Jurnal Riset Matematika (JRM)
e-ISSN 2798-6306 | p-ISSN 2808-313X
https://journals.unisba.ac.id/index.php/JRM

Tersedia secara online di
Unisba Press
https://publikasi.unisba.ac.id/

Penerapan Metode *K-Nearest Neighbor* untuk Prediksi Harga Gas Alam Menggunakan *Python*

Selby Diva Qirani, Icih Sukarsih*

Prodi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

ARTICLE INFO

Article history :
Received : 28/03/2024
Revised : 28/05/2024
Published : 13/06/2024

Volume : 4
No. : 1
Halaman : 57-64
Terbitan : **Juli 2024**

ABSTRAK

Parameter yang dibutuhkan untuk melakukan prediksi menggunakan KNN adalah nilai *p* (banyak kolom input), *k* (banyak data terdekat), serta fungsi jarak yang digunakan. Tujuan penelitian ini adalah untuk mencari parameter terbaik dalam memprediksi harga gas alam berdasarkan nilai *Root Mean Squared Error* (RMSE) yang paling kecil dengan bantuan bahasa pemrograman *Python* dan juga penerapan metode *K-Fold Cross Validation.* Penelitian ini menggunakan data per bulan harga gas alam dunia tertinggi pada bulan tersebut dengan evaluasi tingkat kesalahan yang digunakan adalah *Mean Absolute Percentage Error* (MAPE). Tujuan dari penelitian ini adalah melakukan prediksi harga gas alam menggunakan metode *K-Nearest Neighbor* (KNN) serta bagaimana pengaruh penghapusan data pencilan terhadap nilai *Mean Absolute Percentage Error* (MAPE) dari prediksi harga gas alam. Ditemukan data pencilan pada harga gas alam selama periode 15 bulan terletak di antara selang waktu bulan September 2021 hingga bulan Desember 2022. Parameter terbaik dengan nilai MAPE paling kecil yaitu $p = 2$, $k = 2$, dan fungsi jarak Chebysev. Parameter ini memiliki nilai MAPE sebesar 14,5306% dengan tanpa pencilan serta tanpa menerapkan *K-Fold Cross Validation*. Pada bulan Juni, prediksi harga gas alam adalah $2,8215, sedangkan harga aktualnya yaitu $2,839. Kesalahan relatif yang diukur oleh Mean Absolute Percentage Error (MAPE) pada prediksi ini sebesar 0,6164%.

Kata Kunci : KNN; Prediksi; RMSE.

ABSTRACT

The required parameters for making predictions using KNN include the value of p (number of input columns), k (number of nearest data points), and the distance function used. The objective of this research is to identify the optimal parameters for predicting natural gas prices based on minimizing the Root Mean Squared Error (RMSE) value, utilizing the Python programming language. Additionally, the study involves the implementation of the K-Fold Cross Validation method for a comprehensive evaluation of predictive performance. The research utilizes monthly data on the world's highest natural gas prices, with the evaluation of error rates conducted through the Mean Absolute Percentage Error (MAPE). The objective of this study is to predict natural gas prices using the KNN method and to analyze the impact of outlier data removal on the MAPE of natural gas price predictions. The optimal parameters, yielding the smallest MAPE, are determined to be p = 2, k = 2, and the Chebyshev distance function. These parameters result in an MAPE value of 14,5306%, with no outliers and without applying K-Fold Cross Validation. In the month of June, the predicted natural gas price is $2,8215, while the actual price is $2,839. The relative error, measured by the MAPE for this prediction, is 0,6164%.

Keywords : KNN; Prediction; RMSE.

Corresponding Author : *sukarsih@unisba.ac.id
Indexed : Garuda, Crossref, Google Scholar
DOI : https://doi.org/10.29313/jrm.v4i1.3602

A. Pendahuluan

K-Nearest Neighbor (KNN) merupakan suatu metode yang awalnya dirancang untuk klasifikasi, tetapi kini diterapkan pada regresi [1]. Terdapat beberapa parameter yang dibutuhkan metode *K-Nearest Neighbor* (KNN) dalam melakukan regresi, yaitu banyaknya kolom input yang disimbolkan dengan *p* atau bisa dikatakan bahwa suatu data dipengaruhi oleh *p* data sebelumnya. Parameter selanjutnya yaitu banyaknya data terdekat yang akan digunakan sebagai acuan nilai prediksi yang kemudian disimbolkan dengan *k*. Parameter lainnya yang digunakan yaitu fungsi jarak untuk menghitung jarak antar setiap data. Terdapat beberapa penelitian, yang menggunakan metode KNN untuk regresi yaitu prediksi harga bahan pokok dan prediksi harga beras [2], [3] prediksi pertumbuhan jumlah penduduk [4], serta prediksi harga emas [5]. Selain itu, masalah lain yang penting untuk diprediksi adalah harga gas alam, mengingat saat kini gas alam kembali menjadi perhatian sebagai dampak invansi yang dilakukan Rusia terhadap Ukraina.

Gas alam adalah campuran gas dari parafin, karbon, hidrogen, dalam bentuk gas dan persentasenya bergantung pada sumber gas alam tersebut [6]. Selama tahun 2022, Indonesia menduduki posisi ke-15 negara teratas dalam produksi gas alam sebanyak 57,7 bcm [7]. Dengan berlimpahnya gas alam yang diproduksi Indonesia, maka gas alam Indonesia tidak hanya dimanfaatkan dalam negeri tetapi juga diekspor. Berdasarkan Peraturan Menteri Perdagangan Republik Indonesia Nomor 45 Tahun 2022 tentang Penetapan Harga Patokan Ekspor Atas Produk Pertambangan Yang Dikenakan Bea Keluar pada Pasal 1 menyebutkan bahwa penetapan Harga Patokan Ekspor (HPE) produk pertambangan salah satunya mengacu pada harga rata-rata tertinggi di bursa Internasional dalam satu bulan terakhir.

Dalam upaya mengevaluasi model secara menyeluruh, penelitian ini mengintegrasikan metode *K-Fold Cross Validation* untuk validasi model. Langkah ini memungkinkan pembagian dataset ke dalam k subset untuk evaluasi menyeluruh terhadap model yang dikembangkan. Prediksi harga gas alam menggunakan metode *K-Nearest Neighbor* membutuhkan banyak langkah yang harus dilakukan sehingga jika dilakukan secara manual akan membutuhkan waktu yang lama. Penggunaan *tools* dan bahasa pemrograman tertentu dapat mempercepat proses prediksi serta dapat mengurangi kemungkinan salah perhitungan. Salah satu bahasa pemrograman yang bisa digunakan dalam pengolahan data adalah *python.* Beberapa keunggulan yang dimiliki *python* di antaranya yaitu merupakan bahasa pemrograman tingkat tinggi sehingga mudah digunakan serta bersifat *open source* [8]. Tujuan dari penelitian ini adalah melakukan prediksi harga gas alam menggunakan metode *K-Nearest Neighbor* (KNN) serta mengetahui bagaimana pengaruh penghapusan data pencilan terhadap nilai *Mean Absolute Percentage Error* (MAPE) dari prediksi harga gas alam.

B. Metode Penelitian

Dalam penelitian ini, digunakan data harga gas alam dunia setiap bulan sejak Januari 2016 hingga Mei 2023 dengan 89 baris data. Harga gas alam yang digunakan merupakan harga gas alam tertinggi padas setiap bulannya.

K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (KNN) untuk melakukan prediksi, yaitu [4]: (1) Menentukan nilai *p* (jumlah kolom input), *k* (jumlah tetangga) dan fungsi jarak; (2) Representasikan ulang data ke dalam format berdasarkan nilai *p*. Jika $\boldsymbol{y_i}$ merupakan nilai setiap baris data *time series* dengan jumlah data *time series* sebanyak $\boldsymbol{n}$ data maka format data dengan *p* kolom input ditunjukkan pada Tabel 1.

Tabel 1. Representasi Dataset dengan *p* kolom input

x_p	x_{p-1}	x_{p-2}	…	x_1	y
y_{i-p}	$y_{i-(p-1)}$	$y_{i-(p-2)}$	…	y_{i-1}	y_i

Nilai y pada kolom x_j adalah $y_{(i-j)}$ dengan $i = p+1, \dots, n$ dan $j = 1, 2, \dots, p$

Menghitung jarak antara kolom input yang akan diprediksi dengan beberapa kolom input sebelumnya. Mengurutkan data berdasarkan jarak terkecil, kemudian menghitung rata-rata target dari *k* data teratas.

Pada penelitian ini, digunakan 3 fungsi jarak yaitu Euclid, Manhattan, dan Chebyshev.

Fungsi Jarak *Euclidean*

$$d_E(i,j) = \sqrt{\sum_{k=1}^{p}(x_{(i,k)} - x_{(j,k)})^2} \quad (1)$$

Fungsi Jarak *Manhattan*

$$d_M(i,j) = \sum_{k=1}^{p}\left|x_{(i,k)} - x_{(j,k)}\right| \quad (2)$$

Fungsi Jarak *Chebyshev*

$$d_C(i,j) = max_{k=i}^{p}\left|x_{(i,k)} - x_{(j,k)}\right| \quad (3)$$

Keterangan :

$d_E(i,j)$ = jarak *Euclidean* antara data baris ke-*i* dan data baris ke-*j*
$d_M(i,j)$ = jarak *Manhattan* antara data baris ke-*i* dan data baris ke-*j*
$d_c(i,j)$ = jarak *Chebyshev* antara data baris ke-*i* dan data baris ke-*j*
p = banyaknya kolom input
$x_{(i,k)}$ = data pada baris ke-*i* dan kolom x_k
$x_{(j,k)}$ = data pada baris ke-*j* dan kolom x_k

Dengan baris ke-*j* merupakan baris yang akan dicari nilai prediksinya serta *i* = 1,2,…, *j*-1.

Cross Validation

Cross Validation adalah sebuah teknik validasi model untuk menilai bagaimana hasil statistik analisis akan menggeneralisasi kumpulan data independen. Teknik ini utamanya digunakan untuk melakukan prediksi model dan memperkirakan seberapa akurat sebuah model prediktif ketika dijalankan dalam praktiknya. Salah satu teknik dari *cross validation* adalah *k-fold cross validation,* yaitu memecah data menjadi k bagian set data dengan ukuran yang sama [9]. Penerapan *k-fold cross validation* dalam penelitian ini bertujuan untuk meningkatkan generalisasi evaluasi data. Dalam *k-fold cross validation*, setiap bagian data mendapatkan posisi sebagai data uji, dan proses evaluasi dilakukan secara berulang dengan bagian set data yang digunakan secara bergantian. Pada penelitian ini, proses prediksi harga gas alam akan dilakukan dengan menerapkan *k-Fold cross validation* serta tanpa menerapkan *k-fold cross validation.*

Evaluasi Tingkat Kesalahan

Pada penelitian ini akan digunakan dua evaluasi tingkat kesalahan yaitu *Root Mean Squared Error* (RMSE) dan *Mean Absolute Percentage Error* (MAPE). Tujuan dari menghitung tingkat kesalahan adalah untuk mengevaluasi akurasi hasil prediksi. RMSE digunakan ketika mencari parameter *K-Nearest Neighbor* (KNN) terbaik dan MAPE digunakan untuk mengevaluasi parameter terbaik terhadap *data testing.* MAPE merupakan rata-rata dari keseluruhan persentase kesalahan (selisih) antara data aktual dengan data hasil prediksi [10]. Rumus untuk RMSE dan MAPE yaitu:

$$RMSE = \sqrt{\frac{\sum_1^n (y_i - \hat{y}_i)^2}{n}} \quad (4)$$

$$MAPE = \frac{1}{n}\sum_1^n \left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \quad (5)$$

Keterangan :
n = banyaknya data
$\hat{y}_i$ = nilai prediksi data ke-i
$\boldsymbol{y_i}$ = nilai sebenarnya data ke-i
Kriteria keakuratan MAPE ditunjukkan pada Tabel 2.

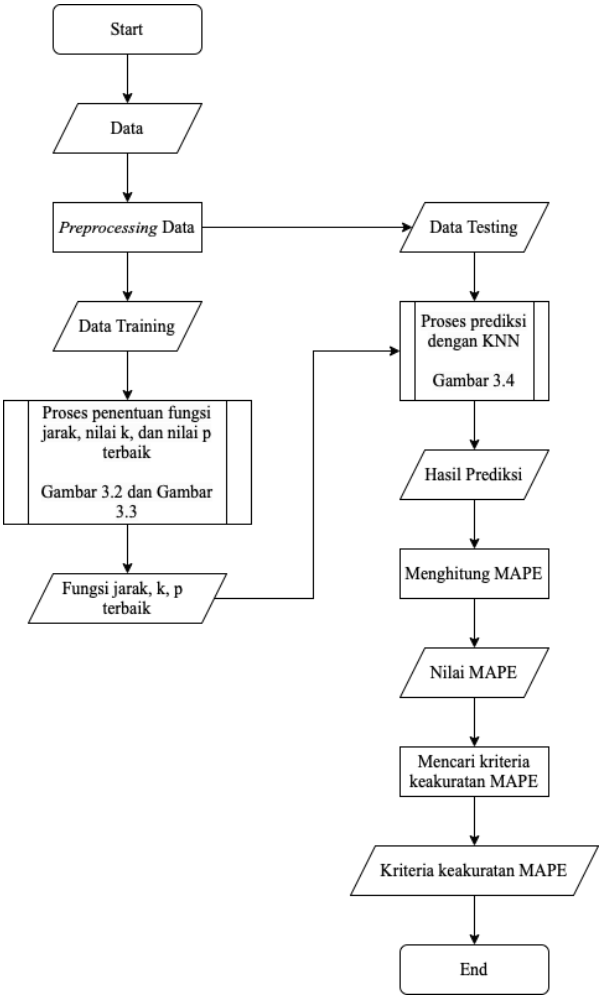
Tabel 2. Kriteria Keakuratan MAPE

Nilai MAPE	Akurasi Peramalan
MAPE ≤ 10%	Sangat Baik
10% < MAPE ≤ 20%	Baik
20% < MAPE ≤ 50%	Cukup
MAPE ≥ 50%	Buruk

Tahapan Penelitian

Pada penelitian ini, akan dilakukan 4 percobaan dengan masing-masing data sebagai berikut : (1) Data dengan pencilan dengan menerapkan *k-fold cross validation;* (2) Data dengan pencilan tanpa menerapkan *k-fold cross validation*; (3) Data tanpa pencilan dengan menerapkan *k-fold cross validation*; (4) Data tanpa pencilan tanpa menerapkan *k-fold cross validation*

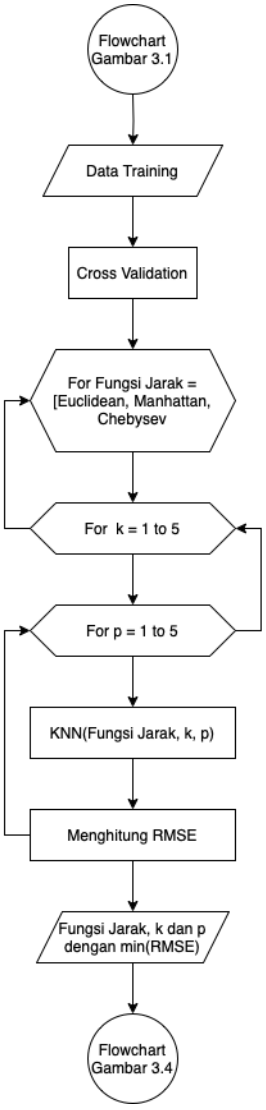
Dari 4 data yang akan digunakan pada penelitian ini, masing-masing data akan dilakukan proses seperti ditunjukkan *flowchart* pada Gambar 1.



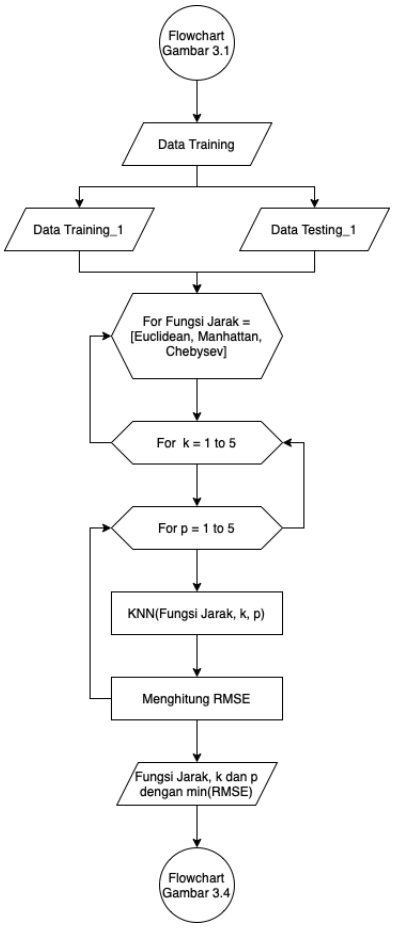
Gambar 1. *Flowchart* Langkah-Langkah Penelitian

Pada penelitian ini data akan dibagi menjadi dua bagian yaitu data *training* dan data *testing.* Data *training* digunakan untuk mencari nilai *k,* nilai *p,* dan fungsi jarak terbaik. Sedangkan, data *testing* digunakan untuk menguji keakuratan nilai *k,* nilai *p,* dan fungsi jarak yang diterapkan pada metode *K-Nearest Neighbor* terhadap dataset untuk memprediksi harga gas alam.

Proporsi pembagian data *training* dan data *testing* bersifat subjektif tergantung peneliti dengan data *testing* yang dapat digunakan yaitu 10% sampai 30% dari data terakhir [11]. Pada penelitian ini persentase yang akan digunakan untuk data *training* 80% serta 20% untuk data *testing.* Pengambilan persentase 80% dan 20% ini supaya data *testing* tidak terlalu sedikit maupun tidak terlalu banyak. proses pencarian nilai *k*, nilai *p,* dan fungsi jarak terbaik ditunjukkan pada flowchart Gambar 2 untuk proses dengan menerapkan *k-fold cross validation* sedangkan untuk proses tanpa menerapkan *k-fold cross validation* ditunjukkan pada flowchart Gambar 3.



Gambar 2. Langkah Mencari Parameter Terbaik Dengan *K-Fold Cross Validation*



Gambar 3. Langkah Mencari Parameter TerbaikTanpa *K-Fold Cross Validation*

C. Hasil dan Pembahasan

Hasil Parameter Terbaik Setiap Data

Parameter terbaik dari setiap data ditunjukkan pada Tabel 3.

Tabel 3. Parameter Terbaik dan Nilai RMSE

Data		Parameter Terbaik	RMSE
Dengan Pencilan	Menerapkan *K-Fold Cross Validation*	p = 1, k = 4, Fungsi Jarak Euclidean, Manhattan, Chebyshev	0,5167
	Tidak menerapkan *K-Fold Cross Validation*	p = 5, k = 1, Fungsi Jarak Euclidean	0,7764
Tanpa Pencilan	Menerapkan *K-Fold Cross Validation*	p = 1, k = 4, Fungsi Jarak Euclidean, Manhattan, Chebyshev	0,3659
	Tidak menerapkan *K-Fold Cross Validation*	p = 2, k = 2, Fungsi Jarak Chebyshev	0,3530

Selanjutnya, masing-masing parameter terbaik pada Tabel 3 akan diuji akurasinya dengan menerapkan parameter tersebut ke data testing. Hasil pengujian ditunjukkan pada Tabel 4.

Tabel 4. Nilai MAPE Pada Masing-Masing Parameter

Data		Parameter Terbaik	MAPE (%)
Dengan Pencilan	Menerapkan *K-Fold Cross Validation*	p = 1, k = 4, Fungsi Jarak Euclidean, Manhattan, Chebyshev	23,4937
	Tidak menerapkan *K-Fold Cross Validation*	p = 5, k = 1, Fungsi Jarak Euclidean	32,8770
Tanpa Pencilan	Menerapkan *K-Fold Cross Validation*	p = 1, k = 4, Fungsi Jarak Euclidean, Manhattan, Chebyshev	17,4092
	Tidak menerapkan *K-Fold Cross Validation*	p = 2, k = 2, Fungsi Jarak Chebyshev	14,5306

Parameter terbaik yang memiliki nilai MAPE yang paling kecil dibanding tiga parameter lainnya yaitu p = 2, k = 2, Fungsi Jarak Chebyshev yaitu data tanpa pencilan serta tanpa penerapan *K-Fold Cross Validation.* Selain itu parameter p = 5, k = 1, Fungsi Jarak Euclidean mempunyai nilai MAPE tertinggi dibandingkan tiga parameter lainnya.

Prediksi Harga Gas Alam Bulan Juni 2023

Kemudian, akan dilakukan prediksi harga gas alam pada bulan Juni 2023 menggunakan setiap parameter terbaik. Hasil prediksi harga gas alam pada bulan Juni 2023 ditunjukkan pada Tabel 5. Diketahui harga gas alam bulan Juni 2023 adalah $2,839.

Tabel 5. Prediksi Harga Gas Alam Bulan Juni 2023

Data *Training*		Parameter	Harga Prediksi ($)	Persentase Error (%)
Dengan Pencilan	Menerapkan *K-Fold Cross Validation*	p = 1, k = 4, Fungsi Jarak Euclidean, Manhattan, Chebyshev	2,7045	4,7376
	Tidak menerapkan *K-Fold Cross Validation*	p = 5, k = 1, Fungsi Jarak Euclidean	2,475	12,8214
Tanpa Pencilan	Menerapkan *K-Fold Cross Validation*	p = 1, k = 4, Fungsi Jarak Euclidean, Manhattan, Chebyshev	2,7045	4,7376
	Tidak menerapkan *K-Fold Cross Validation*	p = 2, k = 2, Fungsi Jarak Chebyshev	2,8215	0,6164

Parameter p = 2, k = 2, Fungsi Jarak Chebyshev menghasilkan persentase error yang paling kecil dibandingkan tiga parameter lainnya yaitu sebesar 0,6164% dengan prediksi harga gas alam sebesar $2,8215.

D. Kesimpulan

Prediksi harga gas alam menggunakan metode *K-Nearest Neighbor* (KNN) membutuhkan tiga parameter yaitu banyaknya kolom input (p) atau bisa disebut dengan p data yang mempengaruhi nilai prediksi, banyaknya data terdekat yang mempengaruhi nilai prediksi (k), serta fungsi jarak.

Berdasarkan nilai *Mean Absolute Percentage Error* (MAPE) prediksi, data harga gas alam tanpa pencilan memberikan hasil prediksi harga gas alam yang lebih baik dibanding data dengan menyertakan pencilan. Tanpa penerapan *K-Fold Cross Validation* menghasilkan hasil prediksi yang lebih baik dengan nilai MAPE 14,5306% dengan parameter p = 2, k = 2, fungsi jarak Chebyshev. Dengan menerapkan *K-Fold Cross Validation* parameter terbaik yang didapat adalah p = 1, k = 4, fungsi Jarak Euclidean, Manhattan, Chebyshev dengan nilai MAPE 17,4092%. Kategori akurasi dari kedua parameter tersebut tergolong baik.

Daftar Pustaka

[1] M. Nanja and Purwanto, "Metode K-Nearest Neighbor Berbasis Forward Selection Untuk Prediksi Harga Komoditi Lada," *Jurnal Pseudocode*, vol. 2, no. 1, 2015, [Online]. Available: www.ejournal.unib.ac.id53

[2] S. D. Purwanto, A. C. Fauzan, A. Wahyudi, and F. Y. Mardana, "Sistem Prediksi Harga Kebutuhan Bahan Pokok Nasional Menggunakan Metode K-Nearest Neighbour," *Seminar Nasional Ilmu Komputer*, pp. 481–488, 2016.

[3] Y. 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, 2020.

[4] D. Sekar Seruni, M. Tanzil Furqon, and R. Cahya Wihandika, "Sistem Prediksi Pertumbuhan Jumlah Penduduk Kota Malang Menggunakan Metode K-Nearest Neighbor Regression," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 4, pp. 1075–1082, Apr. 2020.

[5] P. B. Utomo, E. Utami, and S. Raharjo, "Implementasi Metode K-Nearest Neighbor dan Regresi Linear Dalam Prediksi Harga Emas," *Jurnal Informasi Interaktif*, vol. 4, no. 3, pp. 131–200, 2019, [Online]. Available: http://e-journal.janabadra.ac.id/

[6] C. Aktemur, "An overview of natural gas as an energy source for various purposes," *International Journal of Engineering Technologies IJET*, vol. 3, no. 3, pp. 91–104, Sep. 2017, doi: 10.19072/ijet.300750.

[7] "Statistical Review of World Energy 2023 | 72 nd edition," 2023.

[8] R. R. Saragih, "Pemrograman dan Bahasa Pemrograman."

[9] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 5, pp. 577–584, 2018, doi: 10.25126/jtiik20185983.

[10] F. Aditya, D. Devianto, and Maiyastri, "Peramalan Harga Emas Indonesia Menggunakan Metode Fuzzy Time Series Klasik," *Jurnal Matematika UNAND*, vol. 8, no. 2, pp. 45–52, 2019.

[11] S. AISYAH, S. WAHYUNINGSIH, and F. AMIJAYA, "Peramalan Jumlah Titik Panas Provinsi Kalimantan Timur Menggunakan Metode Radial Basis Function Neural Network," *Jambura Journal of Probability and Statistics*, vol. 2, no. 2, pp. 64–74, Nov. 2021, doi: 10.34312/jjps.v2i2.10292.