Jurnal CoreIT, Vol.11, No.1, June 2025
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

APPLICATION OF K-NEAREST NEIGHBOR REGRESSION METHOD FOR RICE YIELD PREDICTION

Lestari Handayani[1*], Alif Alfarabi B.[2], Tasya Aprilia[3], Indah Wulandari[4], Jasril[5], Siti Ramadhani[6], Elvia Budianita[7]

Department of Informatics Engineering, Universitasi Islam Negeri Sultan Syarif Kasim Riau, Indonesia
lestari.handayani@uin-suska.ac.id(*), 12050112688@students.uin-suska.ac.id, 12050123544@students.uin-suska.ac.id, 12050120353@students.uin-suska.ac.id, jasril@uin-suska.ac.id, siti.ramadhani@uin-suska.ac.id, elvia.budianita@uin-suska.ac.id

Abstract. Rice plants with the Latin name Oryza Sativa are food plants that are widely used as the main food crop in various countries, one of which is Indonesia. Indonesia is ranked 4th as the largest rice consuming country in the world. This requires the availability of rice to be maintained. Unstable rice production can be a problem. One of the districts that has experienced a decline in rice production in recent years is the district of Lima puluh kota located in West Sumatra province. This requires prediction of rice production so that it can be used as a benchmark for the future. This study uses data on rice production in fifty cities from 2013 to 2023. The method used to predict is k-nearest neighbor regression (KNN Regression). The data division uses rasio 90 : 10. In testing the data used is divided into 2, namely normal data and data that has been normalized. The test results produce the smallest mean absolute percentage error (MAPE) value of 6.98% on normal data, the value of k is 6 with data division using k-fold 5. Based on the resulting MAPE value, it can be said that KNN Regression can predict rice production results very accurately.

Keywords: K-Nearest Neighbor Regression, Mean Absolute Percentage Error, Rice, Prediction

Received June 2024 / **Revised** January 2025 / **Accepted** May 2025

This work is licensed under a Creative Commons Attribution 4.0 International License.

INTRODUCTION

Food is a basic and most essential need for humans to sustain life and life. Food and nutrition development is one of the efforts to realize food security, this has been regulated in food law number 7 of 1996 concerning food and government regulation number 68 of 2002 concerning food security [1]. Food security is a condition where an area can guarantee its population to get enough, decent and quality food. Indonesia is included in a country where almost the entire population makes rice as a staple [2]. Rice plants have long been cultivated and originated from two continents, namely the Asian continent and the West African continent [3]. Rice plants will produce rice which is the main source of carbohydrates [4]. to maintain food sustainability in one way is to make food barns [5]. Food barns are useful for regional food reserves and can cope in the event of a food crisis when food crop production decreases.

Rice production in Indonesia has increased and decreased, such as a decrease in 2021 and then an increase in 2022. In 2021, Indonesia produced 54.42 million tons of milled dry grain (MDG) of rice which decreased by 0.43%, which is equivalent to 233.91 thousand tons from 2020 [6]. In 2022 rice production increased where the total rice production in that year amounted to 54.75 tons of MDG which increased by 0.61% which is equivalent to 333.68 thousand tons from 2021 [7]. Rice producing areas in Indonesia mostly come from the island of Java, but there is a province in Sumatra which is also one of the largest rice producers, namely West Sumatra. Rice production in West Sumatra has decreased in the last few years which can be in several districts in West Sumatra such as the district of lima puluh kota.

The right solution can be found if there is a decrease in production. One of the data mining methods that can be used to make predictions is K-Nearest Neighbor Regression (KNN-Regression). KNN-Regression is an algorithm that groups data based on the nearest neighbor [3]. Grouping is done by determining the value of k, namely the neighborhood and the resulting Mean Absolute Percentage Error (MAPE) value will be a measure of the error rate of the model that has been made. Previous research using the KNN-Regression method in predicting the amount of coconut oil production resulted in the lowest RMSE of 0.109 [4]. Other research that predicts outgoing goods at TB. Wijaya Bangunan using KNN-Regression resulted in the cobra sherlock spoon hinge getting an RMSE value of 3.55 which means it produces the best accuracy results [5].

DOI: 10.24014/coreit.v11i1.30907

Previous research that predicted closing stock prices at PT Adaro Energy Indonesia using KNN-Regression using 11 attributes resulted in a Root Mean Square Error (RMSE) value of 35.02, R-Squared (R2) = 0.99 and MAE = 24.54 [6].
The research conducted aims to help solve the problem of predicting food crop production in the form of rice using mining methods on rice production in the district of 50 cities. Prediction is done using the KNN-Regression method because it is effective for data that has a lot of noise and high training data [4]. The prediction of rice production in District 50 cities is carried out for policy determination in order to meet the food needs of the region.

METHODS

This research uses data on rice production from the Food Crops and Horticulture Office of District of lima puluh kota from 2013 to 2023.

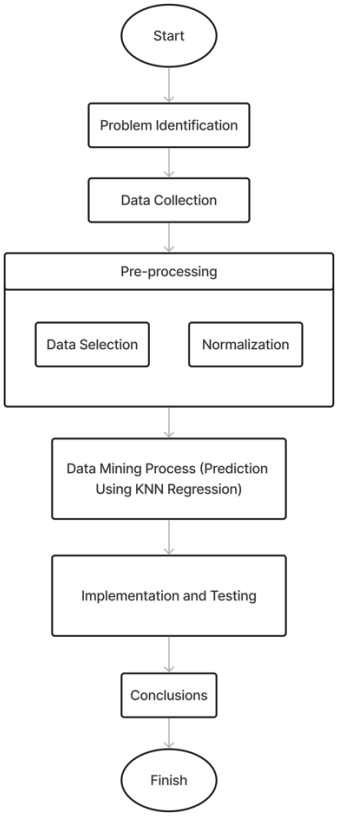

Figure 1. Research method

A. Data Pre-processing

Pre-processing is the initial stage in the data mining cycle. Pre-processing is a stage of preparing data where raw data will be converted through several processes into a form of data that is easy to understand. In this research, the data pre-processing stage only goes through 2 stages, namely data selection and data normalization.

1. Data Selection

Data selection is a process to determine the appropriate data source [7]. In this research, the data used is secondary data. The secondary data used in this study are secondary data on rice production results obtained from the agriculture and horticulture office of Lima Puluh Kota Regency. The data used consists of 2 variables, namely planting and production. The data used is data from 2013 to 2023.

2. Normalization Data

Normalization process is a process that aims to reduce the possibility of data anomalies and inconsistencies in the data. The main purpose of normalizing data is so that data that has small dimensions can still represent the original data without losing data characteristics [7].

Table 1. Normalized data

TANAM	PRODUKSI
-0,679874391	0,170233947
-0,717111008	0,638606595
0,213804423	0,729902651
-0,102017256	0,105356041
…….	…….
-0,44093943	-0,455006515
-0,691941813	-0,524339783
-0,221657128	0,138834474
-0,177524841	-0,08813093

B. K-Nearest Neighbor Regression

The K-Nearest Neighbor Regression algorithm is an algorithm that groups data based on its neighborliness, the grouping of data depends on the value of k. to evaluate the model can use evaluation methods such as MAPE. After getting the MAPE value, it can be used to measure the error in the model that has been made [8]. Here's how KNN Regression works:

- Determining the value of k
- Calculate the distance of the new data point to all data points in the training dataset. Calculate the distance using the distance function.
- Determine the nearest neighbor based on the closest distance value
- Calculate predictions using the K-Nearest Neighbor Regression formula.

Here is the KNN Regression formula:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i \quad (1)$$

Keterangan:
$\hat{y}$ = Predicted data
k = Number of neighbors
y_i = i-data

C. Euclidean Distance

The distance function is a method used to measure the distance of a data point to another data point in a data set or measure the distance of two points in Euclidean, the Euclidean plane itself includes Euclidean two dimensions or even more [9]. Distance functions are usually used when doing data processing and machine learning.

$$\mathrm{d_g(i,j)} = \sqrt{\sum(\mathrm{x_i} - \mathrm{x_j})^2} \quad (2)$$

Description:
$d_g(i,j)$ = Euclidean distance from i-data to j-data
x_i = i-data
x_j = j-data

D. Mean Absolute Percentage Error

MAPE is one of the model evaluation methods. MAPE is a method used to calculate the error value in percentage form. The use of MAPE in prediction models is used to see the level of accuracy of the prediction results produced by the model and actual data [10]. The following is the MAPE formula:

$$MAPE = \frac{1}{n}\sum \left|\frac{y - y'}{y}\right| \times 100\% \quad (1)$$

Description:
MAPE = MAPE value
y = Actual data
y' = Predicted data
N = Number of data

When the resulting MAPE value is between 10% to 20%, it is included in the good category. When the resulting MAPE value is between 20% - 50% then the model is included in the sufficient category and when the resulting MAPE value is greater than 50% then the model is included in the bad category in making predictions. The following is a table of criteria for the MAPE value range:

Table 2 MAPE Criteria

MAPE Value	Criteria
<10%	Very good
10%-20%	Good
20%-50%	Fair
>50%	Poor

RESULT AND DISCUSSION

The data used is 132 data. The data division uses a ratio of 90 : 10, with a total of 118 training data and a total of 14 test data. The research dataset is processed using the K-Nearest Neighbor Regression method. This research uses a scenario to find the most effective k value that produces the lowest MAPE value. Here are some parameter limitations in this test:

- Range of k values: 1 - 50
- Data: Normal and Normalized

The following MAPE value is generated according to the scenario performed:

	MAPE	
K Value	Data Without Normalization	Normalized Data
1	9,70%	9,70%
2	9,40%	9,40%
3	7,62%	7,62%
4	7,83%	7,83%
5	7,71%	7,71%
6	7,35%	7,35%
7	7,17%	7,17%
8	7,05%	7,05%
9	7,12%	7,12%
10	7,16%	7,16%
11	7,06%	7,06%

12	6,98%	6,98%
13	6,83%	6,83%
14	6,80%	6,80%
15	6,78%	6,78%
16	6,82%	6,82%
17	6,826%	6,826%
18	6,75%	6,75%
19	6,73%	6,73%
20	6,67%	6,67%
21	6,62%	6,62%
22	6,58%	6,58%
23	6,57%	6,57%
24	6,54%	6,54%
25	6,49%	6,49%
26	6,48%	6,48%
27	6,53%	6,53%
28	6,51%	6,51%
29	6,52%	6,52%
30	6,51%	6,51%
31	6,48%	6,48%
32	6,49%	6,49%
33	6,50%	6,50%
34	6,49%	6,49%
35	6,45%	6,45%
36	6,42%	6,42%
37	6,425%	6,425%
38	6,423%	6,423%
39	6,43%	6,43%
40	6,44%	6,44%
41	6,46%	6,46%
42	6,45%	6,45%
43	6,426%	6,424%
44	6,425%	6,425%
45	6,41%	6,41%
46	6,440%	6,440%
47	6,43%	6,43%
48	6,39%	6,39%
49	**6,36%**	**6,36%**
50	6,38%	6,38%

Based on the table above, it can be seen that the average k value that gets the best value is k = 49 using normalized data. The smallest MAPE obtained is 6.36% using normalized data. The following is a comparison graph between the actual data and the predicted data from the model.

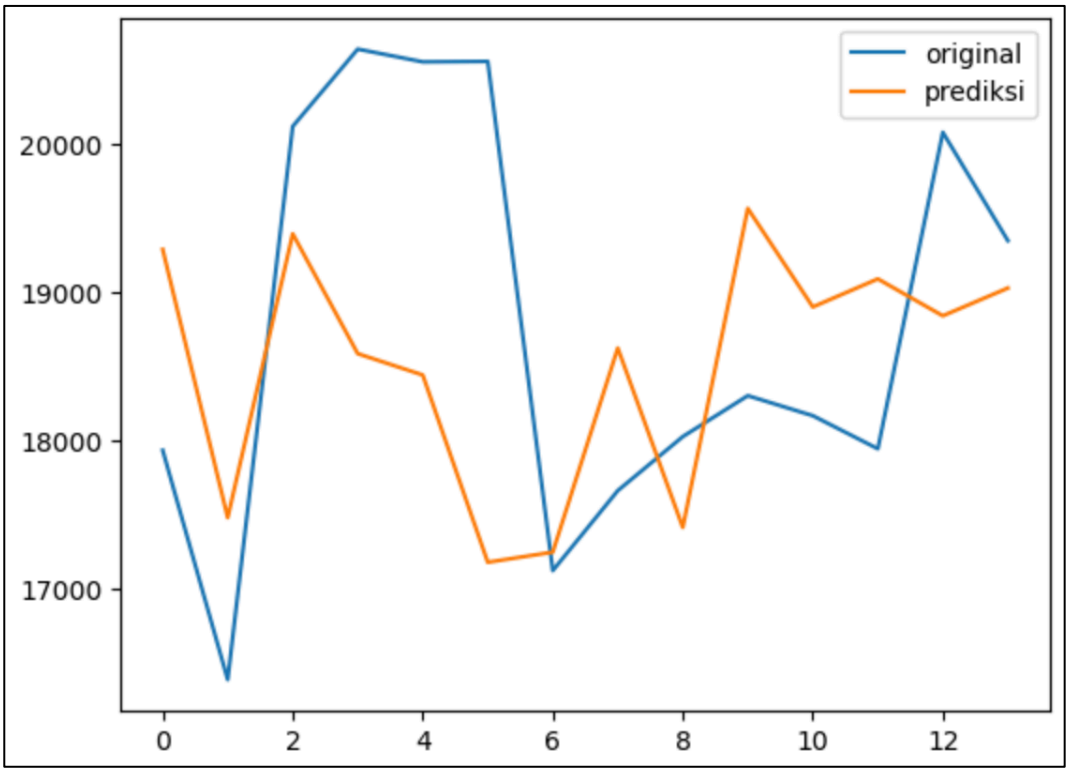

Figure 2. Graph of predicted and actual data

CONCLUSION

In this study, modeling was carried out using the K-Nearest Neighbor Regression algorithm to predict rice production results. Modeling is done using the value of neighborliness with a range of 1 - 50, data division using a ratio of 90: 10 and using two kinds of data, namely data that has been normalized and data that is not normalized. Based on the results of the experiments carried out, the smallest MAPE value was obtained at a neighborhood value of 49 using normalized data. The resulting MAPE value is 6.36%. With the results obtained, it can be concluded that this model can predict very accurately.

REFERENCES

[1] R. Chaireni, D. Agustanto, R. A. Wahyu, and P. Nainggolan, "Ketahanan Pangan Berkelanjutan," *Jurnal Kependudukan dan Pembangunan Lingkungan*, vol. 1, no. 2, pp. 70–79, 2020, [Online]. Available: jkpl.ppj.unp.ac.id/index.php/JKPL/article/view/13

[2] M. N. Fawaiq, A. Jazuli, and M. M. Hakim, "Prediksi Hasil Pertanian Padi Di Kabupaten Kudus Dengan Metode Brown'S Double Exponential Smoothing," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 4, no. 2, p. 78, 2019, doi: 10.29100/jipi.v4i2.1421.

[3] E. Triyanto, H. Sismoro, and A. D. Laksito, "Implementasi Algoritma Regresi Linear Berganda Untuk Memprediksi Produksi Padi Di Kabupaten Bantul," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 4, no. 2, pp. 66–75, 2019, doi: 10.36341/rabit.v4i2.666.

[4] Mergono Adi Ningrat, Carolina Diana Mual, and Yohanis Yan Makabori, "Pertumbuhan dan Hasil Tanaman Padi (Oryza sativa L.) pada Berbagai Sistem Tanam di Kampung Desay, Distrik Prafi, Kabupaten Manokwari," *Prosiding Seminar Nasional Pembangunan dan Pendidikan Vokasi Pertanian*, vol. 2, no. 1, pp. 325–332, Sep. 2021, doi: 10.47687/snppvp.v2i1.191.

[5] ketahananpangan.probolinggokab.go.id, "LUMBUNG PANGAN MASYARAKAT, DUKUNGAN DAN PERANNYA DALAM PEMBANGUNAN KETAHANAN PANGAN DAERAH ," ketahananpangan.probolinggokab.go.id. Accessed: Jun. 17, 2024. [Online].

Available: https://ketahananpangan.probolinggokab.go.id/2022/12/06/lumbung-pangan-masyarakat-dukungan-dan-perannya-dalam-pembangunan-ketahanan-pangan-daerah/

[6] bps.co.id, "Luas Panen, Produksi, dan Produktivitas Padi Menurut Kabupaten/Kota Hasil Kerangka Sampel Area (KSA) 2021-2023 ," bps.co.id. Accessed: Jun. 03, 2024. [Online]. Available: https://sumbar.bps.go.id/subject/53/tanaman-pangan.html#subjekViewTab3

[7] P. Dan, "Produksi Padi Di Indonesia 2022," *Badan Pusat Statistik*, pp. 15–20, 2022, [Online]. Available: https://www.bps.go.id/publication/2023/08/03/a78164ccd3ad09bdc88e70a2/luas-panen-dan-produksi-padi-di-indonesia-2022.html

[8] Mukhlisin, M. Imrona, and D. T. Murdiansyah, "Prediksi Harga Beras Premium dengan Metode Algoritma K-Nearest Neighbor," *e-Proceeding of Engineering*, vol. 7, no. 1, pp. 2714–2724, 2019.

[9] I. C. R. Drajana, "Prediksi Jumlah Produksi Coconut Oil Menggunakan k-Nearest Neighbor dan Backward Elimination bagian dari pohon digunakan manusia , sehingga tumbuhan ini dianggap," *Tecnoscienza*, vol. 13, no. 1, pp. 51–64, 2018.

[10] N. K. Suparman, B. A. Dermawan, and T. N. Padilah, "Prediksi Barang Keluar TB. Wijaya Bangunan Menggunakan Algoritma KNN Regression dengan RStudio," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 6, no. 2, pp. 90–97, 2021, doi: 10.14421/jiska.2021.6.2.90-97.

[11] F. D. N. Q. Januar, K. T. B. Artani, and N. W. Utami, "Analisis Dan Prediksi Penutupan Harga Saham Pada Pt Adaro Energy Indonesia Tbk Menggunakan Algoritma K-Nearest Neighbor Regression," *Jurnal Riset Akuntansi*, vol. 13, no. 2, pp. 199–216, 2023.

[12] ori.hhs.gov, "Responsible conduct in data management." Accessed: Jun. 21, 2024. [Online]. Available: https://ori-hhs-gov.translate.goog/education/products/n_illinois_u/datamanagement/dstopic.html?_x_tr_sl=en&_x_tr_tl=id&_x_tr_hl=id&_x_tr_pto=tc

[13] R. Rahmadini, Enjel Erika LorencisLubis, Aji Priansyah, Yolanda R.W.N, and Tuti Meutia, "Penerapan Data Mining Untuk Memprediksi Harga Bahan Pangan Di Indonesia Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Mahasiswa Akuntansi Samudra*, vol. 4, no. 4, pp. 223–235, 2023, doi: 10.33059/jmas.v4i4.7074.

[14] W. W. Pribadi, A. Yunus, and A. Sartika Wiguna, "PERBANDINGAN METODE K-MEANS EUCLIDEAN DISTANCE DAN MANHATTAN DISTANCE PADA PENENTUAN ZONASI COVID-19 DI KABUPATEN MALANG," 2022.

[15] I. Nabillah and I. Ranggadara, "Mean Absolute Percentage Error untuk Evaluasi Hasil Prediksi Komoditas Laut," *JOINS (Journal of Information System)*, vol. 5, no. 2, pp. 250–255, Nov. 2020, doi: 10.33633/joins.v5i2.3900.