by a MySQL-based database system, where data from various sources such as weekly reports and Excel files are collected, cleaned, classified, and normalized using the Min-Max Scaling method [5]. Seed data is classified into two main groups, namely leaf vegetable seeds and fruit vegetable seeds, to facilitate the transformation process and the preparation of structured datasets [6].

This study provides a practical contribution to the development of decision support systems in agricultural supply chain management. It is hoped that the prediction results produced by this model can be utilized by farmers, distributors, and agribusiness actors in developing more efficient production and distribution strategies, as well as reducing the risk of excess or shortage of stock in the field.

2 Methodology

2.1 Dataset. Data was obtained from sales records during the period from January to November 2022. Data includes the number

[1]Corresponding Author.

Table 1 Dataset

Month	Leaf Vegetable Seeds	Fruit Vegetable Seeds
January	120	150
February	200	180
March	240	210
April	210	190
May	190	175
June	130	160
July	220	200
August	180	170
September	250	230
October	270	240
November	260	225

Table 2 Dataset

Month	Leaf Vegetable Seeds	Fruit Vegetable Seeds
January	0	0
February	0.533333333	0.333333333
March	0.8	0.666666667
April	0.6	0.444444444
May	0.466666667	0.277777778
June	0.066666667	0.111111111
July	0.666666667	0.555555556
August	0.4	0.222222222
September	0.866666667	0.888888889
October	1	1
November	0.933333333	0.833333333

of vegetable seed units sold per month and is grouped into two categories: leaf vegetable seeds and fruit vegetable seeds, it show on Table 1.

2.2 Linear Regression. The simple linear regression method is used to find the relationship between time (month) and sales volume [7]. This approach is particularly useful when the goal is to identify consistent trends over a specified period and forecast future outcomes based on historical data patterns [8]. By modeling the data with a linear equation, it becomes possible to estimate future sales with a degree of confidence, enabling better inventory planning, production scheduling, and strategic decision-making in the agricultural supply chain [9]. The linear regression formula is:

$$Y = a + bX \tag{1}$$

Where:
Y = Sales Amount
X = Month-n
a = Intercept
b = Regression Coefficient

Calculating Constants (a):

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \tag{2}$$

Calculating the Coefficient (b):

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \tag{3}$$

2.3 Evaluation. To assess the performance of the simple linear regression model used in predicting vegetable seed sales, an evaluation process was carried out using two approaches, namely Mean Absolute Deviation (MAD) and prediction accuracy against actual data for December 2022.

2.3.1 Mean Absolute Deviation (MAD). Mean Absolute Deviation (MAD) is a statistical evaluation method used to measure the accuracy of a predictive model by calculating the average absolute difference between the actual value and the predicted value. MAD provides an overview of how much error the model's prediction is in the same units as the original data, regardless of the direction of the error (positive or negative) [10].

The MAD formula is as follows:

$$MAD = \frac{1}{n}\sum_{n}^{t=1}[yt - \widehat{y}t] \tag{4}$$

Where:
yt = Actual value at period - t
$\widehat{y}t$ = Actual value at period - t
n = Total number of periods (months)

A lower MAD value indicates that the model has better predictive ability because it produces smaller errors [9]. In the context of this study, MAD is used to evaluate the extent to which a simple linear regression model is able to approach the reality of vegetable seed sales from month to month. By calculating MAD from actual and predicted data, the overall effectiveness of the model can be determined. The use of MAD also helps in comparing the performance of several predictive models in the future, such as Random Forest or LSTM, quantitatively.

2.3.2 December Month Prediction Accuracy. Prediction Accuracy is a statistical measure used to assess how close a model's predictions are to the actual values that occur. In the context of regression models, prediction accuracy indicates how well a model is able to estimate realistic values that can be used in decision making. The prediction accuracy is also calculated specifically for December 2022, by comparing the predicted value to actual sales as follows:

$$\text{Accuracy} = \left(1 - \frac{|y - \widehat{y}|}{y}\right) \times 100\% \tag{5}$$

Where:
y = Actual value of December sales
$\widehat{y}$ = Predicted result of December sales

In this study, the prediction accuracy is calculated specifically for December 2022 using the prediction results from a simple linear regression model. If the predicted value is very close to the actual value, then the accuracy percentage will be close to 100%, indicating that the model has good performance. Conversely, if the predicted value is far from the realization, then the accuracy will decrease, and the model needs to be refined.

The use of Prediction Accuracy is very important in modern agricultural systems, because it helps farmers and distributors make more informed decisions regarding stock provision, seed delivery, and planting planning, thereby reducing potential losses due to mismatches between demand and supply.

2.4 Data Extraction and Processing System.

2.4.1 Dataset. The data comes from the internal recording system of farmers and seed distributors, in the form of .xls documents and weekly reports. All data is stored in the bibitDB database using MySQL and has gone through a normalization process using the Min-Max Scaling method so that each feature is in the range of 0 to 1 to support the analysis process and predictive modeling more optimally as seen on Table 2.

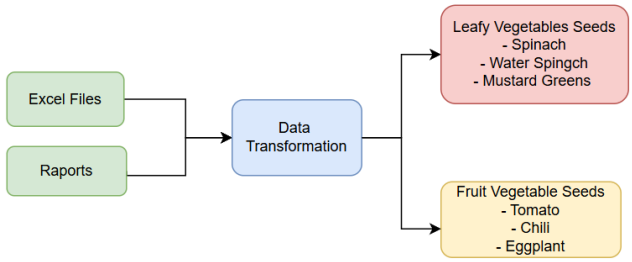

Fig. 1 Data Transformation and Classification Stages

2.5 Transformation and Classification. In this stage, seed sales data originating from various formats such as Excel files (.xls), weekly farmer reports, and system input results, are transformed so that the data is ready to be used in the predictive analysis process. One important step in the transformation is data classification based on the type of seed product sold. The data is classified into two large groups, namely:

a) Leafy Vegetable Seeds, which include products such as spinach, kale, and mustard greens. This type of seed generally has a short planting cycle and fluctuating seasonal demand.
b) Fruit Vegetable Seeds, which include products such as tomatoes, chilies, and eggplants. These seeds usually have a longer planting cycle and are related to household consumption patterns and traditional markets.

This classification aims to facilitate predictive modeling and provide more targeted segmentation for analysis needs. Data transformation, as seen on Figure 1 includes combining data sources into one integrated schema, data cleaning, and converting quantitative values into a ready-to-process numeric format. The result of this stage is a well-structured and classified dataset, which is then continued to the normalization stage and simple linear regression analysis for sales prediction purposes [11].

3 Results and Discussion

3.1 Data analysis. Based on the results of calculations on seed sales data during January–November, a regression equation was obtained:

$$Y = 217.49 + 4.51X \tag{6}$$

With $X = 12$ (December), sales prediction:

$$Y = 217.49 + 4.52(12) = 271.61$$

If the actual data for December is 255 units, then the accuracy is:

$$\text{Akurasi} = \left(\frac{|271.61 - 255|}{271.61}\right) \times 100\% \approx 94\%$$

These results show that the simple linear regression model is quite effective in predicting vegetable seed sales.

3.2 Visualization and Model Evaluation. The predictive performance of the simple linear regression model is further assessed through visual and tabular representations that highlight the comparison between actual and predicted sales data [12–14]. The line chart displays actual and predicted vegetable seed sales from January to December. The orange dashed line shows a rising trend, helping farmers anticipate stock needs based on seasonal patterns.

Figure 2. Presents a visualization of the comparison between actual vegetable seed sales data and the predicted results generated

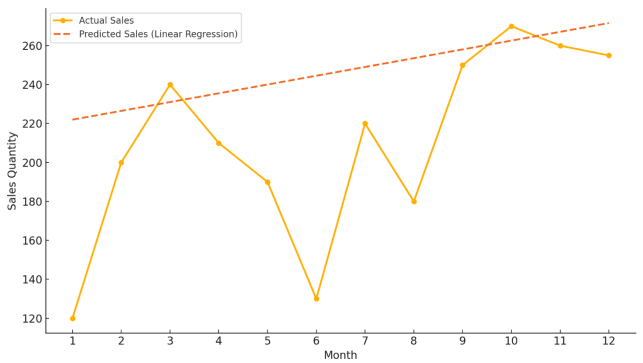

Fig. 2 Linear Regression of Monthly Vegetable Seed Sales

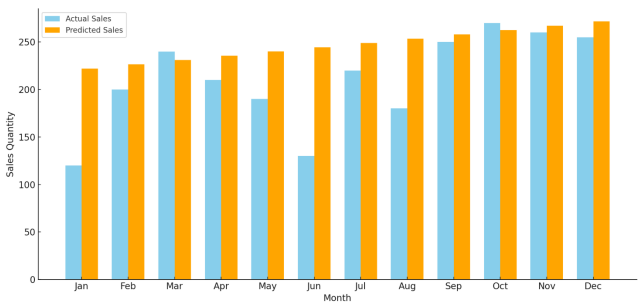

Fig. 3 Comparison of Actual and Predicted Sales

through a simple linear regression method [15]. This graph maps sales data from January to December in one year, with the horizontal axis representing time (month) and the vertical axis showing the number of sales units. The solid yellow line represents actual sales, while the dashed orange line shows the predicted results from the model. The difference between the two lines illustrates the deviation of the prediction from the sales realization. This graph is used to assess the extent to which the regression model is able to map historical trends and estimate future values. From the graph, it can be seen that the predicted pattern has a consistent upward trend, although fluctuations in actual data indicate seasonal dynamics or other external factors. This visualization is supporting evidence in evaluating the effectiveness of predictive models in the context of stock management and supply planning in the agricultural sector.

Table 3 Presents a month-by-month comparative data between the actual sales volume and the predicted results of a simple linear regression model for each month for one year. This table is an important element in evaluating the performance of the prediction model, by including a column that shows the difference or error between the actual and predicted values. Error is calculated as the result of subtracting the actual value from the predicted value, which is then interpreted to assess how much the model deviates from reality. A small error value indicates a more accurate level of prediction, while a large error value indicates the need for model improvement. Presenting data in a table like this makes it easier to analyze quantitatively and identify patterns of prediction discrepancies in a more structured manner. This table plays an important role as a basis for improving seed distribution and procurement strategies in a dynamic modern agricultural system.

Figure 3. Complements the previous analyses by offering a direct side-by-side visual comparison a visualization in the form of a bar chart that compares the actual sales value and the predicted sales value of vegetable seeds for each month throughout the year. The blue color represents the actual sales data, while the orange color depicts the predicted results of a simple linear regression

Table 3 Comparison of Actual Sales Value and Predicted R based on Linear Regression for Each Month

Month	Actual Sales	Predicted Sales	Error (Actual - Predicted)
January	120	2220	-1020
February	200	22651	-2651
March	240	23102	898
April	210	23553	-2553
May	190	24004	-5004
June	130	24455	-11455
July	220	24906	-2906
August	180	25357	-7357
September	250	25808	-808
October	270	26259	741
November	260	2671	-71
December	255	27161	-1661

model. This visualization provides a clear picture of the model's consistency in following real sales patterns, as well as identifying months where there is a significant deviation between the actual value and the predicted results. This diagram helps analysts and agribusiness actors to more easily evaluate the performance of the predictive model comprehensively and intuitively. In addition, this graph is also useful in the model validation process and the preparation of efficient seed stock distribution planning strategies. With this visualization, the analysis results are not only presented in numerical form, but also in a communicative and informative form to support data-based decision making.

4 Conclusions

A simple linear regression-based sales prediction model has proven effective in modelling vegetable seed sales trends. This model can be used by distributors or agricultural business actors to plan supplies more efficiently, thereby reducing the risk of losses due to overstock or shortages. In the future, the use of more complex predictive models such as Random Forest or LSTM can be explored to consider seasonality, price, and weather factors. In addition, integration with geographic information systems (GIS) and real-time data from agricultural sensors can improve prediction accuracy. The integration of Internet of Things (IoT) technology to monitor land and environmental conditions also has the potential to enrich the input variables in the model. Thus, the sales prediction system is not only reactive, but also adaptive and responsive to complex and changing agricultural dynamics.

References

[1] Said, Z., Vigneshwaran, P., Shaik, S., Rauf, A., and Ahmad, Z., 2025, "Environmental and Sustainability Indicators Climate and carbon policy pathways for sustainable food systems," Environmental and Sustainability Indicators, **27**(November 2021), p. 100730.

[2] Li, J. et al., 2012, "Product Sales Forecasting," Foods.

[3] Guido, Z. et al., 2020, "Climate Risk Management Farmer forecasts: Impacts of seasonal rainfall expectations on agricultural decision-making in Sub-Saharan Africa," Climate Risk Management, **30**(August 2019), p. 100247.

[4] Chung, C., 2017, "A Sales Forecast Model for Short-Life-Cycle Products: New Releases at Blockbuster," Production and Operations Management, Originally published September 2012.

[5] Linoff, G. S., 2007, *Data Analysis Using SQL and Excel*.

[6] Mendes, F. C., Sarna, P., Emelyanov, P., and Dunlop, C., 2023, *Database Performance at Scale*.

[7] Montgomery, D. C., Peck, E. A., and Vining, G. G., 2012, *Linear Regression Analysis*.

[8] Harianto, F. J. and Abdulloh, F. F., 2023, "Linear Regression Algorithm Analysis to Predict the Effect of Inflation on the Indonesian Economy," Indonesian Journal of Computer Science, **12**(4), pp. 1673–1681.

[9] Mustapha, O. O., Sithole, T., and Mustapha, O. O., 2025, "Forecasting Retail Sales using Machine Learning Models," American Journal of Statistical and Actuarial Sciences, **6**(1), pp. 35–67.

[10] Mohammed, S. et al., 2025, "The effects of data quality on machine learning performance on tabular data," Information Systems, **132**(March), p. 102549.

[11] Mumuni, A. and Mumuni, F., 2025, "Automated data processing and feature engineering for deep learning and big data applications: A survey," Journal of Information Intelligence, **3**(2), pp. 113–153.

[12] Dong, C., 2024, "Application of Multiple Linear Regression on Sales Prediction," Unknown, **45**, pp. 159–164.

[13] Azure, I., 2024, "Predictive modeling for industrial productivity: Evaluating linear regression and decision tree regressor approaches," Unknown, **2**(4), pp. 1–12.

[14] Sharif, M. S., unknown, "A Comparative Study of Sales Prediction Using Machine Learning Models: Integration of PySpark and Power BI," Unknown.

[15] Pratama, M. A., 2023, "Utilizing Linear Regression for Predicting Sales of Top- Performing Products," Unknown, **01**(03), pp. 174–180.