JAMBURA JOURNAL OF PROBABILITY AND STATISTICS
Volume 4 Nomor 1, Mei 2023
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jps
DOI: https://doi.org/10.34312/jjps.v4i1.18948

PERBANDINGAN KINERJA METODE REGRESI *K-NEAREST NEIGHBOR* DAN METODE REGRESI LINEAR BERGANDA PADA DATA BOSTON HOUSING

Lutfi Sivana Ihzaniah[1], Adi Setiawan [2], Rachel Wulan N. Wijaya[3]
[1,2,3] Program Studi Matematika, Fakultas Sains dan Matematika, Universitas Kristen Satya Wacana, Jl. Diponegoro No. 52-60, Salatiga 50711, Jawa Tengah, Indonesia

e-mail: 662019018@student.uksw.edu

Abstrak
Penelitian ini dibuat guna untuk melihat perbandingan kinerja metode yang lebih baik antara metode regresi KNN (*K-Nearest Neighbor*) dan metode regresi linear berganda pada data Boston Housing. Kinerja metode yang dimaksudkan di sini adalah MAE, RMSE, MAPE dan R^2. Metode KNN merupakan metode memprediksi sesuatu berdasarkan contoh pelatihan terdekat dari suatu objek. Sedangkan regresi linear berganda merupakan teknik peramalan dengan melibatkan lebih dari satu variabel bebas. Perbandingan kedua metode tersebut didasarkan dari hasil ukuran kebaikan *Mean Absolute Percent Error* (MAPE). Pada penelitian ini definisi jarak yang digunakan adalah jarak *Euclidean* dan jarak *Minkowski*. Dalam penelitian ini digunakan persentase data uji sebesar 20%, 30%, dan 40% untuk kedua metode tersebut. Nilai K pada metode KNN mendefinisikan banyak tetangga terdekat yang akan diperiksa untuk menentukan nilai suatu variabel terikat, pada penelitian ini menggunakan nilai K dari 1 sampai 10 untuk setiap data uji dan definisi jarak. Perolehan nilai MAPE terbaik metode regresi KNN adalah 12,89% pada saat K = 3 untuk jarak *Euclidean* dan 13,22% pada saat K = 3 untuk jarak *Minkowski* sedangkan hasil nilai MAPE terbaik untuk metode regresi linear berganda yaitu sebesar 17,17%. Metode yang terbaik antara kedua metode tersebut adalah metode regresi KNN dilihat dari perolehan nilai MAPE metode regresi KNN yang lebih kecil dibandingkan dengan nilai MAPE metode regresi linear berganda.
Kata Kunci: *Regresi KNN, Regresi Linear Berganda, Mean Absolute Percent Error (MAPE)*

Abstract
This research was made in order to see which method performance is better between the KNN (K-Nearest Neighbor) regression method and the multiple linear regression method on Boston Housing data. The method performace referred here is MAE, RMSE, MAPE, and R^2. The KNN method is a method to predict something based on the closest training examples of an object. Meanwhile, multiple linear regression is a forecasting technique involving more than one independent variable. The comparison of the two methods is based on the results of the Mean Absolute Percent Error (MAPE). In this research the definitions of distance used are Euclidean distance and Minkowski distance. The K value in the KNN method defines the number of nearest neighbors to be examined to determine the value of a dependent variable, in this research we use K values from 1 to 10 for each test data and definition of distance. In this research, the percentage of test data used was 20%, 30%, and 40% for both methods. The best MAPE value obtained by the KNN regression method was 12,89% at K = 3 for Euclidean distance and 13,22% at K = 3 for Minkowski distance. Meanwhile the best MAPE value for the multiple linear regression method is 17,17%. The best method between the two methods is the KNN regression method as seen from the MAPE value of the KNN regression method which is smaller than the MAPE value of the multiple linear regression method.
Keywords*: KNN Regression, Multiple Linear Regression, Mean Absolute Percent Error (MAPE)*

Received: 2023-02-22 | Accepted: 2023-05-31 | Available Online: 2023-05-31

1. PENDAHULUAN

Rumah merupakan kebutuhan primer bagi masyarakat, rumah juga dapat dijadikan bahan investasi di masa depan (Saiful et al., 2021). Penentuan harga rumah harus diperhitungkan baik kepada penjual maupun pembeli. Penentuan harga ini bisa berdasarkan berbagai spesifikasi rumah (Labib et al., 2021). Memperkirakan harga rumah sangat penting bagi calon pemilik rumah, pengembang maupun pelaku pasar *real estate* lainnya (Begum et al., 2022).

Untuk memperkirakan harga rumah dapat menggunakan beberapa metode peramalan. Peramalan adalah kegiatan memperkirakan apa yang akan terjadi di masa depan. Peramalan atau prediksi banyak digunakan sebagai alat atau pertimbangan dalam mengambil keputusan pembelian (Ayuni & Fitrianah, 2019). Untuk melakukan prediksi dapat memanfaatkan beberapa metode seperti metode SVM (*Support Vector Machine*) untuk regresi, D*ecision Tree* untuk regresi, ANN (*Artificial Neural Network*) untuk regresi dan lain-lain. Namun demikian pada penelitian ini akan menggunakan metode regresi KNN dan metode regresi linear berganda. Metode KNN dipilih karena lebih sederhana dibanding metode lainnya, dapat diterapkan pada data yang besar, dan memiliki akurasi yang tinggi (Permana et al., 2021). Sedangkan metode regresi linear dipilih karena algoritmanya sederhana sehingga cenderung lebih efisien dan memiliki akurasi yang tinggi (Utomo et al., 2019). Untuk mendapatkan prediksi harga rumah, perlu dilakukan perbandingan beberapa algoritma model. Tujuannya adalah untuk menghasilkan kesalahan minimum dan akurasi yang tinggi (Pathak & Chaudhari, 2021).

Penelitian sebelumnya mengenai perbandingan metode regresi KNN dan regresi linear berganda dijelaskan berikut ini. Pada penelitian Priambodo et al (2019) yang membandingkan metode regresi KNN dengan metode *neural network* dan metode regresi linear berganda dalam memprediksi GDP di Indonesia, mendapatkan hasil bahwa metode regresi KNN lebih baik dari pada metode *neural network* dan metode regresi linear berganda yang ditinjau dari nilai MAPE regresi KNN sebesar 6,53% untuk K = 3 dan nilai MAPE regresi linear berganda sebesar 91,59%. Penelitian oleh Al-Dosary et al (2019) yang memprediksi penggunaan bahan bakar traktor yang menggunakan metode regresi KNN dan regresi linear berganda juga menunjukkan bahwa regresi KNN lebih baik dari pada regresi linear berganda dilihat dari nilai MAE regresi linear berganda sebesar 3,12 dan nilai MAE regresi KNN sebesar 1,78 untuk K = 5. Kedua penelitian tersebut memperoleh hasil yang sama dengan penelitian ini. Sedangkan penelitian oleh Utomo et al (2019) yang memprediksi harga emas menggunakan metode KNN dan regresi linear menunjukkan regresi linear lebih baik dari pada metode KNN dengan perolehan RMSE regresi linear sebesar 0,05807 dan KNN sebesar 1,29292. Hal itu menunjukkan perolehan hasil penelitian oleh Utomo et al (2019) berbeda dengan penelitian ini. Namun demikian, penelitian-penelitian tersebut belum melakukan perbandingan kinerja metode yang digunakan dan hanya menggunakan salah satu dari beberapa definisi jarak. Dalam penelitian ini, dilakukan perbandingan kinerja metode regresi KNN dan regresi linear ganda pada data Boston Housing. Dalam hal ini, kinerja yang dimaksudkan adalah MAE, RMSE, MAPE dan R^2. Melihat penelitian-penelitian sebelumnya penulis tertarik untuk membandingkan kinerja metode regresi KNN dan regresi linear berganda menggunakan data Boston Housing.

2. METODE PENELITIAN

Penelitian ini memprediksi median harga rumah di Boston menggunakan data Boston Housing yang memiliki 506 baris, 13 atribut kontinu dan 1 atribut kategoris (Lydia et al., 2019). Dalam memprediksi menggunakan metode regresi KNN dan metode regresi linear

berganda perlu menentukan proporsi data latih (*training data*) dan data uji (*testing data*). Perhitungan kesalahan prediksi difokuskan pada ukuran kebaikan MAPE dan didukung ukuran kebaikan lain seperti RMSE, MAE, dan R-Squared. Proses pengolahan data dilakukan menggunakan *software* RStudio.

2.1 Regresi Linear Berganda

Regresi linear berganda merupakan teknik peramalan dengan melibatkan lebih dari satu variabel bebas (Mauladi et al., 2020). Pada regresi linear berganda memerlukan model regresi (Ayu et al., 2022), model regresinya sebagai berikut :

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (1)$$

dengan

Y : variabel terikat,

$x_1, x_2, \ldots, x_p$: variabel bebas,

α : konstanta,

$\beta_1, \beta_2, \ldots, \beta_p$: nilai koefisien regresi.

Untuk mendapatkan nilai koefisien regresi dapat menggunakan persamaan contoh berikut dengan 2 variabel:

$$\begin{aligned} &\alpha n + \beta_1 \sum x_1 + \beta_2 \sum x_2 = \sum Y \\ &\alpha \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_2 x_1 = \sum x_1 Y \\ &\alpha \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2 = \sum x_2 Y \end{aligned} \quad (2)$$

Melalui persamaan (2) dapat diketahui koefisien regresi menggunakan metode eliminasi setelah itu masukkan nilai β_1, β_2 ke model regresi. Persamaan regresi itu nanti yang akan digunakan untuk memprediksi variabel tak bebas (Ayu et al., 2022).

Untuk memperjelas bagaimana algoritma regresi linear berganda itu bekerja diberikan contoh berikut ini. Misalkan disajikan cuplikan sederhana data Boston Housing pada Tabel 1 dengan 2 variabel bebas (CRIM dan NOX) dan 1 variabel tak bebas (MEDV). CRIM sebagai x_1, NOX sebagai x_2, dan MEDV sebagai Y. Diambil 90% data latih yaitu data ke-1 sampai data ke-9 dan 10% data uji yaitu data ke-10. Disini akan dicari prediksi nilai MEDV untuk data ke-10. Mengacu pada persamaan (2) langkah pertama adalah mencari koefisien regresi dengan persamaan berikut:

$$\begin{aligned} &9\alpha + \beta_1(0{,}63627) + \beta_2(4{,}422) = 245{,}1 \\ &\alpha(0{,}63627) + \beta_1(0{,}081549) + \beta_2(0{,}321827) = 15{,}550608 \\ &\alpha(4{,}422) + \beta_1(0{,}321827) + \beta_2(2{,}182386) = 119{,}1841 \end{aligned}$$

Dengan metode matriks ditentukan koefisien regresi dengan cara:

$$\begin{bmatrix} 9 & 0{,}63627 & 4{,}422 \\ 0{,}63627 & 0{,}081549 & 0{,}321827 \\ 4{,}422 & 0{,}321827 & 2{,}182386 \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 245{,}1 \\ 15{,}550608 \\ 119{,}1841 \end{bmatrix}$$

Hitung determinannya,

$$M_1 = \begin{bmatrix} 245{,}1 & 0{,}63627 & 4{,}422 \\ 15{,}550608 & 0{,}081549 & 0{,}321827 \\ 119{,}1841 & 0{,}321827 & 2{,}182386 \end{bmatrix} \quad M_2 = \begin{bmatrix} 9 & 245{,}1 & 4{,}422 \\ 0{,}63627 & 15{,}550608 & 0{,}321827 \\ 4{,}422 & 119{,}1841 & 2{,}182386 \end{bmatrix}$$

$$M_3 = \begin{bmatrix} 9 & 0{,}63627 & 245{,}1 \\ 0{,}63627 & 0{,}081549 & 15{,}550608 \\ 4{,}422 & 0{,}321827 & 119{,}1841 \end{bmatrix} \quad M = \begin{bmatrix} 9 & 0{,}63627 & 4{,}422 \\ 0{,}63627 & 0{,}081549 & 0{,}321827 \\ 4{,}422 & 0{,}321827 & 2{,}182386 \end{bmatrix}$$

$$\alpha = \frac{\det M_1}{\det M} = \frac{0.1983925514595320}{0.0024327794379204} = 81{,}550$$

$$\beta_1 = \frac{\det M_2}{\det M} = \frac{-0.0524195256060003}{0.0024327794379204} = -21.547$$

$$\beta_2 = \frac{\det M_3}{\det M} = \frac{-0.2613988582867880}{0.0024327794379204} = -107.449$$

Masukkan nilai-nilai koefisien regresi tersebut ke persamaan regresi sehingga menjadi:

$$81{,}550 - 21.547x_1 - 107.449x_2 = Y$$

Masukkan nilai $x_1$dan x_2 data ke-10 pada Tabel 1 ke dalam persamaan regresi yang terlah didapatkan diatas.

$$99 - 21.547(0{,}17004) - 107.449(0{,}524) = Y$$

$$21.583 = Y$$

Jadi, prediksi nilai MEDV untuk data ke-10 adalah sebesar **21,583**. Untuk menguji kinerja metode ini dapat menggunakan MAPE dengan rumus seperti persamaan (5) atau beberapa uji kebaikan yang lain.

2.2 Regresi KNN

Metode KNN merupakan metode memprediksi sesuatu berdasarkan contoh pelatihan terdekat dalam ruang fitur. Pada regresi KNN perlu dicari nilai K (tetangga terdekat) dan definisi jarak yang ingin digunakan (Priambodo et al., 2019). Algoritma Regresi KNN sebagai berikut:

1. Pilih parameter K (banyaknya tetangga terdekat),
2. Hitung jarak data yang akan ditentukan prediksinya dengan data latih,
3. Urutkan jarak yang diperoleh dari langkah 2 dalam urutan naik. Ambil titik data terkecil sejumlah K titik (Setiawan, 2022),
4. Hitung rata-rata sejumlah K titik data terdekat tersebut (Akbar & Kusumodestoni, 2020).

Beberapa definisi jarak yang digunakan pada regresi KNN adalah sebagai berikut:

1. Jarak Euclidean

Jarak *Euclidean* merupakan pengukuran jarak antara dua titik yang digambarkan dalam garis lurus (Santoso & Kusumaningsih, 2018). Rumus jarak Euclidean dapat dilihat pada persamaan (3).

$$d(x,y) = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2} \quad (3)$$

2. Jarak Minkowski

Jarak *Minkowski* adalah bentuk umum dari jarak *Euclidean* dan jarak *Manhattan* (Azwar et al., 2021). Rumus jarak *Minkowski* dapat dilihat pada persamaan (4).

$$d(x,y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^{\frac{1}{p}}} \quad (4)$$

Dalam melakukan prediksi tentu perlu menghitung kesalahan prediksinya. Terdapat beberapa cara yang digunakan yaitu:

1. Mean Absolute Percent Error (MAPE)

Mean Absolute Percent Error (MAPE) merupakan rata-rata absolut antara nilai peramalan dan nilai aktual yang dinyatakan sebagai persentase (Putro et al., 2018). Rumus nilai MAPE dapat dilihat pada persamaan (5).

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \frac{|A_t - F_t|}{A_t} \quad (5)$$

2. RMSE

RMSE atau *Root Mean Squared Error* adalah akar kuadrat dari rata-rata kuadrat nilai aktual dan nilai prediksi (Tatachar, 2021). Rumus RMSE dapat dilihat pada persamaan (6).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}} \quad (6)$$

3. MAE

Mean Absolute Error (MAE) adalah rata-rata nilai aktual dan nilai prediksi yang bernilai mutlak positif (Azmi et al., 2020). Rumus MAE dapat dilihat pada persamaan (7).

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |A_t - F_t| \quad (7)$$

4. R-Squared

R-Squared (R^2) atau disebut koefisien determinasi berfungsi untuk mengetahui hubungan antara variabel bebas dan variabel tak bebas secara simultan. R-Squared berkisar antara 0 sampai 1 (Budilaksana et al., 2021). Rumus R-Squared dapat dilihat pada persamaan (8).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (8)$$

Untuk memperjelas bagaimana algoritma regresi KNN itu bekerja diberikan contoh berikut ini. Misalkan disajikan cuplikan sederhana data Boston Housing pada Tabel 1 dengan 2 variabel bebas (CRIM dan NOX) dan 1 variabel tak bebas (MEDV). Diambil 10% data uji yaitu data ke-10 dan 90% data latih yaitu data ke-1 sampai ke-9. Contoh ini menggunakan definisi jarak *Euclidean* dengan K = 3. Setelah persiapan data dilakukan, dihitung jarak data

uji dan data latih. Cara menghitung jarak *Euclid* menggunakan persamaan (3) untuk data ke-10 dan data ke-1 yaitu:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$d(x,y) = \sqrt{(0.17004 - 0.00632)^2 + (0.524 - 0.538)^2}$$

$$= 0.164317$$

Dari perhitungan diatas jarak antara data uji dan data latih ke-1 adalah sebesar 0,164317. Hitung seterusnya sampai data latih ke-9. Langkah selanjutnya adalah mengurutkan jarak terkecil sejumlah K data. Dari Tabel 1 sejumlah K = 3 data terkecil yaitu data ke 8, 9, dan 7. Langkah terakhir adalah menghitung rata-rata sejumlah K data dengan cara Pred MEDV = $\frac{27,1+16,5+22,9}{3} = \mathbf{22,17}$. Jadi, dengan K = 3 dapat dihasilkan nilai prediksi MEDV sebesar **22,17.** Untuk menguji kinerja metode ini dapat menggunakan MAPE dengan rumus seperti persamaan (5) atau beberapa uji kebaikan yang lain.

Tabel 1. Data contoh

NO	CRIM	NOX	MEDV	*Euclid* (KNN)
1	0.00632	0.538	24	0.164317
2	0.02731	0.469	21.6	0.15296
3	0.02729	0.469	34.7	0.152979
4	0.03237	0.458	33.4	0.152673
5	0.06905	0.458	36.2	0.120644
6	0.02985	0.458	28.7	0.154949
7	0.08829	0.524	22.9	0.08175
8	0.14455	0.524	27.1	0.02549
9	0.21124	0.524	16.5	0.0412
10	**0.17004**	**0.524**	**18.9**	

3. HASIL DAN PEMBAHASAN

Dari data Boston Housing ditentukan variabel independennya (x) yaitu: CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, BLACK, LSTAT sedangkan variabel dependennya (y) yaitu MEDV. Pada penelitian ini dipilih proporsi data uji sebesar 20%, 30%, dan 40% dengan 100 kali simulasi. Pengolahan data dilakukan menggunakan *software* RStudio.

3.1 Regresi Linear Berganda

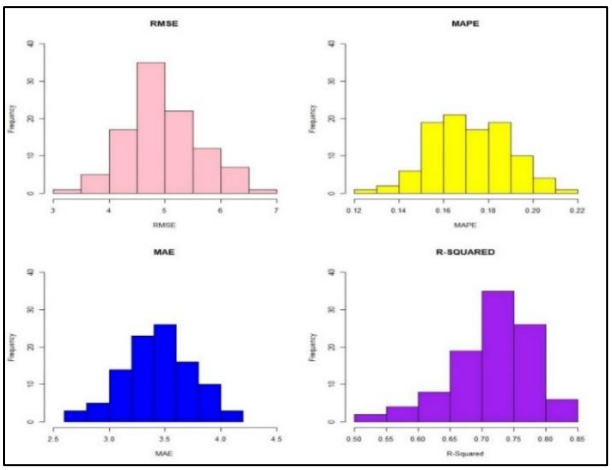
Penelitian ini menggunakan metode regresi linear berganda pada data Boston Housing untuk memprediksi median harga rumah (MEDV) di Boston. Dilakukan 100 kali simulasi menggunakan proporsi data uji sebesar 20%, 30%, dan 40% yang dipilih secara random. Pada penelitian metode regresi linear berganda berturut-turut menghasilkan nilai MAPE sebesar 17,17%; 17,45%; 17,35%. Didapat pula nilai RMSE, MAE, dan R-Squared seperti pada Tabel 2.

Dari Tabel 2 tingkat akurasi terbaik didapat pada data uji 20%. Dilihat dari Tabel 4 pula semakin besar proporsi data uji, akurasinya semakin menurun. Semakin kecil nilai MAPE maka RMSE dan MAE akan semakin mengecil sedangkan R-Squared semakin membesar mendekati 1.

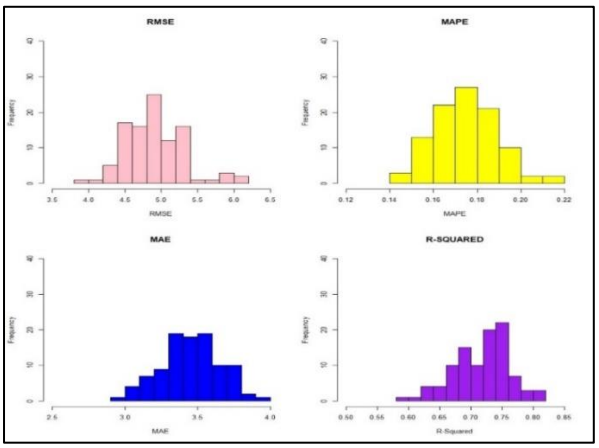
Histogram RMSE, MAPE, MAE, R-Squared untuk data uji 20%, 30%, dan 40% secara berturut-turut terlihat pada Gambar 1, Gambar 2, dan Gambar 3. Melihat perolehan nilai RMSE, MAPE, MAE, dan R-Squared pada histogram-histogram dapat diketahui histogram RMSE, MAPE, dan MAE cenderung simetris sehingga data cenderung berdistribusi normal. Sedangkan histogram R-Squared cenderung miring ke kiri sehingga datanya memiliki kecenderungan tidak berdistribusi normal.

Tabel 2. Nilai-nilai ukuran kebaikan metode regresi linear berganda

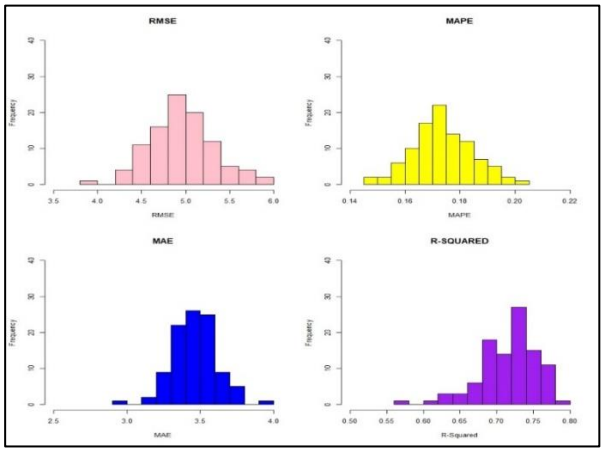
Data Uji	RMSE	MAPE	MAE	R-Squared
20%	4.94	17.17%	3.43	0.72
30%	4.91	17.45%	3.45	0.72
40%	4.97	17.35%	3.46	0.72



Gambar 1. Histogram dengan data uji 20%



Gambar 2. Histogram dengan data uji 30%



Gambar 3. Histogram dengan data uji 40%

3.2 Regresi KNN

Pada penelitian ini akan dilakukan menggunakan metode regresi KNN untuk memprediksi median harga rumah (MEDV) di Boston dengan 100 kali simulasi. Perhitungan *error* difokuskan menggunakan ukuran kebaikan *Mean Absolute Percent Error* (MAPE). Dipilih nilai K = 1 sampai K = 10 dengan proporsi data uji sebesar 20%, 30%, dan 40% yang dipilih secara random.

a. Jarak Euclidean

Dihasilkan nilai-nilai akurasi perhitungan regresi KNN menggunakan jarak *Euclidean* secara berturut-turut tertera pada Tabel 3 dan Tabel 4. Berdasarkan tabel-tabel tersebut terlihat bahwa pada proporsi data uji 20% nilai MAPE terkecil pada saat K = 3 yaitu sebesar 12,89%, pada proporsi data uji 30% nilai MAPE terkecil pada saat K = 3 yaitu sebesar 13,31%, dan pada proporsi data uji 40% nilai MAPE terkecil terjadi pada saat K = 3 yaitu sebesar 13,78%.

Pada keseluruhan tabel nilai akurasi jarak *Euclidean* memiliki kecenderungan bahwa semakin kecil nilai MAPE maka nilai RMSE dan MAE akan semakin mengecil sedangkan R-Squared akan semakin membesar mendekati 1.

Tabel 3. Nilai-nilai ukuran kebaikan dengan data uji 20% & 30%

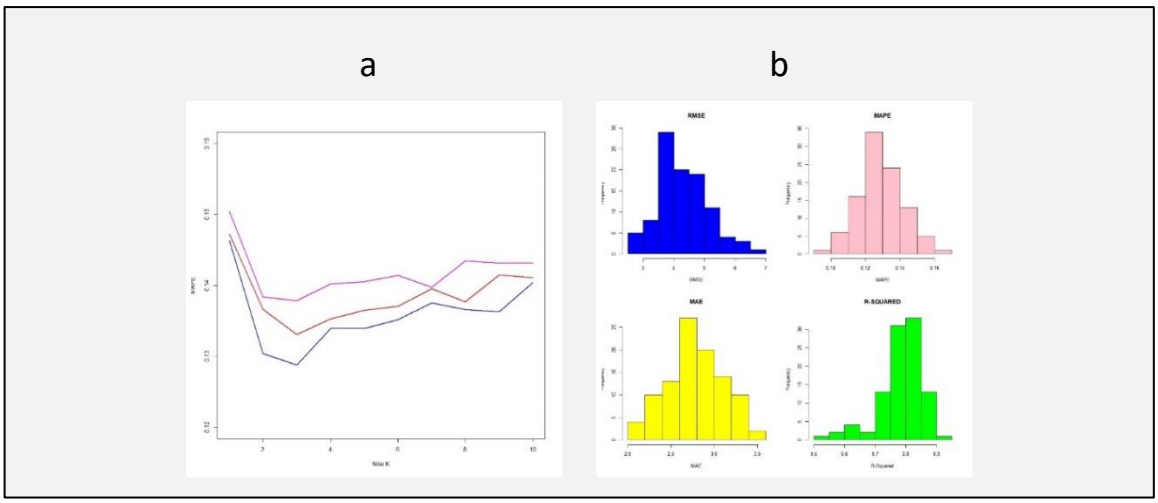
Nilai K	***Euclid* (20%)**				***Euclid* (30%)**			
	RMSE	MAPE	MAE	R-Squared	RMSE	MAPE	MAE	R-Squared
1	4.74	14.64%	3.01	0.74	4.74	14.73%	3.04	0.75
2	4.24	13.04%	2.74	0.79	4.46	13.67%	2.85	0.77
3	4.31	12.89%	2.78	0.79	4.41	13.31%	2.84	0.78
4	4.46	13.39%	2.86	0.77	4.58	13.53%	2.91	0.76
5	4.61	13.39%	2.90	0.76	4.74	13.65%	2.98	0.75
6	4.52	13.52%	2.89	0.76	4.64	13.71%	2.94	0.76
7	4.73	13.75%	2.99	0.75	4.79	13.95%	3.05	0.75
8	4.67	13.66%	2.99	0.76	4.63	13.77%	2.97	0.76
9	4.81	13.63%	3.05	0.76	4.81	14.15%	3.07	0.76
10	4.72	14.04%	3.04	0.77	4.71	14.11%	3.04	0.76

Tabel 4. Nilai-nilai ukuran kebaikan dengan data uji 40%

Nilai K	*Euclid* (40%)			
	RMSE	MAPE	MAE	R-Squared
1	5.02	15.04%	3.17	0.72
2	4.66	13.84%	2.95	0.75
3	4.72	13.78%	3.00	0.75
4	4.86	14.02%	3.05	0.73
5	4.81	14.05%	3.04	0.74
6	4.86	14.14%	3.08	0.74
7	4.83	13.97%	3.06	0.74
8	4.81	14.35%	3.11	0.75
9	4.74	14.31%	3.08	0.75
10	4.76	14.32%	3.10	0.75

Dihasilkan kurva variasi K dari nilai MAPE untuk data uji 20%, 30%, dan 40% menggunakan definisi jarak *Euclidean* yang terlihat pada Gambar 4 (a). Dari Gambar 4 (a) dapat dilihat bahwa apabila nilai K semakin besar maka kurva akan memiliki kecenderungan naik yang artinya semakin besar nilai K maka semakin buruk tingkat akurasinya. Begitupun dengan data ujinya jika semakin besar nilai akurasi cenderung semakin memburuk.

Gambar 4 (b) merupakan histogram nilai-nilai akurasi RMSE, MAPE, MAE, dan R-Squared dengan definisi jarak *Euclidean* menggunakan proporsi data uji 20% dan K = 3 dengan 100 kali simulasi. Diambil nilai pemusatan rata-rata sehingga diperoleh nilai rata-rata RMSE, MAPE, MAE, dan R-Squared berturut-turut sebesar 4,31; 12,89%; 2,78; dan 0,79. Histogram MAPE dan MAE cenderung simetris sehingga cenderung berdistribusi normal, sedangkan histogram RMSE miring ke kanan dan histogram R-Squared miring ke kiri sehingga cenderung tidak berdistribusi normal.



Gambar 4. (a) Kurva variasi nilai K untuk data uji 20% (biru), 30% merah, 40% (ungu), (b) Histogram regresi KNN jarak *Euclidean*

b. Jarak Minkowski

Dilakukan perhitungan regresi KNN menggunakan jarak *Minkowski* dengan data uji 20%, 30%, dan 40% yang dipilih secara random sebanyak 100 kali simulasi. Dipilih nilai K = 1 sampai K = 10 dan nilai p mengikuti nilai K yaitu p = 1 sampai p = 10. Perolehan nilai

MAPE terkecil pada proporsi data uji 20% yaitu pada saat K = 3 sebesar 13,22%, pada proporsi data uji 30% pada saat K = 2 sebesar 13,45%, dan pada proporsi data uji 40% pada saat K = 2 sebesar 13,93% tertera berturut-turut pada Tabel 5 dan Tabel 6.

Pada keseluruhan tabel nilai akurasi jarak *Minkowski* memiliki kecenderungan bahwa semakin kecil nilai MAPE maka nilai RMSE dan MAE akan semakin mengecil sedangkan R-Squared akan semakin membesar mendekati 1.

Tabel 5. Nilai-nilai akurasi dengan data uji 20% & 30%

Nilai K	***Minkowski* (20%)**				***Minkowski* (30%)**			
	RMSE	MAPE	MAE	R-Squared	RMSE	MAPE	MAE	R-Squared
1	4.69	14.56%	3.03	0.75	4.76	14.73%	3.08	0.75
2	4.44	13.50%	2.83	0.77	4.44	13.45%	2.84	0.77
3	4.37	13.22%	2.81	0.78	4.61	13.62%	2.92	0.76
4	4.54	13.40%	2.89	0.76	4.71	13.88%	3.00	0.75
5	4.62	13.95%	2.99	0.75	4.83	14.45%	3.09	0.74
6	4.78	14.33%	3.05	0.73	4.90	14.62%	3.14	0.73
7	4.73	14.57%	3.09	0.74	4.94	14.80%	3.18	0.73
8	4.86	14.68%	3.13	0.74	5.05	15.28%	3.28	0.72
9	4.97	15.22%	3.23	0.72	5.09	15.51%	3.30	0.71
10	5.06	15.20%	3.28	0.71	5.05	15.79%	3.32	0.71

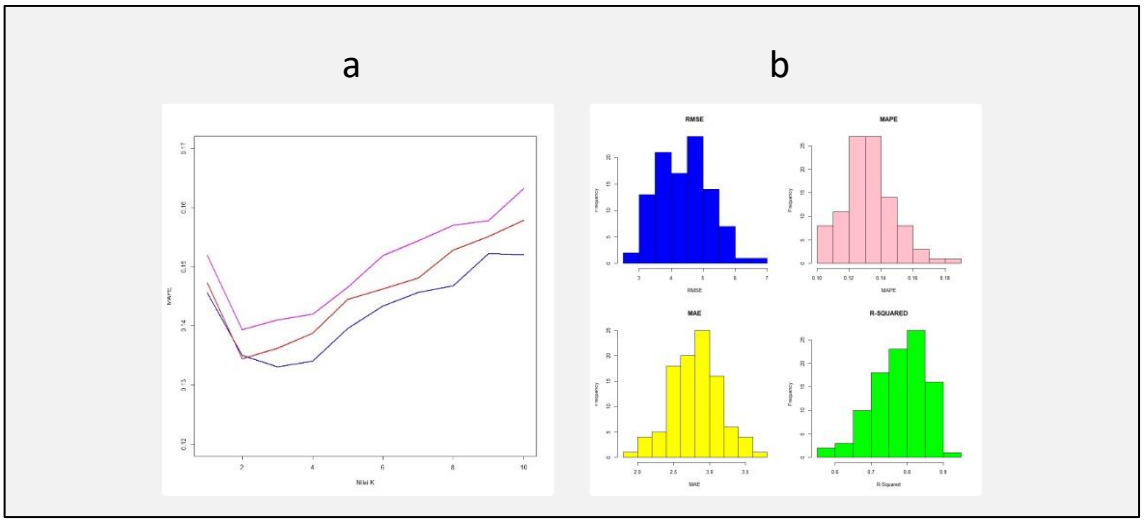
Tabel 6. Nilai-nilai akurasi dengan data uji 40%

Nilai K	***Minkowski* (40%)**			
	RMSE	MAPE	MAE	R-Squared
1	4.84	15.19%	3.14	0.74
2	4.59	13.93%	2.95	0.76
3	4.73	14.10%	3.01	0.74
4	4.79	14.20%	3.07	0.73
5	4.85	14.65%	3.13	0.72
6	5.06	15.19%	3.27	0.71
7	5.11	15.44%	3.31	0.71
8	5.21	15.70%	3.37	0.70
9	5.21	15.78%	3.39	0.70
10	5.37	16.32%	3.49	0.69

Dihasilkan kurva variasi K dari nilai MAPE untuk data uji 20%, 30%, dan 40% menggunakan definisi jarak *Minkowski* yang terlihat pada Gambar 5 (a). Berdasarkan kurva Gambar 5 (a) dapat dilihat bahwa apabila nilai K semakin besar maka nilai MAPE akan cenderung naik yang artinya akurasinya semakin buruk. Semakin besar data uji akurasinya juga cenderung memburuk.

Gambar 5 (b) merupakan histogram nilai-nilai akurasi RMSE, MAPE, MAE dan R-Squared dengan definisi jarak *Minkowski* menggunakan data uji 20% dan K = 3 dengan 100 kali simulasi. Diambil nilai pemusatan rata-rata sehingga diperoleh nilai rata-rata RMSE, MAPE, MAE dan R-Squared berturut-turut sebesar 4,37; 13,22%; 2,81; dan 0,78. Pada histogram RMSE dan MAPE cenderung simetris sehingga cenderung berdistribusi normal, sedangkan histogram MAE dan R-Squared cenderung miring ke kiri sehingga cenderung tidak berdistribusi normal. Dalam hal ini, histogram-histogram dari nilai-nilai akurasi

RMSE, MAPE, MAE dan R-Squared cukup halus (*smooth*) baik untuk metode regresi linear ganda maupun metode KNN.



Gambar 5. (a) Kurva variasi nilai K untuk data uji 20% (biru), 30% merah, 40% (ungu), (b) Histogram regresi KNN jarak *Minkowski*

Pada percobaan regresi linear berganda tingkat akurasi terbaik pada data uji 20%. Pada percobaan regresi KNN untuk definisi jarak *Euclidean* dan jarak *Minkowski* tingkat akurasi terbaik juga pada data uji 20%. Dengan uji Kolmogorov-Smirnov hasil-hasil nilai MAPE untuk kedua metode memiliki p-*value* kurang dari 5% sehingga hasil-hasil nilai MAPE kedua metode memiliki perbedaan yang signifikan. Perolehan nilai-nilai MAPE pada metode regresi linear berganda dan regresi KNN termasuk dalam kriteria nilai MAPE yang baik karena terletak pada rentang 10%-20%. Pada penelitian ini juga dilakukan percobaan menggunakan data uji 10% dan percobaan 1000 kali simulasi untuk seluruh definisi jarak pada metode regresi KNN namun terjadi *error missing values* pada data uji, untuk itu perlu dilakukan penelitian lebih lanjut.

4. KESIMPULAN

Dari penelitian yang sudah dilakukan, berdasarkan nilai MAPE dapat ditarik kesimpulan bahwa metode regresi KNN lebih baik dari pada metode regresi linear berganda dalam memprediksi data Boston Housing. Hal tersebut dapat dilihat dari nilai MAPE regresi KNN untuk keseluruhan definisi jarak dan variasi K yang mana lebih kecil dari nilai MAPE regresi linear berganda.

DAFTAR PUSTAKA

Akbar, A. S., & Kusumodestoni, R. H. (2020). Optimization of K Value and Lag Parameter of K-Nearest Neighbor Algorithm on the Prediction of Hotel Occupancy Rates. *Jurnal Teknologi Dan Sistem Komputer*, *8*(3), 246–254. https://doi.org/10.14710/jtsiskom.2020.13648

Al-Dosary, N. M. N., Al-Hamed, S. A., & Mohamed Aboukarima, A. (2019). K-Nearest Neighbors Method for Prediction of Fuel Consumption in Tractor-Chisel Plow Systems. *Engenharia Agricola*, *39*(6), 729–736. https://doi.org/10.1590/1809-4430-Eng.Agric.v39n6p729-736/2019

Ayu, W., Sinaga, L., Sumarno, S., & Sari, I. P. (2022). Penerapan Metode Regresi Linier

Berganda Untuk Estimasi Jumlah Penduduk Pada Kecamatan Gunung Malela. *Journal of Machine Learning and Artificial Intelligence*, *1*(1), 55–64. https://doi.org/10.55123/jomlai.v1i1.143

Ayuni, G. N., & Fitrianah, D. (2019). Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti Pada PT XYZ. *Jurnal Telematika*, *14*(2), 79–86. https://journal.ithb.ac.id/telematika/article/view/321

Azmi, U., Hadi, Z. N., & Soraya, S. (2020). ARDL METHOD : Forecasting Data Jumlah Hari Terjadinya Hujan Di NTB. *Jurnal Varian*, *3*(2), 73–82. https://doi.org/10.30812/varian.v3i2.627

Azwar, M., Hidayat, S., Yudha, F., Informatika, J., Magister, P., Industri, F. T., & Indonesia, U. I. (2021). Teknik Audio Forensik Dengan Metode Minkowski Untuk Pengenalan Rekaman Suara Pelaku Kejahatan. *CyberSecurity Dan Forensik Digital*, *4*(1), 1–12.

Begum, A., Kheya, N. J., & Rahman, Z. (2022). Housing Price Prediction with Machine Learning. *International Journal of Innovative Technology and Exploring Engineering*, *3075*(3), 42–46. https://doi.org/10.35940/ijitee.C9741.0111322

Budilaksana, D., Sukarsa, I. M., Agung, A., & Agung, K. (2021). Implementing k-Nearest Neighbor Methods to Predict Car Prices. *Jurnal Ilmiah Merpati*, *9*(1), 58–71.

Labib, M., Damayanti, S. A., Zaki, H. N., Muhayat, T., & Wirawan, R. (2021). Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression. *Jurnal Informatik*, *4221*(3), 238–245.

Lydia, E. L., Bindu, G. H., Sirisham, A., & Kiran, P. P. (2019). Electronic Governance of Housing Price Using Boston Dataset Implementing Through Deep Learning Mechanism. *International Journal of Recent Technology and Engineering*, *7*(6), 560–563.

Mauladi, K. F., Informatika, T., Teknik, F., Lamongan, U. I., Ujian, N., & Absolute, M. (2020). Perbandingan Metode Regresi Linear Dan Neural Network Backpropagation Dalam Prediksi Nilai Ujian Nasional Siswa SMP Menggunakan Software R. *Joutica*, *5*(1).

Pathak, S. M., & Chaudhari, P. A. K. (2021). Comparison of Machine Learning Algorithms for House Price Prediction Using Real Time Data. *International Journal of Engineering Research and Technology*, *10*(12), 300–305.

Permana, A. P., Ainiyah, K., & Holle, K. F. H. (2021). Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(3), 178–188. https://doi.org/10.14421/jiska.2021.6.3.178-188

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

Putro, B., Furqon, M. T., & Wijoyo, S. H. (2018). Prediksi Jumlah Kebutuhan Pemakaian

Air Menggunakan Metode Exponential Smoothing. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(11), 4679–4686.

Saiful, A., Andryana, S., & Gunaryati, A. (2021). Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning Dengan Algoritma Linear Regression. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *8*(1), 41–50. https://doi.org/10.35957/jatisi.v8i1.701

Santoso, P. Y., & Kusumaningsih, D. (2018). Algoritma K-nearest Neighbor Dengan Menggunakan Metode Euclidean Distance Untuk Memprediksi Kelulusan Ujian Nasional Berbasis Desktop SMA Negeri 12 Tangerang. *Skanika 2018*, *1*(1), 123–129.

Setiawan, A. (2022). Perbandingan Penggunaan Jarak Manhattan, Jarak Euclid, dan Jarak Minkowski dalam Klasifikasi Menggunakan Metode KNN Pada Data Iris. *Jurnal Sains Dan Edukasi Sains*, *5*(1), 28–37. https://doi.org/10.24246/juses.v5i1p28-37

Tatachar, A. V. (2021). Comparative Assessment of Regression Models Based On Model Evaluation Metrics. *International Journal of Innovative Technology and Exploring Engineering*, *8*(9), 853–860.

Utomo, P. B., Utami, E., & Raharjo, S. (2019). Implementasi Metode K-Nearest Neighbor Dan Regresi Linear Dalam Prediksi Harga Emas. *Jurnal Informasi Interaktif*, *4*(3), 155–159.