Published online on the page : https://ejournal.uinbukittinggi.ac.id/index.php/ijokid/

Knowbase : International Journal of Knowledge in Database

| ISSN (Print) 2798-0758 | ISSN (Online) 2797-7501 |

Optimization Of Agricultural Production In South Sumatera Using Multiple Linear Regression Algorithm

Dedi Setiadi[1,], Sasmita[2], Yogi Isro Mukti[2]*

[1,2,3]*Institut Teknologi Pagar Alam,Kota Pagar Alam,* Indonesia

Article Information

Article History:

Submitted: November 23, 2024
Revision: November 29, 2024
Accepted: December 13, 2024
Published: December 24, 2024

Keywords

Agricultural
CRISP-DM
Machine Learning
Multiple Linear Regression
Optimization

Correspondence

E-mail: dedisetiadi1212@gmail.com*

A B S T R A C T

Rice is one of the agricultural commodities in South Sumatra whose productivity level still fluctuates. In 2000, rice production reached 1,863,643.00 kg, then increased to 3,272,451.00 kg, in 2010, but decreased again in 2020 to 2,696,877.46 kg. This instability is influenced by various factors such as land area, rainfall, pest attacks, and fertilizer use. This study aims to optimize rice production by applying machine learning using multiple linear regression algorithms, and the CRISP-DM method, with the stages being business understanding, data understanding, data preparation, modeling, evaluation, and implementation. Data of 1,000 records obtained from farmers were analyzed using Google Collaboratory, resulting in an intercept of -3836,2639, and coefficients for land area of 5,7336, rainfall of 1,2710, pests of 6,1153, urea of 1,6226, and phonska of 1,2581. To evaluate the accuracy of rice production predictions based on these independent variables, calculations were made on the RMSE value and analysis of the coefficient of determination. The results were that the RMSE value was recorded at 17065084,9641, and the coefficient of determination (R^2) was 0,6487, indicating that around 64,87 % of the variability in rice production can be explained by independent variables such as land area, rainfall, pest attacks, use of urea fertilizer, and phonska, while the remaining 35,13 % was influenced by other factors.

1. Introduction

Indonesia is a country consisting of many islands, and is one of the largest archipelagic countries in the world [1] consists of 38 provinces stretching from Sabang to Merauke, one of which is the province of South Sumatra. This province is known for having a region with very large agricultural potential. [2] Due to having quite fertile land spread across various regions and the natural beauty of its tropical forests, South Sumatra has become one of the main agricultural centers in Indonesia, which contributes significantly to national food production, especially in the plantation, rice field and horticulture sectors. [3] One of the superior agricultural commodities produced is rice with good quality, but rice productivity in the province of South Sumatra is not yet stable and still fluctuates from year to year, as can be seen from data obtained from the website, https://www.kaggle.com/datasets/ardikasatria/datasettanamanpadisumatera, In 2000, the amount of rice production was 1,863,643.00 kg, in 2010 the amount of rice production increased to 3,272,451.00 kg, and in 2020 the amount of rice production decreased to 2,696,877.46 kg, as shown in Table 1 below :

Tabel 1. Rice Production Data in South Sumatra

No	Year	Production (kg)	Harvest area (m^2)
1	2000	1.863.643,00	555.427,00
2	2001	1.723.433,00	511.928,00
3	2002	1.899.849,00	561.724,00
4	2003	1.977.345,00	570.010,00
5	2004	2.260.794,00	625.013,00
6	2005	2.320.110,00	626.849,00
7	2006	2.456.251,00	646.927,00
8	2007	2.753.044,00	691.467,00
9	2008	2.971.286,00	718.797,00
10	2009	3.125.236,00	746.465,00
11	2010	3.272.451,00	769.478,00
12	2011	3.384.670,00	784.820,00
13	2012	3.295.247,00	769.725,00
14	2013	3.676.723,00	800.036,00
15	2014	3.670.435,00	810.900,00
16	2015	4.247.922,00	872.737,00
17	2016	4.881.089,00	615.184,00
18	2017	4.807.430,00	621.903,00
19	2018	2.994.191,84	581.574,61
20	2019	2.603.396,24	539.316,52
21	2020	2.696.877,46	551.320,76

The instability of rice production is caused by various factors, [4] namely land area, rainfall, pests and fertilizer use. [5] Rice farming has a strategic role in supporting food security [6] in South Sumatra. The problem faced is that farmers have not been able to optimize the factors that cause the rise and fall of rice production, resulting in instability in their production results, which has an impact on food availability in South Sumatra or nationally. Various methods are used to increase rice production, one of which is by utilizing computer technology, namely artificial intelligence, which is a development of computer technology [7] which has the ability to carry out various tasks in various fields of life well [8] and offers great potential to increase productivity such as in the agricultural sector, especially rice. Machine learning is a method in artificial intelligence [9] [10] that aims to imitate human abilities in optimizing problem solving. [11]

Machine learning can be used to analyze data and produce models that can be used to predict [12] agricultural yields using the right algorithm, one of the algorithms for predicting agricultural yields is multiple linear regression. This algorithm [13] can be used to predict agricultural production results, especially rice commodities, by analyzing data on the relationship between various variables, [14] such as land area, rainfall, pests and fertilizer use by farmers, by utilizing historical data on rice production in South Sumatra and also related factors. The results of this study can provide innovative solutions to increase rice production in South Sumatra. By understanding more deeply about the factors that influence rice production results, and can develop more effective and sustainable strategies, and help optimize rice production results in South Sumatra so as to support national food security.

Machine learning, [15] used to analyze multivariate factors, where several independent variables can be identified and their impact on rice production can be measured. [16] Machine learning [17] can be used to predict rice production [18] based on various factors, such as land area, rainfall, pests, and fertilizer use, so that it can help farmers plan and optimize their rice production. This algorithm is a mathematical model used to describe the relationship between several independent variables and dependent variables.

In this case, these independent variables can include elements such as land area, rainfall, pests, and fertilizer use, while the dependent variable is the harvest. Through the application of multiple linear regression, the complex interactions between these factors can be identified and analyzed more deeply, resulting in a model that is able to predict results with higher accuracy. This model is a very valuable tool in the decision-making process for smarter and more optimal agricultural land management, thereby helping to increase the effectiveness and efficiency of agricultural practices. [19] Multiple linear regression allows researchers and farmers to evaluate the effect of each independent variable on production yields

simultaneously. For example, by knowing how the combination of rainfall and fertilizer use affects crop yields, farmers can allocate resources more wisely to maximize yields, even under less than ideal conditions. Multiple linear regression is not only a prediction tool, but also a mechanism for understanding the dynamic relationship between environmental factors and agricultural yields. [20] In other words, this method allows us to estimate the desired results based on a series of factors that influence it. [21]

Machine learning can be used to predict [22] rice production, so it can help farmers to plan and optimize their rice production, and reduce the risk of crop failure. [23] Based on research conducted by Diyanti et al., entitled "Prediction of Rice Harvest Results in 2023 Using Linear Regression Method in Indramayu Regency", [24] this study produces predictions of rice production results with a linear regression algorithm, and also in research conducted by Devi Wulandari and Rumini entitled "Modeling and Prediction of Rice Production Using Linear Regression", [25] from The two previous studies in general only provide predictions of rice production for the following year without providing in-depth input on the factors that influence rice production results. This reduces the potential for using these predictions in strategic decision making. However, the state of the art in this study lies in the development of a web-based prediction application or system that not only functions to predict rice production results, but also allows detailed analysis of the factors that influence production. With this system, users can gain more comprehensive insights into key variables, such as weather, soil quality, and the use of agricultural technology, as well as get advice to improve agricultural efficiency and productivity, [26] which can be used by anyone, anywhere and anytime needed online[27].

A prediction system is [28] a system that can be used to predict a future condition by referring to past information and/or present information, which is computer-based. [29] This application was built using machine learning technology, to predict rice production results in South Sumatra, using a multiple linear regression algorithm. This application works by studying historical data on rice production in the region, including various influencing variables, such as land area, rainfall, pests and fertilizer use. Through the learning process from historical data, the application is able to provide more accurate predictions regarding future production results [30]. Thus, farmers and decision makers in the agricultural sector can anticipate changes in production and take strategic steps to increase agricultural productivity[31]. The prediction system is made by including a multiple linear regression algorithm that describes the relationship between independent variables[32]. This algorithm is a novelty in this study, where previous studies only used a linear regression algorithm with one independent variable, which was carried out by Diyanti et al. And also the system [33] that has been created is based on a simple web [34] so that it makes it easier for users, namely farmers, to run it [35], so that farmers can predict the results of rice production that will be planted, by knowing what factors can influence increasing rice production, so that it is hoped that rice production can be more optimal and support national food security.

2. Method

The method used in this study is CRISP-DM which stands for Cross-Industry Standard Process for Data Mining. [36] This method serves as a standard approach to managing data mining projects, offering a structured framework that helps in the process of extracting information or knowledge from data, as in Figure 1.

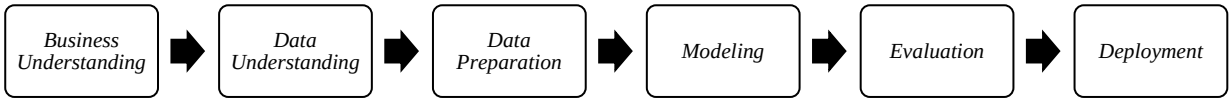

Figure 1. CRISP-DM Method

CRISP-DM is a widely recognized methodology, [37], due to its flexibility in various industries, including agriculture, to organize and manage data mining projects. In the context of rice production prediction, CRISP-DM facilitates comprehensive data analysis with structured stages from start to finish of the process, ensuring that each stage from data collection to interpretation of results runs in a structured manner. By using the CRISP-DM approach, understanding agricultural factors such as rainfall, fertilizer use, or pest attacks is an important part of the business understanding stage. After that, data preparation is carried out to ensure that the data analyzed is valid and relevant, followed by modeling using the selected algorithm, such as multiple linear regression. This method ensures that the analysis is carried out thoroughly and is oriented towards the end goal, [38] namely providing results that can be used in decision making to increase rice production.

CRISP-DM allows for clear guidance in each stage, making it an effective and efficient method in processing and analyzing big data, and helps bridge the gap between raw data and actionable information for improved production results. The explanation of the stages is as follows : [39]

- Business understanding, The initial stage involves understanding the agricultural context and the various factors that influence production results, to ensure that the model applied is in accordance with the problems faced by farmers in the field. Next, data obtained from farmers and other sources are prepared and processed for analysis using a multiple linear regression model. This step also includes determining the objectives of rice production predictions, such as estimating harvest results, identifying factors that influence production results, developing agricultural strategies, and understanding the needs of farmers and related parties. Based on the results of observations and interviews, it was found that land area, rainfall, pest attacks, and the use of urea and phonska fertilizers are the main factors that influence rice production..
- Data understanding, once the business objectives are understood, the next step is to understand the available data. This includes collecting initial data, exploring data characteristics, and identifying data quality and limitations. Exploratory data analysis is performed to find patterns, outliers, or inconsistencies that may affect the analysis in the next stage.
- Data preparation, this stage includes various activities aimed at producing a final data set that is ready to use. [40] The data obtained at this stage is usually in raw form, which often has low quality, for example there are missing values, errors, or inconsistencies, [41] therefore, a data cleaning process is needed first. The cleaning steps include deleting duplicate data, filling in missing data, fixing inconsistent data, and correcting typing errors. In addition, the data is also further processed to ensure that its quality is adequate before being analyzed. Core activities in this stage include selecting relevant variables, cleaning data from anomalies or missing data, and transforming data to suit the model to be used. This data preparation stage is often the most time-consuming step in the CRISP-DM process.
- Modelling, this stage is when algorithms or statistical methods are applied to the prepared data. At this stage, various models such as regression, decision trees, clustering, or artificial neural networks are selected and applied. The team then tests and validates these models to find the model that best suits the business problem at hand.
- Evaluation, once the model is built, an evaluation phase is conducted to assess the quality and performance of the model in the context of business objectives. Evaluation involves testing the model against test or validation data and comparing the results to predetermined metrics. At this stage, researchers will also determine whether the model has answered the business problem correctly or whether additional steps are needed.
- Deployment, the final stage of CRISP-DM is the implementation of the model into operational systems. This includes disseminating the results of the analysis to stakeholders or integrating the model into business applications for use in day-to-day decision making. Documentation and maintenance of the model are also important parts of this stage to ensure that the model remains relevant and can be updated when needed.

3. Results and Discussion

The application of multiple linear regression algorithm in this study in the context of machine learning provides the ability to analyze the relationship between various variables that affect rice production, such as rainfall, land area, fertilizer use, and the number of pests. Multiple linear regression was chosen because of its ability to capture the linear relationship between several independent variables with the dependent variable, namely rice production yield. This prediction process not only provides an estimate of production results, but also allows for better decision making in the management of agricultural resources.

The CRISP-DM method ensures that the entire process runs in a structured manner and focuses on the final goal, namely obtaining accurate and useful predictions, with systematic and structured steps, as follows :

3.1. Business Understanding

The first stage, namely understanding the agricultural context and factors that influence production results, ensures that the model used is relevant to the problems faced by farmers in the field. After that, data collected from farmers and other sources are prepared and processed to be entered into a multiple linear regression model. Identifying the objectives of rice production prediction, such as estimating production results, identifying factors that influence rice production results, and planning agricultural strategies. and understanding the needs of farmers and stakeholders related to rice production predictions. And from the results of observations and interviews conducted, there are factors that influence rice production, namely land area, rainfall, pests and the use of urea and phonska fertilizers.

3.2. Data Understanding

Data collection on rice production was conducted through direct interview methods and distributing questionnaires to rice farmers in the city of Pagar Alam and its surroundings, which is one of the areas with the highest rice production levels in South Sumatra. This questionnaire focused on collecting information related to various factors that affect rice production. The data that was successfully collected was then presented in tabular form to facilitate analysis. The process of collecting data from farmers through interviews and questionnaires is very important to obtain direct information from the main source, namely farmers, who have a deep understanding of field conditions and factors that affect crop yields. By choosing the city of Pagar Alam and its surroundings as the research location, because this area is known as one of the significant rice production centers in South Sumatra, this study is expected to provide a representative picture of the condition of rice farming in the region.

3.3. Data Preparation

Data preprocessing, such as normalization, is an important step in statistical analysis because raw data is often not in a form that is ready for direct processing. Normalization, as one of the techniques, ensures that the variables used have the same range of values, so that the multiple linear regression algorithm can function optimally. This step also includes checking the validity of the data to ensure that there are no missing values, duplicate data, or anomalies that can affect the results of the analysis. There are several steps in data preparation that are carried out, the following are the stages that are carried out.

The data selection stage is an important initial step in the data analysis process, where relevant data is selected from a larger data set. This selection must be adjusted to the desired modeling objectives, so that only significant and relevant variables are included. In today's technological era, tools such as Google Colaboratory are often used to facilitate this process. With Google Colaboratory, we can write and run code collaboratively in a cloud environment that is integrated with Python and various data science libraries. The implementation of coding in Google Colaboratory allows researchers to select data more efficiently and flexibly, the coding is as follows :

```
The coding of select data
  import pandas as pd
  import numpy as np
  from sklearn.model_selection import train_test_split
  from sklearn.linear_model import LinearRegression
  from sklearn.metrics import mean_squared_error, r2_score
  import matplotlib.pyplot as plt
  import seaborn as sns
  df = pd.read_excel('Data_Padi_Baru.xlsx')
df
```

The results can be seen in Table 2, below :

Table 2. Rice Production Data and Factors Influencing It

Record	Harvest Yield (Y)	Land Area (X1)	Rainfall (X2)	Pest (X3)	Urea (X4)	Phonska (X5)
0	10000	1000	405	10	150	150
1	20000	2000	356	15	300	250
2	20000	1500	305	10	200	250
3	500	60	97	10	100	50
4	20000	1700	279	25	300	150
5	10000	1000	272	10	150	150
6	20000	2000	356	15	300	250
7	20000	1000	305	10	200	250
8	500	60	167	10	100	50
9	20000	1000	83	25	300	150
10	10000	1200	97	10	150	150
11	20000	2000	279	15	300	250
12	500	60	272	10	100	50
13	20000	1000	247	25	300	150
14	10000	900	356	10	150	150
15	20000	1200	305	10	200	250
16	500	60	167	10	100	50
17	20000	1000	83	25	300	150
18	10000	700	97	10	150	150
19	600	800	279	15	200	200
21	10000	1000	97	10	150	150
22	20000	1500	279	15	300	250
23	20000	1800	272	15	300	250
24	20000	1000	247	10	200	250
...	...	...	...	...	...	...
...	...	...	...	...	...	...
998	5500	450	313	10	100	50
999	8000	750	182	15	200	100
1000	10000	1000	405	10	150	150

From the table above, it can be seen that the variables "land area, rainfall, pests, urea and phonska" are independent variables (predictors), while the variable "harvest yield" is the dependent variable (response), using a total of 1,000 research data records. In the context of this study, independent variables or predictors are factors that are expected to influence rice harvest yields. Land area, for example, represents the available physical resources, while rainfall is an important environmental factor for plant growth. The presence of pests is a challenge that can disrupt production, while the use of fertilizers such as urea and phonska is an effort to increase soil fertility and, ultimately, crop yields. By using 1,000 data records, this study has a large enough database to train a multiple linear regression model, so it is expected to provide more accurate and representative analysis results. This dataset size also increases the validity of the predictions produced, because it covers a wider variety of agricultural conditions and related factors that influence rice production

3.4. Modelling

The selection of multiple linear regression algorithm as a predictive model for rice production estimation is a strategic step. This model is trained using data that has been processed and prepared in advance, by utilizing the Google Collaboratory platform as a development and analysis tool. In this phase, the data mining process is carried out on the dataset that has gone through the preprocessing stage, ensuring that the data is ready for further analysis. Using Google Collaboratory as a platform provides several advantages, such as access to higher computing power and the ability to share and collaborate with teams in real-time. The data

mining step at this stage is very important to ensure that important patterns in the data can be identified. This process not only prepares the data for analysis, but also helps in finding anomalies or inconsistent data, so that the multiple linear regression model can provide more accurate and reliable predictions. Data mining also allows the identification of historical trends, which can be very useful in modeling the complex relationships between environmental factors and rice production. At this stage, the dependent variable is determined as Harvest Yield (Y) and the independent variables (X) are Land Area, Rainfall, Pests, Urea, and Phonska, with the following coding :

```
The coding of data mining
# Independent variables (X1 until X5)
X = df[['Land area (X1)', ' Rainfall (X2)', ' Pest (X3)', 'Urea(X4)', 'Phonska(X5)']]
# Dependent variable (Y)
y = df['Yields (Y)']
# Splitting the dataset into training and testing data (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Initialization of linear regression model
model = LinearRegression()
# Training the model with training data
model.fit(X_train, y_train)
# Predicting target values for testing data
y_pred = model.predict(X_test)
# Viewing the regression coefficients
print("Regression Coefficient: ", model.coef_)
# Looking at the intercept (intersection point)
print("Intersep: ", model.intercept_)
```

From the above coding, the output message is displayed after the linear regression model is trained on the data set (X and Y). After the learning process is complete, the linear regression coefficients displayed are as in Table 3. Below :

Tabel 3. Coefficient Values on the specified variables

Variables	Value
Constants	-3836,2639
Land area	5,7336
Rainfall	1,2710
Pest	6,1153
Urea	1,6226
Phonska	1,2581

Based on the results of the analysis using multiple linear regression, a regression model was obtained that was able to explain the relationship between independent variables and dependent variables significantly. Multiple linear regression is used because of its ability to handle more than one independent variable, thus providing a more comprehensive view of the factors that influence the dependent variable, as follows :

$$Y = a + b1X1 + b2X2 + b3X3 + b4X4 + b5X5 + e \quad (1)$$

Become :

$$Y = (-3836{,}2639) + (5{,}7336)\ x\ Land\ Area + (1{,}2710)\ x\ Rainfall + (6{,}1153)\ x\ Pest + (1{,}6226)\ x\ Urea + (1{,}2581)\ x\ Phonska + e \quad (2)$$

From the results of this regression analysis, we can understand the role of each independent variable on rice production, with several interesting findings that can be explained further :

- Intercept (-3836.2639): This intercept shows that if all independent variables, such as land area, rainfall, number of pests, and fertilizer use, are zero, then rice production will be at a negative point of -

3836.2639. Although in a practical context this may not be realistic (because these variables are rarely zero), this intercept provides a theoretical basis for starting the calculation of the prediction model.

- Land Area Coefficient (5.7336): This coefficient shows that every additional unit of land area will increase rice production by 5.7336 units, assuming other variables are constant. This confirms the importance of land expansion in increasing productivity. Farmers can use these results to consider how to maximize the land they have in the production process.
- Rainfall Coefficient (1.2710): An increase in rainfall by one unit will increase production by 1.2710 units. This result shows that rainfall plays an important role in supporting the growth of rice plants, although its influence is not as large as other variables such as pests. However, farmers must still pay attention to the ideal distribution and intensity of rainfall in order to maximize production results.Land Area Coefficient (5.7336): This coefficient shows that every additional unit of land area will increase rice production by 5.7336 units, assuming other variables are constant. This confirms the importance of land expansion in increasing productivity. Farmers can use these results to consider how to maximize the land they have in the production process.
- Pest Coefficient (6.1153): This high coefficient indicates that increasing the number of pests will actually increase rice production significantly, by 6.1153 units. This finding may seem contradictory at first glance, because pests are usually considered detrimental. It is possible that the data or model used has anomalies, or there may also be an indirect relationship between the number of pests detected and the intensity of control carried out by farmers, which ultimately increases production.
- Urea Coefficient (1.6226): Urea fertilizer plays a small but significant role in rice production. Increasing urea use by one unit will only increase production by 1.6226 units. This indicates that urea provides benefits, but its contribution may be more effective when combined with other factors such as different types of fertilizers or certain soil conditions.
- Phonska Coefficient (1.2581): The use of Phonska fertilizer, with a relatively large coefficient (1.2581), shows a fairly strong impact on rice production. Each addition of one unit of Phonska will increase production significantly, indicating that this fertilizer has a greater effect than urea, perhaps because its nutrient content is more complex and more in line with the needs of rice plants.

From this analysis, it can be concluded that agronomic factors such as land area and fertilizer use have varying impacts on rice production. Land area and Phonska fertilizer use emerged as the two factors with the most positive influence, indicating that land optimization and proper fertilizer use are important strategies in increasing yields. Meanwhile, the anomaly that emerged regarding the influence of pest numbers highlights the importance of effective pest control in rice production. Although pests are usually considered detrimental, this analysis may reflect that increasing pest numbers trigger the use of more intensive control methods, which can indirectly increase production yields. These results indicate that good land and resource management strategies are essential in a sustainable agricultural system.

Furthermore, although rainfall plays an important role, its impact is relatively small compared to other variables. This implies that in situations with extreme rainfall variations, farmers need to consider more sophisticated irrigation or water management technologies to maintain optimal production results.

3.5. Evaluation

Once we have created a prediction model for rice production, we need to check whether it is really reliable. To do this, we calculate some numbers that show how well our model works. The first number is RMSE, which shows how far our predictions are from the actual results. The smaller the number, the more accurate our predictions are. The second number is R-squared, which shows how much of the rice production our model can explain. The larger the number, the better our model is at explaining rice production.

The coding of Displaying evaluation results

```
# Count Mean Squared Error (MSE)
mse = mean_squared_error(y_test, y_pred)
# Count R-squared (R²)
r2 = r2_score(y_test, y_pred)
# Displaying evaluation results
print(f"Mean Squared Error (MSE): {mse}")
print(f"R-squared (R²): {r2}")
```

From the results of the multiple linear regression analysis carried out on the data, the following are the results of calculating the Root Mean Square Error (RMSE) value and the coefficient of determination (R2) :

- RMSE (Root Mean Square Error): The RMSE value is 17065084.9641. This shows that the average error of the model prediction compared to the actual value is about 17065084 units. The smaller the RMSE value, the better the model is in predicting the crop yield.
- Coefficient of Determination (R2): The R2 value is 0.6487. This means that about 64.87% of the variability in rice yields can be explained by the independent variables in this model (land area, rainfall, pests, use of urea fertilizer, and phonska). The remaining 33.4% of the variability can be caused by other factors not included in this model.

3.6. Deployment

The deployment stage is the final step in compiling a data mining report. This report includes information that describes the knowledge or patterns found from the data mining process. In this study, the resulting pattern involves the use of training and testing data. To measure the level of model accuracy, testing was carried out using the python programming language. From this test, new insights were obtained that multiple linear regression can be applied in predicting agricultural production results, especially in the context of this study. Google colaboratory can also display scatter plots between actual values (y_test) and predicted values (y_pred) as well as residual plots, this diagram shows how well your model predicts the actual value. The straight line (red) represents a perfect prediction, where the actual value is the same as the predicted value. with the following coding :

The coding of display scatter plot

```
plt.figure(figsize=(8,6))
plt.scatter(y_test, y_pred, color='blue')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red', linewidth=2)
plt.title('Scatter Plot: Actual vs Predicted')
plt.xlabel('Nilai Aktual (y_test)')
plt.ylabel('Nilai Prediksi (y_pred)')
plt.show()
```

which looks like Figure 2. Below:

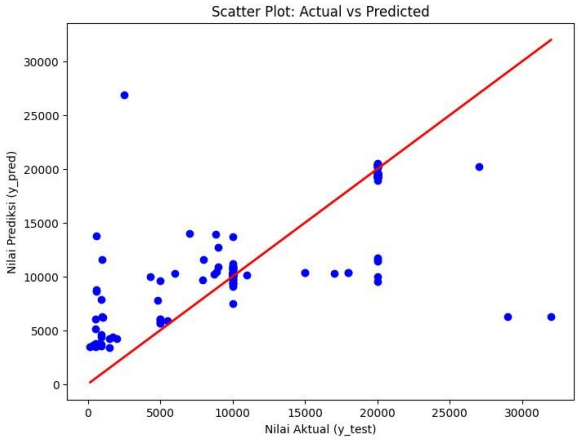

Figure 2. Scatter Plot vs Predicted

The residual plot shows the difference (residual) between the actual and predicted values. Ideally, the residuals should be randomly distributed around the zero horizontal line. If there is a clear pattern, the model may not fit the data (misfit).

The coding of display residual plot

```
# Count residual
residuals = y_test - y_pred
# Make residual plot
plt.figure(figsize=(8,6))
sns.scatterplot(x=y_pred, y=residuals, color='purple')
plt.axhline(y=0, color='red', linestyle='--')
plt.title('Residual Plot')
plt.xlabel('Nilai Prediksi (y_pred)')
plt.ylabel('Residual')
       plt.show()
```

which looks like Figure 3. Below:

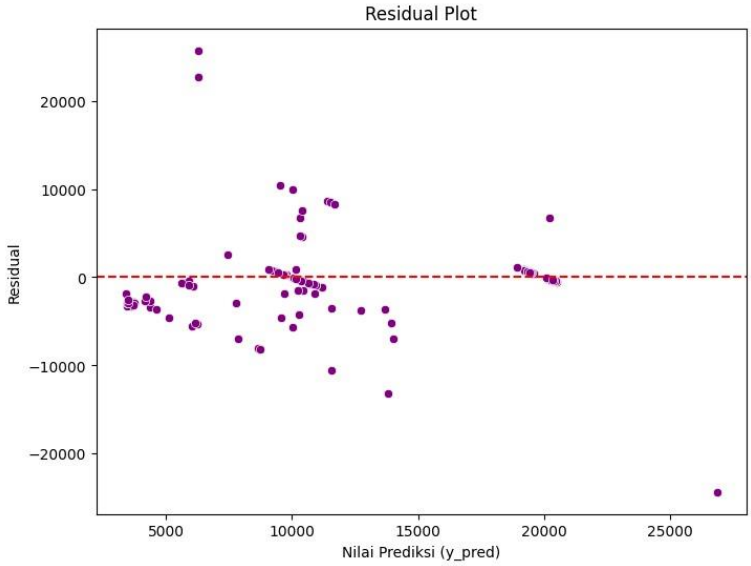

Figure 3. Residual Plot

And then apply the multiple linear regression model into a web-based application to be used in predicting future rice production and designing a strategy for using the prediction results to support better decision making, using a web-based prediction system tool. Where this system is a tool that is easy for farmers to use to find out the results of rice production predictions, which can be accessed anywhere and anytime online so that they can optimize their rice production results.

4. Conclusion

Based on the results of the multiple linear regression model, it was obtained that this model was quite significant with an R-squared of 64.87%, indicating that variables such as land area, pests, and the use of phonska fertilizer had a major impact on yields, while rainfall was not significant in this model, meaning it did not statistically affect yields while the remaining 35.13% was influenced by other factors not discussed in this study. The accuracy of the model using RMSE testing was around 17065084.9641. This multiple linear regression model has an R2 value of 0.6487, indicating that most of the variability in rice yields can be explained by the selected independent variables. Although the RMSE value of 17065084 indicates a prediction error, this value can still be considered quite good considering the complexity of the factors that affect rice production. The relatively high R2 value indicates that this model is reliable enough to be used in predicting rice production based on the variables that have been identified. However, to improve prediction accuracy, it may be necessary to add other variables or further refine the model.

References

[1] N. DWITIYANTI, N. SELVIA, and F. R. ANDRARI, "Penerapan Fuzzy C-Means Cluster dalam Pengelompokkan Provinsi Indonesia Menurut Indikator Kesejahteraan Rakyat," *Fakt. Exacta*, vol. 12, no. 3, pp. 201–209, 2019.

[2] A. Mutolib *et al.*, "Biochar from agricultural waste for soil amendment candidate under different pyrolysis temperatures," *Indones. J. Sci. Technol.*, vol. 8, no. 2, pp. 243–258, 2023.

[3] S. ARISTI, "Analisis Komoditi Unggulan dan Pertumbuhan Subsektor Tanaman Pangan di Provinsi Sumatera Selatan," *J. Agribisnis dan Sos. Ekon. Pertan.*, vol. 8, no. 1, pp. 44–49, 2022.

[4] A. M. A. K. Parewe, M. Mursalim, T. S. Putri, and H. Hermawati, "Application of Case Based Reasoning Using The K-Nearest Neighbor Algorithm in an Expert System for Diagnosing Pests and Diseases of Sugarcane Plants," *Knowbase Int. J. Knowl. Database*, vol. 2, no. 2, pp. 181–189, 2022.

[5] R. RANDIKA, M. SIDIK, and Y. PEROZA, "Analisis faktor-faktor yang mempengaruhi produksi padi sawah di desa sepang kecamatan pampangan kabupaten oki," *Soc. J. Ilmu-Ilmu Agribisnis*, vol. 10, no. 2, pp. 66–71, 2022.

[6] S. MULYATI, K. SALEH, and A. MULYANINGSIH, "Kapasitas Petani Padi Sawah Dalam Mendukung Ketahanan Pangan Keluarga Berkelanjutan di Kabupaten Pandeglang," *J. Agribisnis Terpadu*, vol. 13, no. 2, pp. 266–284, 2020.

[7] Y. A. AL-KHASSAWNEH, "A review of artificial intelligence in security and privacy: Research advances, applications, opportunities, and challenges," *Indones. J. Sci. Technol.*, vol. 8, no. 1, pp. 79–96, 2023.

[8] R. R. RACHMAWATI, "Smart Farming 4.0 Untuk Mewujudkan Pertanian Indonesia Maju, Mandiri, Dan Modern," in *Forum Penelitian Agro Ekonomi*, 2020, pp. 137–154.

[9] K. Lee and H.-Y. Park, "Development of Convergence Course of Artificial Intelligence and Psychology Applying Team Teaching Method," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 5 SE-Articles, pp. 1772–1778, Oct. 2024, doi: 10.18517/ijaseit.14.5.11491.

[10] A. N. Solihat, D. Dahlan, K. Kusnendi, B. Susetyo, and A. S. M. Al Obaidi, "Artificial intelligence (AI)-based learning media: Definition, bibliometric, classification, and issues for enhancing creative thinking in education," *ASEAN J. Sci. Eng.*, vol. 4, no. 3, pp. 349–382, 2024.

[11] A. PINANDITO, S. A. WICAKSONO, and S. H. WIJOYO, "Implementasi Machine Learning dalam Deteksi Risiko Tinggi Diabetes Melitus pada Kehamilan," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 4, pp. 739–746, 2024.

[12] E. P. Saputra, S. Nurajizah, M. Maulidah, N. Hidayati, and T. Rahman, "Komparasi Machine Learning Berbasis Pso Untuk Prediksi Tingkat Keberhasilan Belajar Berbasis E-Learning," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 2, pp. 321–328, 2023.

[13] L. WIKARSA, S. PANDELAKI, and K. SUMAJOUW, "Prediction of the Community Welfare in North Wangurer Village Using Multiple Linear Regression," *J. Pekommas*, vol. 8, no. 2, pp. 107–118, 2023.

[14] Y. Lizar, A. S. Firrizqi, A. Guci, and J. Sunadi, "Data Mining Analysis to Predict Student Skills Using Naïve Bayes Method," *Knowbase Int. J. Knowl. Database*, vol. 3, no. 2, pp. 150–159, 2023.

[15] R. J. Suhatril, R. D. Syah, M. Hermita, B. Gunawan, and W. Silfianti, "Evaluation of Machine Learning Models for Predicting Cardiovascular Disease Based on Framingham Heart Study Data," *Ilk. J. Ilm.*, vol. 16, no. 1, pp. 68–75, 2024.

[16] M. A. SHAFI, "K-means clustering analysis and multiple linear regression model on household income in Malaysia," *Int. J. Artif. Intell*, vol. 12, no. 2, pp. 731–738, 2023.

[17] A. Byna, M. M. Lakulu, I. Y. Panessai, and Nurhaeni, "Machine Learning-Based Stroke Prediction: A Critical Analysis," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 5 SE-Articles, pp. 1609–1618, Oct. 2024, doi: 10.18517/ijaseit.14.5.19527.

[18] D. I. P. DESY, T. W. QUR'ANA, and A. DHARMAWATI, "Pemodelan Spasial untuk Analisa Produksi

Padi Integrasi Machine Learning," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 14, no. 2, pp. 128–137, 2023.

[19] I. U. RAHMAWATI, M. HADDIN, and S. SUHARTONO, "Potensi Sampah Untuk Pembangkit Listrik Tenaga Sampah (PLTSa) Berbasis Metode Regresi Linier Berganda," *J. Tek. Inform. UNIKA St. Thomas*, pp. 1–8, 2023.

[20] M. ADHA, E. UTAMI, and H. HANAFI, "Prediksi Produksi Jagung Menggunakan Algoritma Apriori Dan Regresi Linear Berganda (Studi Kasus: Dinas Pertanian Kabupaten Dompu)," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 3, pp. 803–820, 2022.

[21] A. F. BOY, "Implementasi Data Mining Dalam Memprediksi Harga Crude Palm Oil (CPO) Pasar Domestik Menggunakan Algoritma Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi Sumatera Utara)," *J. Sci. Soc. Res.*, vol. 3, no. 2, pp. 78–85, 2020.

[22] F. A. Vadhil, M. L. Salihi, and M. F. Nanne, "Machine learning-based intrusion detection system for detecting web attacks," *IAES Int. J. Artif. Intell.*, vol. 13, no. 1, pp. 711–721, 2024.

[23] N. Aini, S. A. Wicaksono, and I. Arwani, "Pembangunan Sistem Informasi Perpustakaan Berbasis Web menggunakan Metode Rapid Application Development (RAD)(Studi pada: SMK Negeri 11 Malang)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. e-ISSN*, vol. 2548, p. 964X, 2019.

[24] A. BAHTIAR, "Prediksi Hasil Panen Padi Tahun 2023 menggunakan Metode Regresi Linier di Kabupaten Indramayu," *J. Inform. Terpadu*, vol. 9, no. 1, pp. 18–23, 2023.

[25] D. Wulandari and R. Rumini, "Pemodelan dan Prediksi Produksi Padi Menggunakan Regresi Linear," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 12, no. 4, pp. 1011–1019, 2023.

[26] H. PUTRA and N. U. WALMI, "Penerapan Prediksi Produksi Padi Menggunakan Artificial Neural Network Algoritma Backpropagation," *J. Nas. Teknol. dan Sist. Inf*, vol. 6, no. 2, pp. 100–107, 2020.

[27] C. SUSANTO, T. TAUFIQ, E. HASMIN, and K. ARYASA, "Sistem Pakar Prediksi Penyakit Diabetes Menggunakan Metode K-NN Berbasis Android," *CogITo Smart J.*, vol. 8, no. 2, pp. 359–370, 2022.

[28] Y. I. Sulistya *et al.*, "Prediction and Analysis of Rice Production and Yields Using Ensemble Learning Techniques," *Ilk. J. Ilm.*, vol. 16, no. 2, pp. 115–124, 2024.

[29] D. Setiadi and R. Syahri, "Penerapan Algoritma Naïve Bayes Pada Sistem Prediksi Pengguna Narkoba di Kota Pagar Alam," *JUTIM (Jurnal Tek. Inform. Musirawas)*, vol. 7, no. 1, pp. 1–10, 2022.

[30] B. Triandi, S. Efendi, and H. Mawengkang, "Regression-based Analytical Approach for Speech Emotion Prediction based on Multivariate Additive Regression Spline (MARS).," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 13, no. 6, 2023.

[31] B. A. K. A. NUGROHO and E. Nurfarida, "Prediksi Waktu Kedatangan Pelanggan Servis Kendaraan Bermotor Berdasarkan Data Historis menggunakan Support Vector Machine," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 7, no. 1, pp. 25–30, 2021.

[32] N. NAZERIANDY, Y. SYAHRA, and M. SYAIFUDIN, "Penerapan Data Mining Untuk Memprediksi Penggunaan Daya Listrik Pada PT. PLN (Persero) Rayon Medan Selatan Dengan Menggunakan Metode Regresi Linier Berganda," *J. SAINTIKOM (Jurnal Sains Manaj. Inform. dan Komputer)*, vol. 20, no. 1, pp. 20–27, 2021.

[33] D. Setiadi, Y. I. Mukti, and Y. Widiastiwi, "Decision Support System to Optimize E-tourism in Pagar Alam City," in *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)*, IEEE, 2023, pp. 149–154.

[34] D. Setiadi and Y. I. Mukti, "Electronic Tourism Using Decision Support Systems to Optimize the Trips," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, no. 1, pp. 183–200, 2023.

[35] N. AMALIA and O. P. RACHMAN, "Pengembangan Sistem Informasi Pertanian Berbasis Kecerdasan Buatan (E-Tandur) Dalam Menunjang Pertumbuhan Pertanian Masyarakat Daerah Kabupaten Bandung Dengan Metode Geographic Information System (Gis) Dan Internet Of Things (IOT)," *J. Inform. dan Rekayasa Elektron.*, vol. 5, no. 1, pp. 121–130, 2022.

[36] J. A. SOLANO, D. J. L. CUESTA, S. F. U. IBÁÑEZ, and J. R. CORONADO-HERNÁNDEZ, "Predictive

models assessment based on CRISP-DM methodology for students performance in Colombia-Saber 11 Test," *Procedia Comput. Sci.*, vol. 198, pp. 512–517, 2022.

[37] A. M. M. FATTAH, A. VOUTAMA, N. HERYANA, and N. SULISTIYOWATI, "Pengembangan Model Machine Learning Regresi sebagai Web Service untuk Prediksi Harga Pembelian Mobil dengan Metode CRISP-DM," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 5, pp. 1669–1678, 2022.

[38] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021.

[39] D. Setiadi, S. Sasmita, and M. Yolanda, "Penerapan Algoritma Regresi Linier Berganda Untuk Memprediksi Hasil panen Padi Di Kota Pagar Alam," *Kesatria J. Penerapan Sist. Inf. (Komputer dan Manajemen)*, vol. 5, no. 2, pp. 337–438, 2024.

[40] D. F. Salsabillah, D. E. Ratnawati, and N. Y. Setiawan, "Analisis Sentimen Ulasan Rumah Makan Menggunakan Perbandingan Algoritma Support Vector Machine dengan Naive bayes (Studi Kasus: Ayam Goreng Nelongso Cabang Singosari, Malang)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 107–116, 2024.

[41] B. HUDA *et al.*, "Analisis Sentimen E-Learning X Terhadap Antarmuka Pengguna Menggunakan Kombinasi Multinomial Naive Bayes Dan Pendekatan Design Thinking," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 895–902, 2024.