PADDY WETLAND PRODUCTIVITY ANALYSIS WITH LINEAR REGRESSION OF MACHINE LEARNING APPROACH

Bayu Nugraha¹, Agustina Hotma Uli Tumanggor², Mambang³, Finki Dona Marleny⁴

¹Information Systems, Faculty of Science and Technology, Sari Mulia University
²Industrial Engineering, Faculty of Science and Technology, Sari Mulia University
³Information Technology, Faculty of Science and Technology, Sari Mulia University
⁴Informatics, Faculty of Engineering, University of Muhammadiyah Banjarmasin

Corresponding Author: Bayu Nugraha

ejl. gurnawang@unism.ac.id, agustina.hotma@gmail.com, mambang@unism.ac.id, finkidona@umbjm.ac.id

ABSTRACT

Paddy is one of the priority crops in agricultural production. South Kalimantan is an area that produces Paddy. In paddy productivity in the southern Kalimantan region, there are paddy wetlands and paddy dryland. The need for paddy production in the southern Kalimantan region can increase or decrease every year. The method used in this study is a linear regression algorithm with a machine learning approach. Linear regression analysis basically predicts a variable's value based on its free variables. Linear regression only predicts variables whose data nature is intervals or ratios. Linear regression analysis can be used to examine the relationship between two or more variables. Linear regression can also make additional assumptions between variables through the most suitable lines of straight-line data points. This study is to determine the relationship between harvest area and productivity. As a result of trials using the machine learning approach, linear regression algorithms show a relationship between harvest and production area. The correlation test results can find relationships between data points so that linear regression can be used to predict. From the relationship between harvest area and productivity, a prediction accuracy of 95% was obtained.

Keywords: Paddy, Wetland, Productivity, Linear Regression, Machine Learning

I. INTRODUCTION

Most of the area in southern Borneo is composed of wetlands[1], [2]. Paddy is one of the priority crops in agricultural production[3]. The need for paddy production in the southern Kalimantan region may increase or decrease every year[1]. The increase or decrease in paddy productivity fields in an area can be due to several other supporting factors, including geographical location and climate[4]–[6]. Geographically, the southern Kalimantan region can support an increase in paddy productivity. The development of the area of harvest, production, and productivity of paddy rice for four years (2018-2021) can be seen in Table 1.

<table>
<thead>
<tr>
<th>Year</th>
<th>Harvested Area (ha)</th>
<th>Productivity (qa/ha)</th>
<th>Production (ton)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018</td>
<td>323,091.21</td>
<td>41.09</td>
<td>1,327,492.41</td>
</tr>
<tr>
<td>2019</td>
<td>356,245.95</td>
<td>37.70</td>
<td>1,342,861.82</td>
</tr>
<tr>
<td>2020</td>
<td>292,026.78</td>
<td>38.85</td>
<td>1,134,450.21</td>
</tr>
<tr>
<td>2021</td>
<td>255,760.43</td>
<td>39.69</td>
<td>1,041,862.91</td>
</tr>
</tbody>
</table>

Paddy harvested area in South Kalimantan in 2021 was recorded at 255,760.43 hectares. The largest harvested land is in Barito Kuala Regency, which covers an area of 62,815.09 hectares, or 24.56 percent of the harvest area of South Kalimantan's paddy fields. Meanwhile, the smallest paddy field harvested area is in Banjarbaru City, which is 1,048.91 hectares, or only 0.41 percent of the South Kalimantan paddy field harvest area. Paddy production in South Kalimantan in 2021 was recorded at 1,041,862.91 tons. This figure decreased 9.42 percent compared to 2020 paddy production. The decline in production was in line with a decrease in the harvested area of 11.75 percent[1], [2].

The data obtained above will be analyzed using a machine-learning approach. The method used is the linear regression method. The purpose of this rice wetland productivity analysis is to determine the relationship between variables. Where the approach with machine learning can be whether it can answer the linear relationship of the
relationship of bound variables or free variables, from this, it can be seen whether the method of linear regression can predict.

The method used in this study is a linear regression algorithm with a machine learning approach [7]. This linear regression can make additional assumptions so that data between variables can be correlated through the most precise straight line of the data straight line point [8]. This method can make understanding the relationship between one variable and another easier. Analysis regression was used to examine the relationship between two or more variables in this study to determine the relationship between harvested areas and rice productivity.

II. LITERATURE REVIEW

A. Wetland

Wetlands are lands that are dominated by lands that have more water content. Wetlands can also be used as an agricultural land where agricultural land has the characteristic of being separated by a small embankment to hold water. These wetlands where the main crop can usually be wetland rice fields. Land includes land used as rice fields, both of which are planted with rice, secondary crops, or other seasonal crops[1], [2].

B. Machine Learning

Machine learning is a technique that can be studied on the topic or part of artificial intelligence[9]. Machine learning was developed to be able to learn on its own without direction from the user. Machine learning has several techniques, namely deep learning and supervised and unsupervised learning[10],[11]. Supervised learning techniques are techniques that can receive information that already exists in the data by giving certain labels[12]. Unsupervised learning techniques are techniques applied to machine learning that are used in data that do not have information that can be applied directly. Machine learning has a way of working starting from the stages of data collection, which then the data is explored so that it can be visually seen the data content to make it easier to determine models or techniques that follow data characteristics. The choice of models or techniques from the data exploration process can make better decisions, and then the data can be divided into training data and testing data. Data that has passed modelling can be evaluated to determine whether the data is accurate and follows the selection of machine learning techniques [7].

C. Linear Regression

Linear regression is the simplest algorithm to take a closer look at how the machine learning process is going. Linear regression can be given several predictor variables and continuous response variables[7]. These variables are used to find relationships between other variables that make it possible to predict an outcome[8]. Here is the formula for determining the prediction line:

\[ y = \beta_0 + \beta_1x + u \]  

(1)

The values used for \textit{training} are \( \beta_0 \) and \( \beta_1 \), i.e., the value that affects the position of the line Because \( u \) is not considered and other variables are only \( x \) (input) and \( y \) (output). So that the value of \( \beta_1 \) and \( \beta_0 \) is the weight of the prediction function.

D. Cost Function

To evaluate the quality of the model, many ways can be done, one of which is to use the cost function. Cost loss serves as a tool for measuring the performance of machine learning models. So if there is something unreasonable performance, then this function will give an idea[13]. Alright, so the equation for the cost function is:

\[ J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_{\theta}(x^{(i)}) - y^{(i)})^2 \]  

(2)

E. Gradient Descent

Gradient descent is an algorithm that can optimize to find global and local minimums from differentiated functions. The gradient reduction carried out in this study is an optimization algorithm that aims to adjust the parameters to minimize the cost function[14], [15]. The main update steps for gradient drops are:

\[ \theta_j := \theta_j - a \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (\text{for } j = 0 \text{ and } j = 1 \]  

(3)
So we multiply the derivative of the cost function by the learning rate (\( \alpha \)) and subtract it from the present value of the parameter (\( \theta \)) to get the newly updated parameter (\( \theta \)). The return function is our parameter vector containing the \( \theta \)s values for the hypothesis and is a list containing the cost[13]. These variables are not very important but help to analyze the model better.

### III. RESEARCH METHODOLOGY

#### A. Proposed Method

The method used is a machine-learning approach to linear regression lysis. The proposed method can be seen in figure 1. The stages of the proposed method start from data preparation, and then the data is cleaned so that it can be analyzed according to the machine learning technique of linear regression algorithms. After the data is collected, it can be done through Exploratory data analysis (EDA). Next, enter feature engineering and modelling techniques with linear regression and evaluation measurements to determine the correlation of relationships between variables. Here is an overview of the proposed method.

![Figure 1. Proposed Model](image)

#### B. Data preparation

During the data preparation, information from the data to be explored is needed. Data in the collection from relevant agencies on agricultural production data in south Kalimantan. The data used is data for five years, starting from 2017-2022. Here is the metadata information table:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regency_city</td>
<td>Polynomial</td>
</tr>
<tr>
<td>year</td>
<td>Date</td>
</tr>
<tr>
<td>Harvested Area</td>
<td>real</td>
</tr>
<tr>
<td>Production</td>
<td>real</td>
</tr>
<tr>
<td>Productivity</td>
<td>real</td>
</tr>
</tbody>
</table>

#### C. Exploratory Data Analysis (EDA)

At the EDA stage, it uses the describe() method to return a description of the data in the DataFrame. EDA has a visual form that can contain count information. Visually visible information that serves to see sum values that have no fill or are empty. From the EDA process, you can see the average result of the data, standard deviation, and fundamental data relationships, looking at the maximum or minimum data. Here is a visual overview of the EDA process.
The pairwise relationship plot in a paddy productivity dataset shows a distribution that still leans towards harvested area and production. From the productivity of the results, it seems that it is appropriate.

IV. RESULT AND DISCUSSION

The result variables depicted in figure 3 express a linear relationship. From these results, you can intuitively guess the percentage of harvested area based on productivity. These results can trace the line between the points reading the harvested area value if it can trace the vertical line of the given productivity value. Judging from the results of the regression plot, productivity and harvested areas are less able to find relationships between variables.

Regression analysis is used to examine the relationship between two or more variables in figure 4, which is to determine the relationship between harvest and production area. From the figure, there is a distribution of data that can find the relationship of scattered data between data points and describe linear regression lines. From the regression line, the thing that can be known is how the relationship between the value of the x-axis and the value of the y-axis. Linear regression R can be used to predict when there is a relationship between the values of the axis.
To evaluate the quality of the model can use the cost function. A gradient descent optimization algorithm that aims to adjust the parameter to minimize cost functions. The following are the differences in results using epochs of 1000 with 15000. The following image Fit a Linear Regression Model with Gradient Descent. In the following pictured results, displacement costs decrease as the iteration increases.

The next process is to use a correlation test. This correlation test process uses the Pearson Product Moment Test. This is done to find out the degree of proximity of the data relationship between 2 variables that scale the interval or ratio. This correlation test will return the value of the correlation coefficient, whose value ranges between -1, 0, and 1. A value of -1 means that there is a perfect negative correlation, 0 means that there is no correlation, and a value of 1 means that there is a perfect positive correlation. Here are the stages of the Pearson correlation test.
Machine learning outcomes of linear regression algorithms to produce structure models with good performance evaluation results can be seen in cost functions that decrease as iterations increase. We get a prediction accuracy of 95%. From the results of trials using a machine learning approach, a linear regression algorithm can be stated that there is a relationship between harvest and production area. The correlation test results can find the relationship between the data points and draw a linear regression line. This study concludes that linear regression can be used to predict. The suggestion for further research is that it can be continued with the mapping of agricultural land in the southern Kalimantan province by applying the best model of linear regression. This research can be developed by other methods or incorporating several methods in predicting paddy productivity. Paddy's productivity can be expanded to add variables or parameters to its data so that it can enrich the analysis and knowledge section.

V. CONCLUSION

REFERENCES