Journal of Science and Science Education
Vol. 5 No. 2 November 2021 26–35
ISSN: 2598-3830 (Online)

Comparison between Multiple Linear Regression method and K-Nearest Neighbor method for regression on iris data

Adi Setiawan

Department of Mathematics and Data Science, Faculty of Science and Mathematics, Universitas Kristen Satya Wacana, Salatiga, Jawa Tengah 50711, Indonesia
Corresponding author: adi.setiawan@uksw.edu

ABSTRACT

This study aims to determine the statistics used in regression models such as RMSE, MAPE, MAE and R^2 using the KNN method for regression. The measure of the goodness of the method used is MAPE. The data used is iris data which has been used by many people as an example of data. Variations in the proportion of test data were carried out by 10%, 20%, 30% and 40%. In the proportion of test data of 20%, successively obtained the results that MAPE for case 1, case 2 and case 3 is 5.885 %, 7.778%, 6.979% while in case 4 is 19.341%. As a result, it is obtained that predictions using the KNN method successfully predict/forecast with highly accurate forecasting in case 1, case 2 and case 3 while in case 4 the KNN method predicts with good forecasting.

Keywords: *KNN method, iris data, root mean square of error, mean absolute error, mean absolute percentage error*.

1. Introduction

The Various methods are used in machine learning such as KNN (K-Nearest Neighbor), SVM (Support Vector Machine), Decision Tree, Random Forest, Naïve Bayes, ANN (Jabbar et al., 2021; Jung, 2022). Research on the KNN method for classification has been carried out recently (Arslan & Arslan, 2021; Fu et al., 2019; Hatem, 2022; Nababan et al., 2022; Theerthagiri et al., 2021). However, there are still relatively few studies that discuss the KNN method for regression (Cosenza et al., 2021; Hatipoğlu et al., 2021; Priambodo et al., 2019). Hatipoğlu et al. (2021) in a paper entitled "Prediction of Unemployment Rates in Turkey by k-Nearest Neighbor Regression Analysis", the KNN method for regression is used to predict the unemployment rate in Turkey and in this case the coefficient of determination R^2 = 0.7498. Cosenza et al. (2021) in the paper "Comparison of linear regression, k-nearest neighbor and random forest methods in airborne laser-scanning-based prediction of growing stock", using the KNN method for regression in addition to OLS (Ordinary Linear Regression) and RF (Random Forest) and empirical data, it is concluded that the KNN method is worse than OLS and RF. Furthermore, Bagus Priambodo et al. (2019), in the paper "Predicting GDP of Indonesia Using K-Nearest Neighbor Regression Predicting GDP of Indonesia Using K-Nearest Neighbor Regression" conducted a study to predict GDP (Gross Domestic Product) based on data on rice prices, premium prices, GDP of Japanese country, American GDP, currency exchange rates, Indonesian government consumption, and the value of Indonesia's oil exports in 1980 to 2002 using the KNN method. The results obtained were compared with the Multiple Linear Regression and ANN methods.

In this study, a comparison will be made between the use of the linear regression method and the KNN method for regression on iris data using the MAPE goodness-of-fit measure. Other statistics

such as RMSE (Root Mean Square of Error), MAE (Mean Absolute Error) and coefficient of determination R^2 are also of concern.

2. Materials and Methods

In this article, the KNN method is described in detail while the multiple linear regression analysis used in the regression method as a comparison with the KNN method can be seen further in the literature (Roback & Legler, 2021).

The KNN algorithm for regression is described below.

1. Choose the parameter *k*, namely the number of nearest neighbours.
2. Calculate the distance between the new datum that will be predicted to change the response. In this case, the Euclidean distance is used.
3. Sort the distances obtained in Step 2 in ascending order and determine the smallest distance in the *k*-th order.
4. Determine the corresponding *k*-response variables corresponding to step 3.
5. The prediction of the response variable from the new datum in step 4 is the average of the *k* response variables that have the smallest distance from the new datum.

To give an idea of how the KNN method is for regression, for example, it is known that the training data stated in Table 1.

Table 1. Table of sample data for describing the KNN method for regression.

***Y* (Sepal. length)**	**Sepal. width (X_1)**	**Species**	**Species (X_2 – Numeric)**
5.1	3.5	Setosa	1
4.9	3	Setosa	1
7	3.2	Virginica	2
6.4	3.2	Virginica	2
6.3	3.3	Versicolor	3
5.8	2.7	Versicolor	3

If you have consecutive pairs (Y, X_1, X_2) = (5.9, 3, 3), i.e. a sample that has a sepal length of 5.9, a width of Petal 3 and a Versicolor species and by using k = 4, the prediction of the Y variable will be determined based on the KNN algorithm for regression if given information (X_1, X_2) = (3, 3). To determine the prediction of the Y variable, the following steps are used:

1. Choose k and in this example choose k = 4.
2. Calculate the Euclidean distance between the new datum (X_1, X_2) = (3, 2) which will predict the value of the Y variable with the training data in Table 2.

Table 2. Table of figures for determining the square of the smallest to largest distance with the KNN method for regression.

Y	X_1	X_2	Square of Euclid's distance to new data or test data
5.1	3.5	1	$(3.5-3)^2+(1-2)^2=$ 1.25
4.9	3	1	$(3-3)^2+(1-2)^2=$ 1
7	3.2	2	$(3.2-3)^2+(2-2)^2=$ 0.04
6.4	3.2	2	$(3.2-3)^2+(2-2)^2=$ 0.04
6.3	3.3	3	$(3.3-3)^2+(3-2)^2=$ 1.09
5.8	2.7	3	$(3.3-3)^2+(3-2)^2=$ 1.09

3. Sort the distances in ascending order and determine the smallest distance to the *k*th order with $k = 4$.

Table 3. Table of figures for determining the square of the smallest to largest distance.

Y	X_1	X_2	Square of Euclid's distance to new data or test data	Order from smallest to largest distance
5.1	3.5	1	$(3.5-3)^2+(1-2)^2=$ 1.25	6
4.9	3	1	$(3-3)^2+(1-2)^2=$ 1	5
7	3.2	2	$(3.2-3)^2+(2-2)^2=$ 0.04	1
6.4	3.2	2	$(3.2-3)^2+(2-2)^2=$ 0.04	2
6.3	3.3	3	$(3.3-3)^2+(3-2)^2=$ 1.09	3
5.8	2.7	3	$(3.3-3)^2+(3-2)^2=$ 1.09	4

4. Based on the 3rd step, it is determined whether the associated *Y* value is used in determining the *Y* prediction of new data or not in the KNN method for regression with $k = 4$.

Table 4. Table of figures for determining the square of the smallest distance from 1 to *k*=4.

Y	X_1	X_2	Square of Euclid's distance to new data or test data	Order from smallest to largest distance	Determining whether to enter the KNN method $k = 4$ or not
5.1	3.5	1	$(3.5-3)^2+(1-2)^2=$ 1.25	6	No
4.9	3	1	$(3-3)^2+(1-2)^2=$ 1	5	No
7	3.2	2	$(3.2-3)^2+(2-2)^2=$ 0.04	1	Yes
6.4	3.2	2	$(3.2-3)^2+(2-2)^2=$ 0.04	2	Yes
6.3	3.3	3	$(3.3-3)^2+(3-2)^2=$ 1.09	3	Yes
5.8	2.7	3	$(3.3-3)^2+(3-2)^2=$ 1.09	4	Yes

5. The prediction of the *Y* variable for the new datum or test datum is the average of 7, 6.4, 6.3, 5.8 which is 6.38.

As a result, based on the prediction of the *Y* variable, which is 6.38 and the value of the *Y* variable is 5.9, we get an error = 5.9 – 6.38 = –0.48. Based on the error from the test data, it can be determined MSE, MAE, MAPE and R^2 for the test data in the form of a matrix.

The data used in this study is iris data which has been widely used as an example. This data can be obtained in the basic R program package so there is no need to load a specific program package. Iris data consists of ratio data in the form of Sepal.Length, Sepal.Width, Petal.Length, Petal.Width and Species. In the iris data, there are 3 species, namely Setosa, Versicolor and Virginica so that the book of Jesus provides numerical symbols for the three species, namely 1, 2 and 3. The iris data is

divided into training data and test data so that the training data is relatively more than the test data. . The training data were used to determine the parameters of the KNN method and the test data were used to determine the MSE, MAE, MAPE and R^2 statistics.

In this case, in case 1 the response variable is Sepal.Length, while the independent variables are other variables, namely Sepal.Width, Petal.Length, Petal.Width and Species using the KNN method for regression analysis, while in case 2 the response variable is Sepal.Width, while the response variable is Sepal.Width. The independent variables are other variables, namely Sepal.Length, Petal.Length, Petal.Width and Species. Furthermore, in case 3 the response variable is Petal.Length, while the independent variables are other variables, namely Sepal.Length, Sepal.Width, Petal.Width and Species. Furthermore, in case 4 the response variable is Petal.Width, while the independent variables are other variables, namely Sepal.Length, Sepal.Width, Petal.Length and Species.

In the regression method, the parameters of the regression model are determined based on the training data and based on these parameters the value of the response variable is predicted based on the test data, then used to determine the statistics of MSE, MAE, MAPE and R^2.

3. Results and Discussion

Based on the iris data in case 1, the average results of RMSE, MAPE, MAE and R^2 are presented in Table 1. It can be seen that the MAPE is relatively small, which is less than 10% so that the prediction is considered highly accurate forecasting. Likewise, it can be seen that the average RMSE value is not much different when the proportions of the test data are 10%, 20%, 30% and 40%. Likewise for the average value of MAPE, MAE and R^2.

Table 5. The average results of RMSE, MAPE, MAE and R^2 from the linear regression method and the KNN method for regression in case 1 with variations in the proportions of test data of 10%, 20%, 30% and 40%.

Proportion of Testing Data	RMSE (Reg)	MAPE (Reg)	MAE (Reg)	R^2 (Reg)	RMSE (KNN)	MAPE (KNN)	MAE (KNN)	R^2 (KNN)
10 %	0.3132	4.424 %	0.2579	0.8239	0.4217	5.874 %	0.3425	0.7302
20 %	0.3178	4.447 %	0.2595	0.8401	0.4226	5.885 %	0.3431	0.7296
30 %	0.3196	4.458 %	0.2602	0.8432	0.4215	5.876 %	0.3425	0.7308
40 %	0.3209	4.469 %	0.2607	0.8442	0.4213	5.864 %	0.3421	0.7321

Figure 1 presents a histogram of RMSE, MAPE, MAE and R^2 values if 10 % test data is used. It can be seen in Figure 1 that the histograms of RMSE, MAPE and MAE values tend to be skewed to the right, while the histograms of R^2 values tend to be skewed to the left. This is also supported by the *p*-values of the normality test of the values, namely 0.0000, 0.0000, 0.0000, 0.0000 so that the histogram values are not normally distributed.

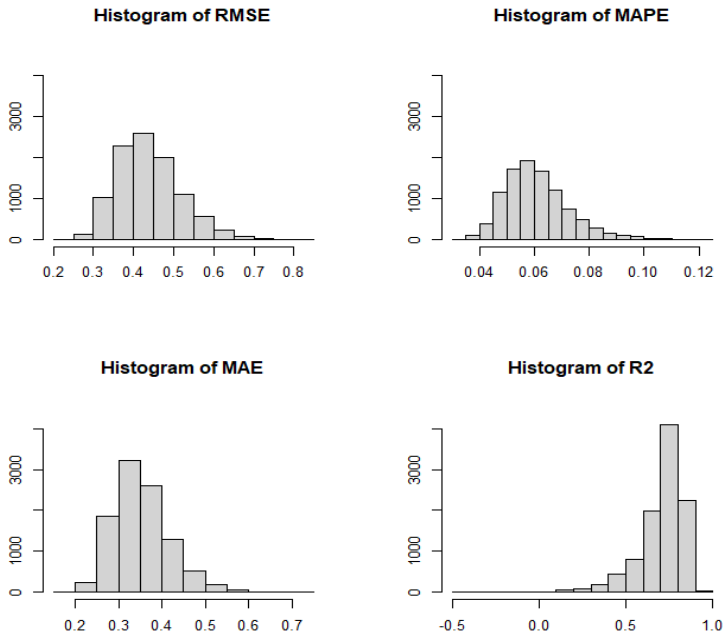

Figure 1. Histogram of RMSE, MAPE, MAE and R^2 values if 10 % test data is used.

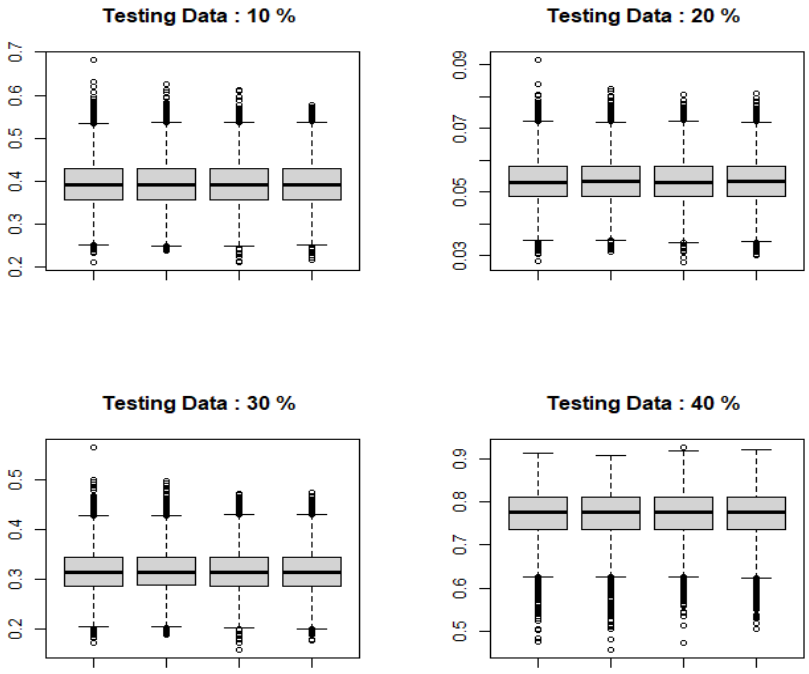

Figure 2. Boxplot comparison of the values of RMSE, MAPE, MAE and R^2 when 10 %, 20 %, 30 % and 40 % test data are used.

Figure 2 presents a Boxplot Comparison of the values of RMSE, MAPE, MAE and R^2 if 10% test data is used. Likewise, a two-sample KS test can be carried out to test whether the RMSE distribution when the proportion of 10% test data is used is the same as when the proportion of 20% of the test data is used. From the test, it was obtained that *p*-value = 0.3305 so that both distributions were the same. The same result can also be done for other couples. It can be seen that there are no significant results when different test data are used.

Table 2, Table 3 and Table 4 respectively state case 2, case 3 and case 4. In case 2, it can be seen that the MAPE value is less than 10% so that the prediction can be said to be good. Likewise, it can be seen that in these cases, the average RMSE values are not significantly different (or not significantly different) when the proportions of the test data are 10%, 20%, 30% and 40%. It can also be found that, with the Wilcoxon test for RMSE values when 10% test data is used, it has the same distribution as the RMSE values when 20% test data is used. Likewise for the average value of MAPE, MAE and R^2. In cases 2 and 3, both methods give highly accurate forecasting results, but in case 4, the multiple linear regression method gives highly accurate forecasting results, while the KNN method for regression gives good forecasting results (Moreno et al., 2013).

Table 6. The average results of RMSE, MAPE, MAE and R^2 from the linear regression method and the KNN method for regression in case 2 with variations in the proportions of test data of 10%, 20%, 30% and 40%.

Proportion of Testing Data	RMSE (Reg)	MAPE (Reg)	MAE (Reg)	R^2 (Reg)	RMSE (KNN)	MAPE (KNN)	MAE (KNN)	R^2 (KNN)
10 %	0.3143	4.957 %	0.2587	0.8238	0.2990	7.793 %	0.2138	0.4975
20 %	0.3177	4.444 %	0.2593	0.8396	0.2992	7.786 %	0.2316	0.4975
30 %	0.3191	4.457 %	0.2599	0.8432	0.2992	7.797 %	0.2320	0.4958
40 %	0.3211	4.476 %	0.2611	0.8443	0.2991	7.792 %	0.2317	0.4984

Table 7. The average results of RMSE, MAPE, MAE and R^2 from the linear regression method and the KNN method for regression in case 3 with variations in the proportions of test data of 10%, 20%, 30% and 40%.

Proportion of Testing Data	RMSE (Reg)	MAPE (Reg)	MAE (Reg)	R^2 (Reg)	RMSE (KNN)	MAPE (KNN)	MAE (KNN)	R^2 (KNN)
10 %	0.3060	8.359 %	0.2455	0.4116	0.2940	6.974 %	0.2270	0.9704
20 %	0.3088	8.344 %	0.2455	0.4541	0.2934	6.979 %	0.2264	0.9706
30 %	0.3113	8.389 %	0.2467	0.4637	0.2936	6.973 %	0.2266	0.9706
40 %	0.3128	8.414 %	0.2473	0.4666	0.2939	6.983 %	0.2268	0.9706

Table 8. The average results of RMSE, MAPE, MAE and R^2 from the linear regression method and the KNN method for regression in case 4 with variations in the proportions of test data of 10%, 20%, 30% and 40%.

Proportion of Testing Data	RMSE (Reg)	MAPE (Reg)	MAE (Reg)	R^2 (Reg)	RMSE (KNN)	MAPE (KNN)	MAE (KNN)	R^2 (KNN)
10 %	0.3038	7.876 %	0.2363	0.9648	0.1831	19.254 %	0.1385	0.9392
20 %	0.3072	7.940 %	0.2367	0.9676	0.1830	19.341 %	0.1386	0.9393
30 %	0.3105	7.996 %	0.2387	0.9677	0.1825	19.273 %	0.1383	0.9395
40 %	0.3124	8.041 %	0.2401	0.9677	0.1828	19.260 %	0.1383	0.9424

Based on the iris data in case 1 using the KNN method for k = 5 and k = 10, the average results of RMSE, MAPE, MAE and R^2 are presented in Table 5. It can be seen that MAPE is relatively small, which is less than 10% so that predictions are considered good. In this case, it can be seen that the average RMSE value is not much different when the proportions of the test data are 10%, 20%, 30% and 40%. Likewise for the average value of MAPE, MAE and R^2.

Table 9. The average results of RMSE, MAPE, MAE and R^2 from the KNN method for regression in case 1 with variations in the proportion of test data 10%, 20%, 30% and 40%.

Proportion of Testing Data	RMSE (k=5)	MAPE (k=5)	MAE (k=5)	R^2 (k=5)	RMSE (k=10)	MAPE (k=10)	MAE (k=10)	R^2 (k=10)
10 %	0.4224	5.887 %	0.3431	0.7295	0.4145	5.726 %	0.3354	0.7438
20 %	0.4212	5.860 %	0.3418	0.7318	0.4148	5.737 %	0.3358	0.7431
30 %	0.4215	5.871 %	0.3422	0.7302	0.4132	5.710 %	0.3344	0.7448
40 %	0.4221	5.880 %	0.3429	0.7297	0.4142	5.729 %	0.3353	0.7435

Table 10. The average results of RMSE, MAPE, MAE and R^2 from the KNN method for regression in case 2 with variations in the proportion of test data 10%, 20%, 30% and 40%.

Proportion of Testing Data	RMSE (k=5)	MAPE (k=5)	MAE (k=5)	R^2 (k=5)	RMSE (k=10)	MAPE (k=10)	MAE (k=10)	R^2 (k=10)
10 %	0.2991	7.757 %	0.2317	0.4986	0.2856	7.464 %	0.2218	0.5460
20 %	0.2990	7.793 %	0.2318	0.4974	0.2856	7.459 %	0.2217	0.5563
30 %	0.2992	7.787 %	0.2316	0.4974	0.2852	7.450 %	0.2215	0.5463
40 %	0.2992	7.797 %	0.2319	0.4959	0.2856	7.449 %	0.2215	0.5458

Table 11. The average results of RMSE, MAPE, MAE and R^2 from the KNN method for regression in case 3 with variations in the proportion of test data 10%, 20%, 30% and 40%.

Proportion of Testing Data	RMSE (k=5)	MAPE (k=5)	MAE (k=5)	R^2 (k=5)	RMSE (k=10)	MAPE (k=10)	MAE (k=10)	R^2 (k=10)
10 %	0.2943	7.001 %	0.2272	0.9704	0.3056	7.005 %	0.2311	0.9685
20 %	0.2943	6.971 %	0.2262	0.9707	0.3043	6.966 %	0.2297	0.9689
30 %	0.2938	6.984 %	0.2268	0.9705	0.3046	6.983 %	0.2303	0.9687
40 %	0.2939	6.986 %	0.2269	0.9706	0.3051	6.982 %	0.2305	0.9687

Table 12. The average results of RMSE, MAPE, MAE and R^2 from the KNN method for regression in case 4 with variations in the proportion of test data 10%, 20%, 30% and 40%.

Proportion of Testing Data	RMSE (k=5)	MAPE (k=5)	MAE (k=5)	R^2 (k=5)	RMSE (k=10)	MAPE (k=10)	MAE (k=10)	R^2 (k=10)
10 %	0.1830	19.310 %	0.1386	0.9392	0.1779	19.149 %	0.1357	0.9426
20 %	0.1831	19.309 %	0.1386	0.9389	0.1778	19.073 %	0.1354	0.9428
30 %	0.1834	19.332 %	0.1386	0.9391	0.1779	19.045 %	0.1356	0.9427
40 %	0.1829	19.266 %	0.1385	0.9392	0.1775	18.989 %	0.1352	0.9431

Figure 3 presents RMSE, MAPE, MAE and R^2 for various values of k where k = 1, 2, ...,15 for p = 10% and case 1. It can be seen that RMSE, MAPE and MAE tend to decrease for an enlarged k and k tends to be flat for a large k, while for R^2 it will tend to be enlarged for an enlarged k. Likewise, in Figure 4 it is presented about RMSE, MAPE, MAE and R^2 for various values of k with k = 1, 2, …., 60 for p = 40 % and case 4. It is seen that RMSE, MAPE and MAE tend to decrease for k increases but for k is large enough it will tend to rise again, while for R^2 it will tend to enlarge for k increases but after k is large enough it will tend to decrease.

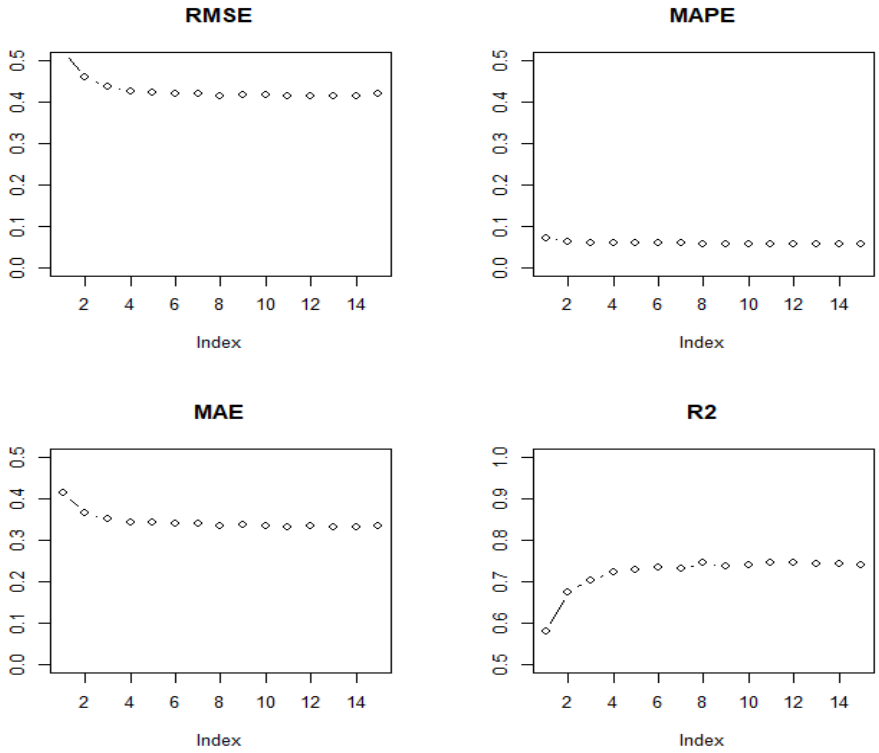

Figure 3. The values of RMSE, MAPE, MAE and R^2 if 10% test data are used for various *k* values with *k* = 1, 2, ….,15 for *p* = 10 % and case 1.

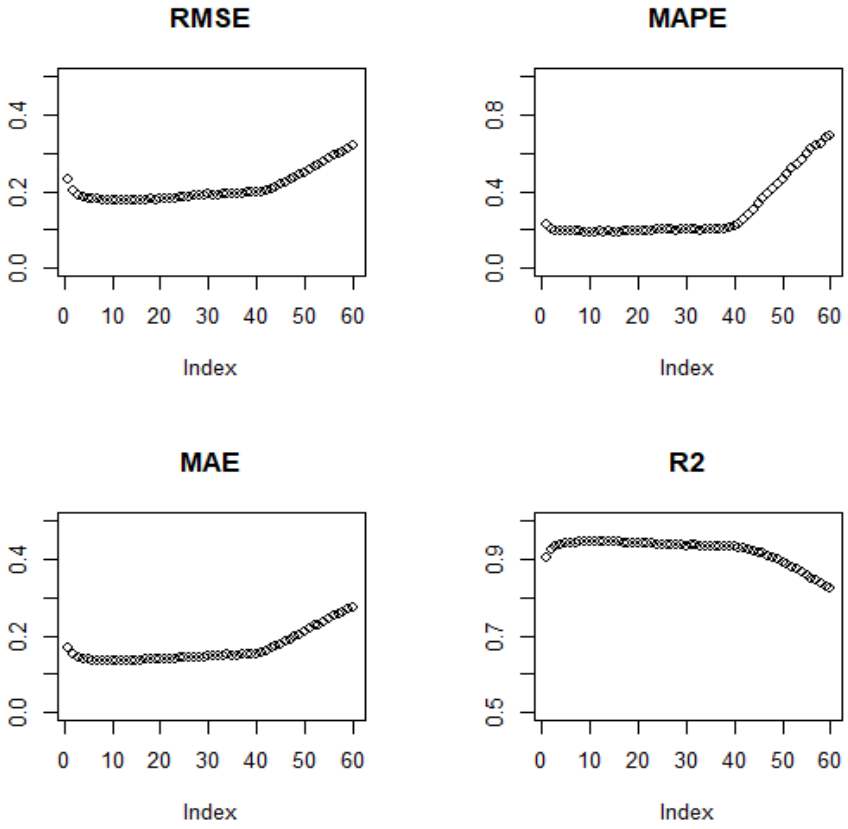

Figure 4. The values of RMSE, MAPE, MAE and R^2 if 10% test data are used for various *k* values with *k* = 1, 2, ….,60 for *p* = 40 % and case 4.

In research (Moreno et al., 2013), the KNN method for regression was used to estimate travel time in Bali by using a good measure of accuracy for regression which was defined as 1 – MAPE so that the smaller the MAPE, the greater the accuracy. In this study, 88.1819% were obtained, so that in MAPE terminology it will be included in good forecasting. On the other hand, in this study, the MAPE goodness-of-fit measure was used. Likewise, this study only uses Euclid's distance, while in the study (Murni et al., 2020), Euclid's distance and L1-normalized distance were used.

4. Conclusion and Remarks

In this study, the following conclusions were obtained:

1. In the iris data with Sepal.Length as the response variable and the proportion of the test data 10%, the linear regression method gives better results, namely with MAPE = 4.424% compared to the KNN method for regression, namely with MAPE = 5.874% so that both can predict with highly accurate forecasting. Analog results were also obtained for case 2 and case 3 and the proportion of test data was 20%, 30% and 40%.
2. In case 4, namely Petal.Width as a response variable and the proportion of test data is 10%, the linear regression method gives better results with MAPE = 7.876% (highly accuracy forecasting) compared to the KNN method for regression, namely MAPE = 19.254% (good forecast)
3. In this study, variations in the value of k in the KNN method were also carried out.

This research was developed by comparing the goodness of the methods that exist in other machine learning with linear regression methods.

References

Arslan, H., & Arslan, H. (2021). A New COVID-19 Detection Method from Human Genome Sequences using CpG Island Features and KNN Classifier. *Engineering Science and Technology, an International Journal*, *24*(4), 839–847. https://doi.org/10.1016/j.jestch.2020.12.026

Cosenza, D. N., Korhonen, L., Maltamo, M., Packalen, P., Strunk, J. L., Næsset, E., Gobakken, T., Soares, P., & Tomé, M. (2021). Comparison of Linear Regression, k-Nearest Neighbour and Random Forest Methods in Airborne Laser-Scanning-based Prediction of Growing Stock. *Forestry: An International Journal of Forest Research*, *94*(2), 311–323. https://doi.org/10.1093/forestry/cpaa034

Fu, Y., He, H. S., Hawbaker, T. J., Henne, P. D., Zhu, Z., & Larsen, D. R. (2019). Evaluating k-Nearest Neighbor (kNN) Imputation Models for Species-Level Aboveground Forest Biomass Mapping in Northeast China. *Remote Sensing*, *11*(17). https://doi.org/10.3390/rs11172005

Hatem, M. Q. (2022). Skin Lesion Classification System using a K-Nearest Neighbor Algorithm. *Visual Computing for Industry, Biomedicine, and Art*, *5*(1). https://doi.org/10.1186/s42492-022-00103-6

Hatipoğlu, Ş., Belgrat, M. A., Degirmenci, A., & Karal, Ö. (2021). Prediction of Unemployment Rates in Turkey by k-Nearest Neighbor Regression Analysis. *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 1–5. https://doi.org/10.1109/ASYU52992.2021.9598980

Jabbar, M. A., Prasad, K. M., Peng, S.-L., Reaz, M. B. I., & Madureira, A. M. (2021). *Machine Learning Methods for Signal, Image and Speech Processing*. River Publishers.

Jung, A. (2022). *Machine Learning: Foundations, Methodologies, and Applications*. Springer.

Moreno, J. J. M., Pol, A. P., Abad, A. S., & Blasco, B. C. (2013). Using the R-MAPE Index as a Resistant Measure of Forecast Accuracy. *Psicothema*, *25*(4), 500–506. https://doi.org/10.7334/psicothema2013.23

Murni, Kosasih, R., Fahrurozi, A., Handhika, T., Sari, I., & Lestari, D. P. (2020). Travel Time Estimation for Destination in Bali Using kNN-Regression Method with Tensorflow. *IOP Conference Series: Materials Science and Engineering*, *854*(1). https://doi.org/10.1088/1757-899X/854/1/012061

Nababan, A. A., Khairi, M., & Harahap, B. S. (2022). Implementation of K-Nearest Neighbors (KNN) Algorithm in Classification of Data Water Quality. *Jurnal Mantik*, *6*(1), 30–35. http://iocscience.org/ejournal/index.php/mantik/article/view/2130

Priambodo, B., Rahayu, S., Hazidar, A. H., Naf'An, E., Masril, M., Handriani, I., Pratama Putra, Z., Kudr Nseaf, A., Setiawan, D., & Jumaryadi, Y. (2019). Predicting GDP of Indonesia using K-Nearest Neighbour Regression. *Journal of Physics: Conference Series*, *1339*(1). https://doi.org/10.1088/1742-6596/1339/1/012040

Roback, P., & Legler, J. (2021). *Beyond Multiple Linear Regression*. CRC Press.

Theerthagiri, P., Jacob, I. J., Ruby, A. U., & Vamsidhar, Y. (2021). Prediction of COVID-19 Possibilities using KNN Classification Algorithm. *International Journal of Current Research and Review*, *13*(6), 156–164.