VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i2.5882

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

# IMPLEMENTATION OF MULTIPLE LINEAR REGRESSION ALGORITHM IN PREDICTING RED CHILI PRICES IN GARUT REGENCY

**Yoga Handoko Agustin[1*]; Fitri Nuraeni[2]; Rika Lestari[3]**

Informatics[1,2,3]
Institut Teknologi Garut, Indonesia[1,2,3]
www.itg.ac.id[1,2,3]
yoga.handoko@itg.ac.id[1*], fitri.nuraeni@itg.ac.id[2], 2006173@itg.ac.id[3]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)



***Abstract***— *Vegetables, including red chili peppers, play an important role in food and economic balance. Significant price fluctuations and inflation are often problems for farmers and traders. Garut Regency, as the center of red chili production in West Java, faces similar challenges. This research aims to implement a Multiple Linear Regression algorithm to predict the price of red chili peppers in the Garut Regency, highlighting the novelty of using a combination of One Hot Encoding, Feature Engineering, Standard Scaler, and Hyperparameter Tuning techniques. The method used is CRISP-DM with 6 stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The data used is the price and production of red chili peppers per week in 2018-2023, with a total of 702 records. This research involved 8 trials with data transformation and normalization scenarios. The model evaluation used MSE, RMSE, MAPE, R-squared, and statistical hypothesis testing metrics. Results showed 5 significantly influential attributes: year, month, production, net harvested area, and productivity. The best model yielded MSE 202,134,650, RMSE 14,217, MAPE 29.16%, and R-squared 0.320. This approach is simpler yet effective and is able to provide fairly accurate predictions. This research is expected to contribute to providing predictive models that help farmers and traders anticipate price fluctuations, as well as provide insights for policymakers in price management.*

***Keywords***: *CRISP-DM, multiple linear regression, prediction, red chili.*

***Intisari—*** *Sayuran, termasuk cabai merah, memiliki peran penting dalam keseimbangan pangan dan ekonomi. Fluktuasi harga yang signifikan sering menjadi masalah bagi petani, pedagang, dan inflasi. Kabupaten Garut, sebagai sentral produksi cabai rawit merah di Jawa Barat, menghadapi tantangan serupa. Penelitian ini bertujuan mengimplementasikan algoritma Regresi Linear Berganda untuk memprediksi harga cabai merah di Kabupaten Garut, dengan menonjolkan kebaruan dalam penggunaan kombinasi teknik One Hot Encoding, Feature Engineering, Standard Scaler, dan Hyperparameter Tuning. Metode yang digunakan adalah CRISP-DM dengan 6 tahapan: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Data yang digunakan adalah harga dan produksi cabai merah per minggu tahun 2018-2023, dengan total 702 records. Penelitian ini melibatkan 8 ujicoba dengan skenario transformasi dan normalisasi data. Evaluasi model menggunakan metrik MSE, RMSE, MAPE, R-squared, dan uji hipotesis statistik. Hasil menunjukkan 5 atribut yang berpengaruh signifikan: tahun, bulan, produksi, luas panen bersih, dan produktivitas. Model terbaik menghasilkan MSE 202,134,650, RMSE 14,217, MAPE 29.16%, dan R-squared 0.320. Pendekatan ini lebih sederhana namun efektif, serta mampu memberikan prediksi yang cukup akurat. Penelitian ini diharapkan dapat berkontribusi dalam menyediakan model prediktif yang membantu petani dan pedagang mengantisipasi fluktuasi harga, serta memberi wawasan bagi pembuat kebijakan dalam pengelolaan harga.*

***Kata Kunci***: *CRISP-DM, regresi linier berganda, prediksi, cabai merah.*



Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5882

## INTRODUCTION

Chili is one type of horticultural commodity vegetable crop that is cultivated, developed, and consumed by the general public for their daily needs. Therefore, the government is very concerned about chili peppers as a horticultural commodity because there is no other vegetable commodity as a substitute[1].

Based on data from the statistic [2] In 2018, cayenne pepper consumption reached 483,650 tons and increased the following year to 513,170 tons in 2019. However, there was a significant decline in 2020, with the amount of consumption decreasing to 479,030 tons. The good news is that consumption gradually increased again in 2021, reaching a total of 528,140 tons of cayenne pepper. In 2022, cayenne pepper consumption in Indonesia jumped to 569,650 tons, an increase of 41,510 tons or about 7.86% compared to the previous year. This figure shows that the consumption of cayenne pepper reached the highest level in the last five years.

Garut Regency is one of the red chili production centers in West Java with a production output of 462,060 quintals in 2020 and an increase to 469,454 quintals in 2021[3]. However, a common problem with this product is the frequent price fluctuations. The frequent sharp rise and fall of chili prices makes it one of the commodities that contribute to inflation every year [1].

Some research that has been done before, such as the first research by [4]. in predicting the price of cayenne pepper using a simple linear regression algorithm. The study obtained prediction modeling results of an error rate of 24.00% or an accuracy rate of 76.00%. The second research conducted by [5] about the prediction of red chili prices using artificial neural networks. The results obtained in this study are that the price of red chili in Batam City always increases at the beginning and end of the year, especially when rainfall affects it. The third research by [6] Regarding predictions using the ARIMA model, the results of predicting the price of red cayenne pepper as a community food need in June 2022 to May 2023 in the Morning Market have a more expensive selling price and the selling price in the Development Market tends to be cheaper. In the ARIMA (3,1,4) model, the smallest MAPE and RMSE values are 20.73% and the ARIMA (3,1,3) model is 20.83%. Further research by [7], about the prediction of red and green cayenne pepper prices using the Lee model fuzzy time series algorithm. The results show that forecasting the price of red cayenne pepper with the MAPE value category is good and forecasting the price of green cayenne pepper with the MAPE value category is very good. The last research by [8] In research using the Long-Short Term Memory method, the prediction of chili prices, whose daily data from 2021-2022 often fluctuates, is discussed. The price of chili peppers in May 2022 was 0.51%, rising to 0.61% in June 2022. The results obtained from January 1, 2021, to July 31, 2022, obtained the smallest MSE result of 0.0155 with a proportion of 70% training data and 30% testing.

Although various methods such as ARIMA, fuzzy time series, and LSTM have been applied in predicting chili prices, these studies have several limitations. For example, Moha Lalapa & Yunus [4] used only one evaluation metric, MAPE, which may not provide a comprehensive picture of the model's performance. Then, some previous studies used less varied data or only used a limited range of data, so the prediction results may be less accurate. In addition, most studies have not considered external factors such as the amount of production that can significantly affect the price of red chili peppers.

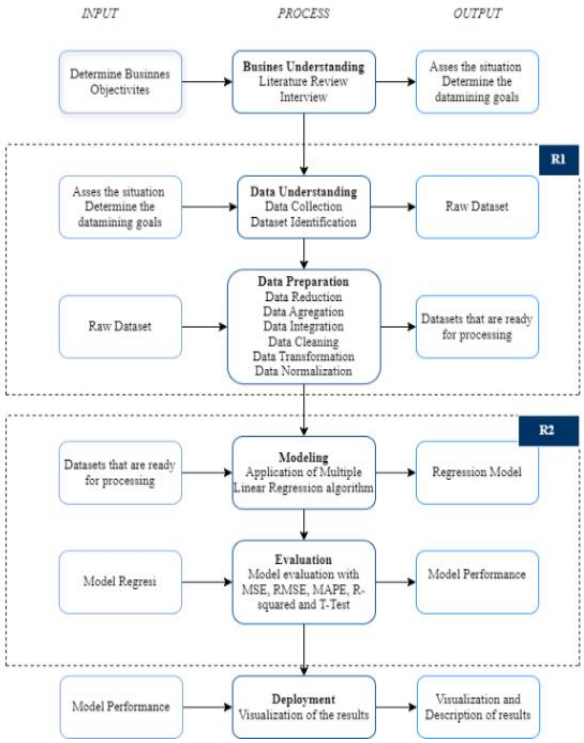
Based on the reference journals above, this research will use a multiple linear regression approach to predict the price of red chili peppers in Garut Regency with some novelty from previous research. Linear Regression is an approach to modeling the relationship between the dependent variable Y and one or more independent variables called X [9]. Based on research by Moha Lalapa & Yunus [4], linear regression is proven to be able to provide 76% accuracy in predicting the price of cayenne pepper, so it is considered effective for predicting agricultural commodities. This research seeks to implement multiple linear regression with improvements in the use of more varied data and the addition of external factors such as red chili production; as well as the application of several evaluation metrics, namely MSE, RMSE, MAPE, and $R^2$, to produce a more comprehensive model.

Hopefully, this research can make a positive contribution to the understanding and management of red chili production and sales in Garut Regency. By using the Linear Regression algorithm, it is expected that farmers and traders can obtain more accurate price predictions, help them plan better strategies, and reduce the impact of price fluctuations on overall inflation.

## MATERIALS AND METHODS

This research uses a framework based on the CRISP-DM method, which describes the stages of data collection and analysis to build models that support decision-making. CRISP-DM allows the creation of data mining models tailored to specific

needs [10] According to [11] there are 6 stages in CRISP-DM. These stages will be further explained in Figure 1.




Source: (Research Result, 2024)

Figure 1. Research Framework

### A. Business Understanding

This stage is the stage of understanding the data by collecting data. In this stage, the required data collection is carried out from the Garut Regency Industry, Trade and ESDM Office (Disperindag ESDM), namely red chili price data from 2018 - 2023 as many as 2,664 records and red chili production data obtained from the Garut Regency Agriculture Office (DISTAN) as many as 108 records. the data collected is data from the Department of Industry, Trade and ESDM (Disperindag ESDM), then identify the dataset by identifying the types of attributes that will be included in the dataset. The type of dataset attribute can be numeric data or nominal data. The output of this stage is raw dataset, which is initial data that has not undergone further processing.

### B. Data Understanding

This stage is the stage of understanding the data by collecting data, the data collected is data from the Department of Industry, Trade and ESDM (DISPERINDAG ESDM), then identifying the dataset by identifying the types of attributes that will be included in the dataset. The type of dataset attributes can be numeric data or nominal data. The output of this stage is the raw dataset, which is the initial data that has not undergone further processing.

### C. Data Preparation

In this data processing stage, there are 6 stages of data processing carried out, namely data reduction which includes determining the attributes that will be used in the analysis. Next is data integration which involves combining all the necessary datasets into one table [12]. Next is data aggregation which is the stage of converting data from a higher aggregate level to a lower level by dividing the data into smaller or more detailed parts. The next stage is data cleansing, which is a process to detect, correct, or delete erroneous or inaccurate records, tables, and databases [13]. The data transformation is the stage where the data is changed to fit the model or algorithm used in the data processing process [14]. this stage involves the process of converting category variables into a numerical format that the modeling algorithm can understand. The last stage is data normalization, this stage is one of the steps taken in the data pre-processing stage. This process involves re-scaling the data values to facilitate further processing [15].

### D. Modelling

At this stage, the appropriate modeling technique is determined to optimize the results. The modeling technique in this study uses the Multiple Linear Regression Algorithm. So that the output produced later is a regression model.

### E. Evaluation

At this stage, an evaluation is carried out to ascertain whether the model results are by the objectives set in the initial phase. This process involves testing the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), R-squared ($R^2$), and Statistical Hypothesis Test (T-test) with the resulting output being model performance.

### F. Deployment

In the dissemination stage, the results of the data mining process are presented in a format that can be easily understood by stakeholders using Google Colab tools. At this stage, the outputs are visualizations and descriptions that explain the model's performance.

## RESULTS AND DISCUSSION

The results of this research use the Cross-Industry Standard Process for Data Mining (CRISP-



Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5882

DM) method. All stages in this research will follow the CRISP-DM methodology flow. These stages will be further explained as follows:

**A. Business Understanding**

At the stage of understanding the business, there are 2 activities, the first is conducting a literature review, from the results of the literature review it was found that several studies using the Linear Regression algorithm have proven to produce predictions with a fairly good level of accuracy in predicting the price of red cayenne pepper, as seen in the research journal by [4] shows that this algorithm is able to provide a good level of accuracy, amounting to 76.00% so this algorithm is appropriate for use in making predictions.

Furthermore, the interview was conducted with the Department of Industry, Trade, and ESDM (Disperindag ESDM). From these interviews, it is known that the price of red chili peppers often experiences significant fluctuations. In addition, Disperindag ESDM does not have an adequate prediction system and does not monitor factors that affect price increases. Therefore, a prediction system is needed to help manage and anticipate changes in the price of red chili peppers.

**B. Data Understanding**

Data Understanding is the process of understanding data. The activities carried out are data collection and dataset identification. The explanation is as follows:

**1. Data Collection**

The data collection process in this study was carried out through two main sources. First, red chili price data is obtained from the Garut Regency Industry, Trade, Energy and Mineral Resources Agency (Disperindag ESDM), which includes daily price data from 2018 to 2023 with a total of 2,664 records. Second, red chili production data was taken from the Garut District Agriculture Office (DISTAN) with a total of 108 records for 2021 to 2023. Data collection is done by utilizing secondary data that has been recorded in official reports from the two related agencies. After the data was obtained, it was found that the production data for curly red chili was only available from 2021 to 2023, so the price data used for this commodity was also limited to the same period. Meanwhile, the price data for cayenne pepper was used from 2018 to 2023. No significant obstacles were reported during the data collection process, but the limited production data in certain periods affected the time span used in this study. And here is a sample of the raw data shown in Table 1.

Table 1. Data Sample

| No | Commodity | Production (Ton) Target | Jan | ... | June |
|---|---|---|---|---|---|
| 1 | Red cayenne pepper | 19.866 | 3.433 | ... | 3.818 |
| 2 | Red cayenne pepper | 19.866 | 4.251 | ... | 6.000 |
| 3 | Red cayenne pepper | 20.050 | 2.582 | ... | 3.301 |
| 4 | Red cayenne pepper | 20.050 | 3.418 | ... | 2.856 |
| ... | ... | | ... | ... | ... |
| 12 | Curly red chili | 3.100 | 241 | ... | 143 |

Source: (Research Results, 2024)

**Dataset Identification**

At this stage, the type of attributes that will be included in the dataset is selected or identified. The type of dataset attribute can be numeric data or nominal data. Here is a sample of attributes used in this study shown in Table 2.

Table 2. Sample Data Identification on Production

| No | Attributes | Data Type | ... | Description |
|---|---|---|---|---|
| 1 | No | *Integer* | ... | Sequence number of the commodity list |
| 2 | Commodities | *String* | ... | Name of commodity type |
| 3 | Target | *Integer* | ... | Commodity production goals or targets for the current year in specific units (e.g. hectares) |
| 4 | Realization | *Integer* | ... | Commodity production realization in the previous year in certain units |
| ... | ... | ... | ... | ... |
| 9 | %(11/4) | *Float* | ... | Percentage of total harvest realization compared to the previous year's realization. |

Source: (Research Results, 2024)

**C. Data Preparation**

The data processing stage involves the process of transforming the raw dataset that has been obtained into data that is ready to be analyzed or modeled. There are 5 activities at this stage, which are as follows:

**1. Data Reduction**

The next stage in data processing is data reduction, which is reducing, reducing, or selecting attributes that will be used in this study. Attributes that are carried out in the reduction stage are price, production, harvest area, and productivity data attributes. The following is a sample of data that has

been reduced on the production attribute shown in Figure 2.

| No | Atribut Sebelum *Data Reduction* |
|---|---|
| 1 | No |
| 2 | Komoditi |
| 3 | Sasaran |
| 4 | Realisasi |
| 5 | Total |
| 6 | Realisasi Produksi (Ton) |
| 7 | Jumlah |
| 8 | % (11/3) |
| 9 | % (11/4) |

| No | Atribut Setelah *Data Reduction* |
|---|---|
| 1 | Realisasi Produksi (Ton) |

Source: (Research Results, 2024)

Figure 2. Illustration of Data Reduction Process in Production

Based on the illustration in Figure 2, only the Production Realization (Ton) attribute is used because production realization provides information about the amount of red chili production that affects the price in the market. Production data is critical to understanding the relationship between supply and price.

## 2. Data Aggregation

In this modeling analysis, data aggregation was carried out on weekly price data to match the monthly production, harvest area, and productivity attributes to ensure alignment and consistency in the analysis. The results of the aggregation of production data are presented in Table 3.

Tabel 3. Data *Aggregation*

| Year | Month | Average Price per Month (IDR) |
|---|---|---|
| 2018 | January | 38.333 |
| 2018 | February | 33.000 |
| 2018 | March | 49.400 |
| 2018 | April | 28.750 |
| ... | ... | ... |
| 2023 | December | 92.500 |

Source: (Research Results, 2024)

## 3. Data Integration

After the data reduction and data aggregation stages, then the data integration stage is carried out, namely combining data from various different sources. The data used in this study are different sources, so data merging must be done. Here are the results of the data that has been combined shown in Table 4.

Table 4. Data Integration Results

| No | Year | Month | Market | ... | Price |
|---|---|---|---|---|---|
| 1 | 2018 | January | Guntur | ... | 38.333 |
| 2 | 2018 | January | Leles | ... | 30.000 |
| 3 | 2018 | January | Kadungora | ... | 45.000 |
| 4 | 2018 | January | Limbangan | ... | 24.333 |
| ... | ... | ... | ... | ... | ... |
| 702 | 2023 | December | Malangbong | | 77.500 |

Source: (Research Results, 2024)

## 4. Data Cleaning

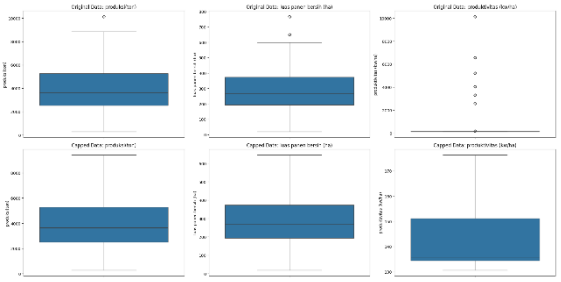
The data cleaning stage is the process of removing or replacing invalid, empty, or duplicate data. At this stage there are 2 activities carried out, namely changing the data type that should be numeric, changing the data type that should be numeric is done because the data that should be numeric is stored in string format. The following are the results of data that has been converted to numeric data type, shown in Table 5.

Table 5. Results of Data That Has Been Converted to Numeric

| Production (Ton) | Net Harvested Area (Ha) | Productivity (Kw/Ha) | Price (Rp) |
|---|---|---|---|
| 286 | 19 | 150.26 | 41250 |
| 286 | 19 | 150.26 | 48750 |
| 286 | 19 | 150.26 | 58750 |
| 286 | 19 | 150.26 | 53750 |

Source: (Research Results, 2024)

In addition to changing the data type to numeric in this data cleaning stage, outliers are also handled by capping using IQR. Handling outliers is done to maintain model stability and accuracy. The following are the results of handling outliers using IQR shown in Figure 3.



Source: (Research Results, 2024)

Figure 3. Comparison before and after handling outliers

## 5. Data Transformation

The data transformation stage includes various steps to convert raw data into a format suitable for analysis. These steps include encoding, using label encoding and one-hot encoding techniques, to convert category variables into a numerical format that can be understood by the modeling algorithm. In addition, this stage also involves Feature Engineering, which is the creation of new features from existing data to improve model performance. The steps of the data transformation stage are as follows:

### *Encoding Data*

In the Data Encoding stage, the two techniques used are Label Encoding and One hot encoding. Label Encoding converts category values into numbers, which makes it easier for some algorithms



Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5882

to process the data. One hot encoding creates each category as a separate binary feature, helping the model capture independent relationships between categories. One hot encoding is more suitable for features that have no order or scale. The following are the transformation results using label encoding and One hot encoding shown in Tables 6 and 7.

Table 6. Label Encoding Transformation

| No | Market Name Attribute | |
|---|---|---|
| | Before | After |
| 1 | guntur | 1 |
| 2 | kadungora | 2 |
| 3 | cikajang | 0 |
| 4 | pameungpeuk | 4 |
| 5 | samarang | 5 |

Source: (Research Results, 2024)

Table 6 is an example of the results of the transformation on the market name attribute where the attribute was originally categorical, namely the names of markets turned into a numeric number 1. Next, One hot encoding is performed, which converts category attributes into binary vectors, the following One hot encoding results are presented in Table 7.

Table 7. One hot encoding transformation

| No | Month Attributes | | |
|---|---|---|---|
| | February | April | December |
| 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 |
| 5 | 1 | 0 | 0 |

Source: (Research Results, 2024)

***Feature Engineering***

In this stage, Feature Engineering technique is also applied which is a technique of engineering features, which are needed to create a suitable data representation for the learning model. This process produces new features that are a combination of the original features, both in the form of squares and interactions between features presented in Table 8.

Table 8. *Feature Engineering*

| No | Feature Name |
|---|---|
| 1 | production_expansion_interaction |
| 2 | production(tons)2 |
| 3 | production (tons) X net harvested area (ha) |
| 4 | net harvested area (ha)2 |
| 5 | production (tons) X productivity (kw/ha) |
| 6 | net harvested area (ha) X productivity (kw/ha) |
| 7 | productivity (kw/ha)2 |

Source: (Research Results, 2024)

**6. Data Normalization**

After the data cleaning and data transformation stages, the next stage is the data normalization stage, this stage includes steps to scale the data into a certain range or a certain distribution. In this stage, data normalization is carried out using 3 techniques, namely Min-Max Normalization, Roust Scaller, and Sacaller Standard. The following are the stages:

***Min-Max Normalization***

Min-max normalization is a data scaling technique that changes the values in a dataset so that they fall within a certain range, usually between 0 and 1. The results of normalization using Min-Max Normalization are presented in Table 9.

Table 9. *Min-Max Normalization*

| No | Productivity Attributes | |
|---|---|---|
| | Before | After |
| 1 | 150.260 | 0.429777 |
| 2 | 150.260 | 0.429777 |
| 3 | 150.260 | 0.429777 |
| 4 | 176.265 | 1.000000 |
| 5 | 176.265 | 1.000000 |

Source: (Research Results, 2024)

Table 9 presents the results of normalizing productivity data using the Min-Max method. This method converts the original diverse production values into a scale between 0 and 1.

***Robust Scaler***

The robust scaler technique is one of the data scaling techniques used in data pre-processing. This technique serves to overcome outliers well, the following robust scaler results are presented in Table 10.

Table 10. Normalization Robust Scaler

| No | Productivity Attributes | |
|---|---|---|
| | Before | After |
| 1 | 150.260 | 0.658555 |
| 2 | 150.260 | 0.658555 |
| 3 | 150.260 | 0.658555 |
| 4 | 176.265 | 2.898966 |
| 5 | 176.265 | 2.898966 |

Source: (Research Results, 2024)

***Standard Scaler***

The standard scaler technique is a data scaling method used to standardize the features in the data so that they have certain properties that make it easier for data analysis. The following are the results of normalization using the Standard scaler technique presented in Table 11.

Table 11. Scaller Standard Normalization

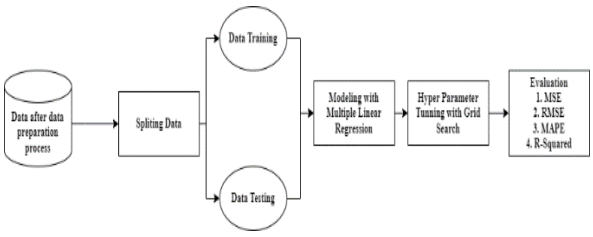
| No | Production Attributes | |
|---|---|---|
| | Before | After |
| 1 | 286.0 | -1.935776 |
| 2 | 286.0 | -1.935776 |
| 3 | 286.0 | -1.935776 |
| 4 | 1390.0 | -1.349065 |
| 5 | 1390.0 | -1.349065 |

Source: (Research Results, 2024)

Table 11 shows that the production values have been successfully normalized using Standard Scaler. This method has transformed the production data so that it has a mean of 0 and a standard deviation of 1.

**D. Modelling**

In this research, the model chosen is Multiple Linear Regression as an algorithm in the prediction process built using the Python programming language with Google Collaboratory tools. The following is the modeling flow shown in Figure 4.




Source: (Research Results, 2024)

Figure 4. Modeling Flow

Figure 4 is the flow of modeling that will be carried out, starting with the results of data reduction, aggregation, integration, cleaning, transformation and normalization. At the transformation stage use two techniques namely Label encoding and One hot encoding, while at the normalization stage use three techniques namely Min-Max Normalization, robust scaler, and standard scaler, which then divides the data with a proportion of 80% test data and 20% training data. Next, the modeling process will be carried out using the Multiple Linear Regression algorithm, then find the optimal value for the hyperparameter using Grid Search, and get the model performance.

In this modeling, 8 trials were conducted to get the best model using MSE, RMSE, MAPE and R-squared evaluations. The following are the best model results based on the Mean Squared Error evaluation presented in Table 12.

Table 12. Best Model Based on MSE

| No | Model Testing | MSE Evaluation |
|---|---|---|
| 1 | LE | 272,433,694 |
| 2 | OHE | 211,348,424 |
| 3 | OHE+MM | 211,348,424 |
| 4 | OHE+RS | 211,348,424 |
| 5 | OHE+FE+RS | 211,348,424 |
| 6 | OHE+FE+SS | 203,677,307 |
| 7 | OHE+FE+SS+HPT | **202,134,650** |
| 8 | OHE+FE+RS+HPT+CV | 224,165,823 |

Source: (Research Results, 2024)

Description:
LE : *Label Encoding*
OHE : *One hot encoding*
MM : *Min-Max Normalization*
RS : *Robust Scaler*
SS : *Standard Scaler*
FE : *Feature Engineering*
CV` : *Cross Validation*

Table 12 presents the performance comparison results of several tested Multiple Linear Regression models. The evaluation results show that the model with a combination of One hot encoding, Feature Engineering, Standard Scaler, Hyperparameter Tuning, and no cross-validation (7th model) has the best performance based on the lowest Mean Squared Error (MSE) value.

Furthermore, the RMSE evaluation was carried out to find the best model with the results presented in Table 13.

Table 13. RMSE evaluation

| No | Model Testing | RMSE Evaluation |
|---|---|---|
| 1 | LE | 16,506 |
| 2 | OHE | 14,538 |
| 3 | OHE+MM | 14,538 |
| 4 | OHE+RS | 14,538 |
| 5 | OHE+FE+RS | 14,538 |
| 6 | OHE+FE+SS | 14,272 |
| 7 | OHE+FE+SS+HPT | **14,217** |
| 8 | OHE+FE+RS+HPT+CV | 14,972 |

Source: (Research Results, 2024)

In Table 4.16, the evaluation results show that the model with the combination of One hot encoding, Feature Engineering, Standard Scaler, and Hyperparameter Tuning (7th model) has the lowest RMSE value. This indicates that this model is able to provide the most accurate price prediction among other models.

Furthermore, to find the best model, an evaluation is also carried out based on the MAPE value with the results presented in Table 14.

Table 14. MAPE evaluation

| No | Model Testing | MAPE % Evaluation |
|---|---|---|
| 1 | LE | 31,55 |
| 2 | OHE | 29,25 |
| 3 | OHE+MM | 29,25 |
| 4 | OHE+RS | 29,25 |
| 5 | OHE+FE+RS | 29,25 |
| 6 | OHE+FE+SS | 29,30 |
| 7 | OHE+FE+SS+HPT | **29,16** |
| 8 | OHE+FE+RS+HPT+CV | 31,18 |

Source: (Research Results, 2024)

Table 14 presents the performance evaluation results of various Multiple Linear Regression models using the Mean Absolute Percentage Error (MAPE) metric. The evaluation results show that the model with the combination of One hot encoding, Feature Engineering, Standard Scaler, and Hyperparameter Tuning (7th model) has the lowest MAPE value.


Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5882

In addition to using MSE, RMSE, and MAPE evaluations, an R-squared evaluation was conducted to determine the best model, with the results presented in Table 15.

Table 15. R-Squared Evaluation

| No | Model Testing | R-Squared Evaluation |
|---|---|---|
| 1 | LE | 0,084 |
| 2 | OHE | 0,289 |
| 3 | OHE+MM | 0,289 |
| 4 | OHE+RS | 0,289 |
| 5 | OHE+FE+RS | 0,289 |
| 6 | OHE+FE+SS | 0,315 |
| 7 | OHE+FE+SS+HPT | **0,320** |
| 8 | OHE+FE+RS+HPT+CV | 0,246 |

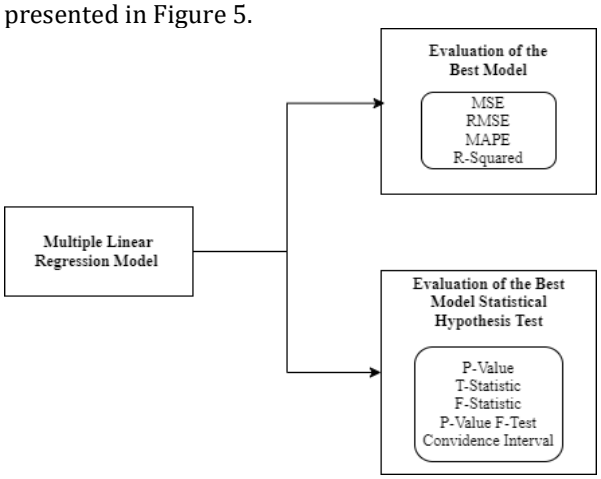
Source: (Research Results, 2024)

In Table 15, the R-squared value shows how well the model can explain data variability. The evaluation results show that the model with the combination of One hot encoding, Feature Engineering, Standard Scaler, and Hyperparameter Tuning (7th model) has the highest R-squared value of 0.320. This means that this model can explain about 32% of the data variability.

Based on the evaluation results using various metrics (MSE, RMSE, MAPE, and R-squared), the model with a combination of One hot encoding, Feature Engineering, Standard Scaler, and Hyperparameter Tuning proved to be the best model in predicting red chili prices. This model consistently shows the best performance in terms of prediction accuracy and ability to explain data variability.

**E. *Evaluation***

The next stage, namely evaluation, where this stage is carried out to determine the quality of the prediction model using Multiple Linear Regression, the evaluation stages that will be carried out are presented in Figure 5.




Source: (Research Results, 2024)

Figure 5: Stages of Evaluation

Figure 5 is an evaluation flow consisting of two stages, namely model performance evaluation using the MSE, RMSE, R-Squared, and MAPE matrix to measure how well the model predicts the value of the dependent variable based on the available independent variables and Statistical Hypothesis Test Evaluation to identify independent variables. The following are the evaluation results:

**1. Model Performance Evaluation**

Model performance evaluation is performed using the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared evaluation matrices. The results of eight model experiments in the modeling stage show that the best model is the seventh model, which is a model that uses One Hot Encoding (OHE), Feature Engineering (FE), Standard Scaler (SS), and Hyperparameter Tuning (HPT). The best evaluation results are shown by the lowest MSE, RMSE, MAPE, and highest R-Squared, as seen in Table 16.

Table 16. Evaluation of the OHE+FE+SS+HPT Model

| Evaluation of the Best Model | | | |
|---|---|---|---|
| MSE | RMSE | MAPE | R-Squared |
| 202,134,650 | 14,217 | 29.16 | 0.320 |

Source: (Research Results, 2024)

Based on Table 16, the low MSE value indicates that the average square of the difference between the predicted value and the actual value is quite small, which means that the model provides a fairly accurate prediction. The low RMSE indicates that the average prediction error of the model is relatively small, while the low MAPE indicates that the average percentage absolute error between the predicted and actual values is quite small. Although the R-Squared of 0.320 is not very high, this result still makes a significant contribution in explaining the fluctuation of red chili prices in Garut Regency.

This result is in line with previous research using ARIMA and Fuzzy Time Series methods which also show accurate prediction results, but have weaknesses in overcoming seasonal factors and inter-variable interactions. The LSTM model, which has been used in similar studies, also produces predictions with good accuracy, but the complexity of its implementation is a challenge. In this case, Multiple Linear Regression with Feature Engineering proved to provide better results than other approaches that rely solely on time series data, with sufficient simplicity and clearer interpretation.

**2. Statistical Hypothesis Test Evaluation**

The Statistical Hypothesis Test is conducted to identify independent variables that have a significant influence on the dependent variable and to understand the uncertainty in the estimation of model coefficients. The results of the Statistical Hypothesis Test are presented in Table 17.

Table 17: Statistical Hypothesis Test Results

| No | Variables | Coefficient | Std Err | T-*Statistic* | P-*Value* | 95% *Confidence Interval* |
|---|---|---|---|---|---|---|
| 1 | const | 42510 | 650,367 | 65,357 | 0,000 | 4.12e+04, 4.38e+04 |
| 2 | month_may | -6864,7216 | 1390,846 | -4,936 | 0,000 | -9597.127, -4132.316 |
| 3 | month_october | -9211,6146 | 2228,265 | -4,134 | 0,000 | -1.36e+04, -4834.047 |
| 4 | red chili commodity_rawit | -126200 | 33300 | -3,791 | 0,000 | -1.92e+05, -6.08e+04 |
| 5 | month_september | -8657,2606 | 2330,653 | -3,715 | 0,000 | -1.32e+04, -4078.545 |
| 6 | no | 109300 | 30000 | 3,646 | 0,000 | 5.04e+04, 1.68e+05 |
| 7 | month_agust | -7022,9395 | 2039,53 | -3,443 | 0,001 | -1.1e+04, -3016.155 |
| 8 | month_december | -8929,7052 | 2664,185 | -3,352 | 0,001 | -1.42e+04, -3695.745 |
| 9 | year | -61330 | 18900 | -3,243 | 0,001 | -9.85e+04, -2.42e+04 |
| 10 | month_april | -3780,4361 | 1186,384 | -3,187 | 0,002 | -6111.162, -1449.710 |
| 11 | month_november | -7669,989 | 2437,623 | -3,147 | 0,002 | -1.25e+04, -2881.124 |
| 12 | production(tons)2 | 17880 | 6013,363 | 2,974 | 0,003 | 6070.030, 2.97e+04 |
| 13 | month_june | -4569,3833 | 1556,85 | -2,935 | 0,003 | -7627.914, -1510.853 |
| 14 | production_expansion_interaction | -6648,0086 | 2291,573 | -2,901 | 0,004 | -1.11e+04, -2146.069 |
| 15 | production(tons) net harvested area (ha) | -6648,0086 | 2291,573 | -2,901 | 0,004 | -1.11e+04, -2146.069 |
| 16 | month_january | 3130,6202 | 1098,421 | 2,85 | 0,005 | 972.703, 5288.538 |
| 17 | market_ kadungora name | 2335,0369 | 840,718 | 2,777 | 0,006 | 683.393, 3986.680 |
| 18 | month_july | -4915,2334 | 1972,616 | -2,492 | 0,013 | -8790.562, -1039.905 |
| 19 | production (ton) productivity (kw/ha) | -73570 | 30000 | -2,451 | 0,015 | -1.33e+05, -1.46e+04 |
| 20 | productivity (kw/ha)2 | 110700 | 49700 | 2,226 | 0,026 | 1.3e+04, 2.08e+05 |
| 21 | market_leles name | 1873,0183 | 851,186 | 2,2 | 0,028 | 200.810, 3545.227 |
| 22 | productivity (kw/ha) | -55480 | 25500 | -2,175 | 0,030 | -1.06e+05, -5377.377 |

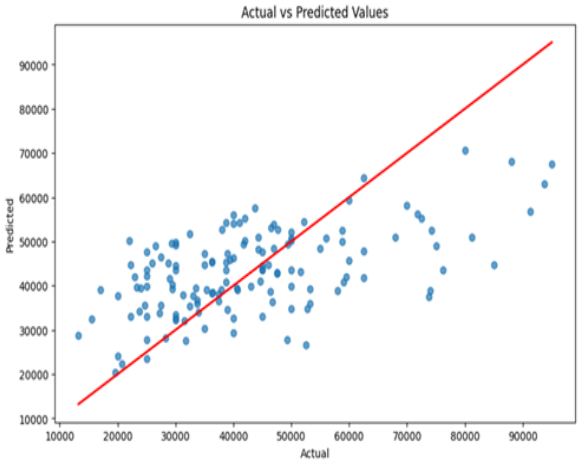
Source: (Research Results, 2024)

Based on Table 17, several independent variables have a significant influence on the price of red chili in the Garut Regency, as indicated by the p-value <0.05 on several variables. For example, the variable month_mei with a coefficient of -6,864.72 and a p-value of 0.000 indicates that the price of red chili tends to be lower by Rp 6,864.72 in May compared to the reference month. Meanwhile, the production variable (tons) with a coefficient of 17,880 and a p-value of 0.003 shows that every increase in production by 1 ton increases the average price of red chili by Rp 17,880. This confirms the importance of taking into account seasonal variables and production variables in predicting red chili prices.

This finding provides important information for policymakers and red chili market players in understanding the factors that affect the price of red chili, both in terms of seasonality and production. These significant variables can be used as a basis for planning and decision-making to improve market efficiency and price stability.

**F. Deployment**

The result dissemination stage is carried out to display the modeling results in the form of visualization so that it is easier for interested parties to understand. Figure 6 shows the visualization results of the red chili price prediction model using the Multiple Linear Regression algorithm.




Source : (Research Results, 2024)

Figure 6. Visualization of Prediction Model

Figure 6 depicts a regression model evaluation graph showing the relationship between actual and predicted values, with blue dots representing pairs of actual and predicted data. The red diagonal line represents the ideal line where the model prediction exactly matches the actual value. Most of the data points follow the trend of this diagonal line, indicating that the model performs reasonably well.

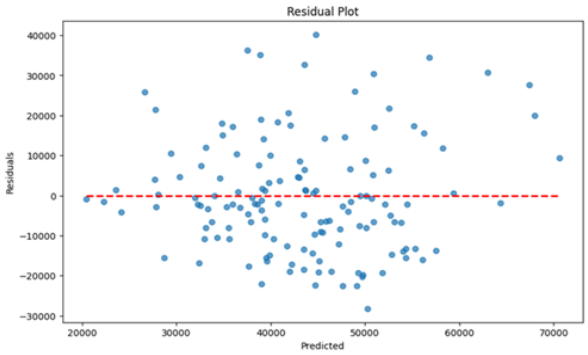
Next, a residual analysis was performed to evaluate the distribution of the prediction error



Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5882

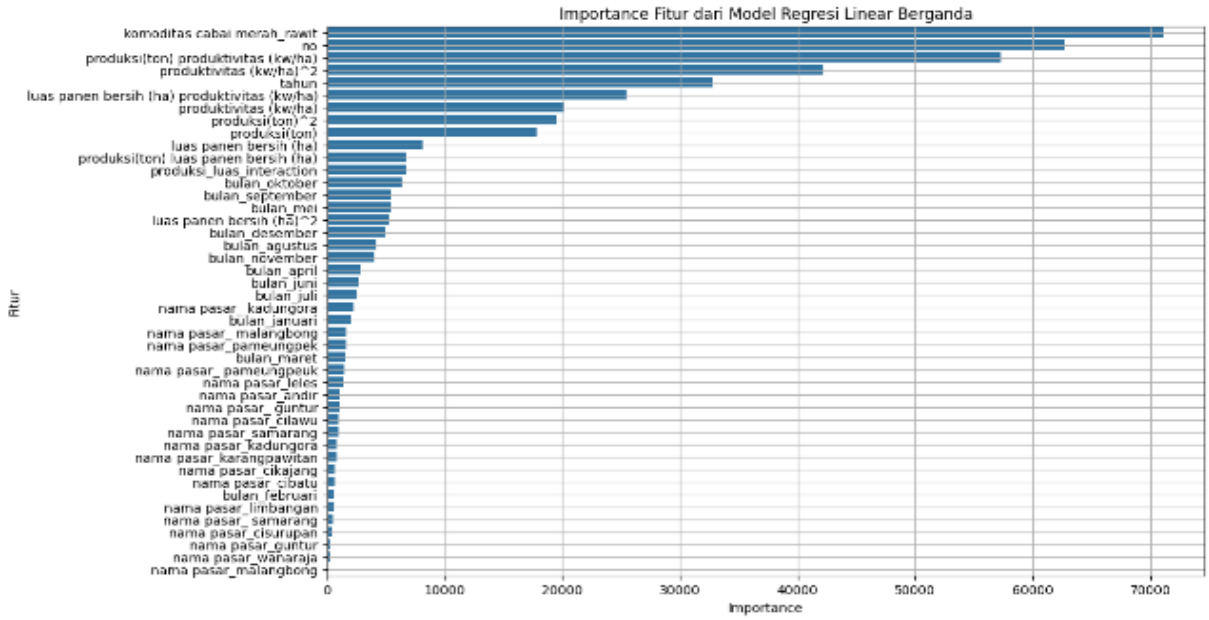
(residual) to ensure that there is no obvious pattern in the error. The results of the residual analysis can be seen in Figure 7.




Source: (Research Results, 2024)
Figure 7. Residual Plot

In Figure 7, the blue dots are randomly scattered around the red line, indicating that the multiple linear regression model is able to capture the data pattern without showing patterns that could potentially cause bias. With a random distribution of residuals, the model is considered good enough to provide accurate predictions.

After building the multiple linear regression model and evaluating its performance, the next step is to identify which features contribute most to the prediction of red chili prices. The results of the analysis are presented in Figure 8.




Source : (Research Results, 2024)
Figure 8. Feature Importance

Figure 8 is a graph that shows the importance of each feature in the multiple linear regression model that has been built. The production (tons), productivity (kw/ha), and net harvested area features have a significant influence on the prediction of red chili prices. The interaction between these variables plays an important role in predicting red chili prices, with results consistent with other studies that emphasize the importance of production and productivity variables in price fluctuations.

These results provide important insights for farmers, traders, and other stakeholders in planning red chili planting and distribution strategies based on more accurate price predictions. The model can also be used as a basis for developing price predictions for other commodities with a similar approach, taking into account the interaction between influential features.

## CONCLUSION

This study successfully predicted the price of red chili peppers in Garut Regency using multiple linear regression, in accordance with the research objective to identify significant factors that affect the price. Based on the T-test, it was found that five attributes had a significant effect, namely year, month, production, net harvested area, and productivity. The best performance was achieved through the combination of One Hot Encoding, Feature Engineering, Standard Scaler, and Hyperparameter Tuning, with MSE of 202,134,650, RMSE of 14,217, MAPE of 29.16%, and R-Squared of

0.320, indicating the model is quite accurate with a prediction error below 30%. These findings are relevant to the research objective of developing a model that can predict prices with high accuracy. However, this study has limitations, such as limited data coverage and linearity assumptions, which can be addressed in future studies with more complex models.

## REFERENCE

[1] A. Hia, R. Nurmalina, and A. Rifin, "Efisiensi Pemasaran Cabai Rawit Merah Di Desa Cidatar Kecamatan Cisurupan Kabupaten Garut," *Forum Agribisnis*, vol. 10, no. 1, pp. 36–45, 2020, doi: 10.29244/fagb.10.1.36-45.

[2] B. P. Statistik, "Rata-Rata Konsumsi per Kapita Seminggu Beberapa Macam Bahan Makanan Penting, 2007-2023," *Badan Pusat Statistik*, 2024. https://www.bps.go.id/id/statistics-table/1/OTUwIzE=/rata-rata-konsumsi-per-kapita-seminggu-beberapa-macam-bahan-makanan-penting-2007-2023.html

[3] Badan Pusat Statistik Kabupaten Garut, "Produksi Tanaman Sayuran Cabai Rawit Menurut Kecamatan di Kabupaten Garut (Kuintal ), 2020-2021," *BPS Kabupaten Garut*, 2021. https://garutkab.bps.go.id/

[4] N. Moha Lalapa and W. Yunus, "Implementasi Metode Regresi Linear Sederhana Untuk Prediksi Harga Cabai Rawit," *J. Ilm. Ilmu Komput. Banthayo Lo Komput.*, vol. 2, no. 2, p. 96, 2023, doi: https://doi.org/10.15548/jostech.v2i1.3802

[5] P. Ekawati, Nia; Wilson, "Prediksi Harga Cabai Merah Menggunakan Jaringan Syarat Tiruan," *Journalinformatics Electron. Eng.*, vol. 1, no. 2, pp. 58–65, 2021, [Online]. Available: https://ejournal.poltektedc.ac.id/index.php/jiee/article/view/537/399

[6] L. Susanti, S. J. Pririzki, Z. Zeleansi, and D. Y. Dalimunthe, "Prediksi Harga Cabai Rawit Merah Sebagai Kebutuhan Pangan Masyarakat Di Kota Pangkalpinang," in *Proceedings of …*, 2022, pp. 140–145. [Online]. Available: https://journal.ubb.ac.id/snppm/article/view/3752

[7] V. Komaria, N. El Maidah, and M. A. Furqon, "Prediksi Harga Cabai Rawit di Provinsi Jawa Timur Menggunakan Metode Fuzzy Time Series Model Lee," *Komputika J. Sist. Komput.*, vol. 12, no. 2, pp. 37–47, 2023, doi: 10.34010/komputika.v12i2.10644.

[8] M. David, I. Cholissodin, and N. Yudistira, "Prediksi Harga Cabai menggunakan Metode Long-Short Term Memory (Case Study : Kota Malang)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 3, pp. 1214–1219, 2023.

[9] M. Bedy Purnama, S.si., *Pengantar Machine Learning*. Bandung: Informatika Bandung, 2019.

[10] J. Brzozowska, J. Pizoń, G. Baytikenova, A. Gola, A. Zakimova, and K. Piotrowska, "Data Engineering In Crisp-Dm Process Production Data – Case Study," vol. 19, no. 3, pp. 83–95, 2023, doi: 10.35784/acs-2023-26.

[11] Y. Yudiana, A. Yulia, and N. Khofifah, "Prediksi Customer Churn Menggunakan Metode CRISP-DM Pada Industri Telekomunikasi Sebagai Implementasi Mempertahankan Pelanggan," vol. 8, no. 1, pp. 1–20, 2023, doi: https://doi.org/10.30631/ijoieb.v8i1.1710

[12] N. Widiawati, B. N. Sari, and T. N. Padilah, "Clustering Data Penduduk Miskin Dampak Covid-19 Menggunakan," vol. 6, no. 1, pp. 55–63, 2022, doi: https://doi.org/10.30871/jaic.v6i1.3266

[13] B. N. Azmi, A. Hermawan, and D. Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," vol. 4, no. 4, pp. 281–290, 2023, doi: https://doi.org/10.35746/jtim.v4i4.298

[14] Y. Mulyanto and A. Algi Fari, "Analisis Keamanan Login Router Mikrotik Dari Serangan Bruteforce Menggunakan Metode Penetration Testing (Studi Kasus: Smk Negeri 2 Sumbawa)," *J. Inform. Teknol. dan Sains*, vol. 4, no. 3, pp. 145–155, 2022, doi: 10.51401/jinteks.v4i3.1897.

[15] I. Permana and F. N. Salisah, "The Effect of Data Normalization on the Performance of the Classification Results of the Backpropagation Algorithm Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation," vol. 2, no. 1, pp. 67–72, 2022, doi: https://doi.org/10.57152/ijirse.v2i1.311