Journal of Mathematics, Computations, and Statistics (hal. 161 – 168)
Vol. 6. No. 2, Oktober 2023
http://www.ojs.unm.ac.id/jmathcos

Penerapan *K-Fold Cross Validation* untuk Menganalisis Kinerja Algoritma *K-Nearest Neighbor* pada Data Kasus Covid-19 di Indonesia

Hardianti Hafid[1, a)]

[1]*Program Studi Satatistika, Universitas Negeri Makassar*

[a)]hardiantihf@unm.ac.id

Abstrak. *Pandemi Covid-19 telah menjadi tantangan global dalam beberapa tahun terakhir. Virus ini telah mempengaruhi sebagian besar aspek kehidupan manusia, termasuk kesehatan, ekonomi, dan masyarak, Indonesia merupakan salah satu negara terdampak yang saat ini telah memasuki masa endemi. Penelitian ini bertujuan untuk menerapkan metode K-Fold Cross Validation untuk menganalisis kinerja algoritma K-Nearest Neighbor (K-NN) pada data kasus Covid-19 di Indonesia, sehingga mampu mengukur sejauh mana model K-NN dapat memprediksi kasus Covid-19 dengan akurat. Hasil yang diperoleh dengan menggunakan 30 Fold cross-validation dan nilai k=5 menunjukkan tingkat akurasi sebesar 68.65% dan nilai kappa sebesar 0.5123. Hasil ini menunjukkan bahwa model K-NN mampu memberikan prediksi yang memadai dan memiliki tingkat kesepakatan yang lebih tinggi. Penelitian ini memberikan pemahaman yang lebih mendalam tentang kinerja algoritma K-NN dalam konteks data kasus Covid-19 di Indonesia, yang dapat digunakan sebagai landasan untuk perbaikan lebih lanjut dalam pemodelan dan pemahaman pada data kasus Covid-19.*
Kata Kunci*: K-Fold Cross Validation, K-Nearest Neighbor (K-NN), Covid-19*

Abstract. *The Covid-19 pandemic has been a global challenge in recent years. This virus has impacted most aspects of human life, including health, the economy, and society. Indonesia is one of the affected countries that has now entered an endemic phase. This research aims to apply the K-Fold Cross Validation method to analyze the performance of the K-Nearest Neighbor (K-NN) algorithm on Covid-19 cases data in Indonesia, in order to measure how accurately the K-NN model can predict Covid-19 cases. The results obtained using 30-Fold cross-validation with a value of k=5 show an accuracy rate of 68.65% and a kappa value of 0.5123. These results indicate that the K-NN model is capable of providing adequate predictions with a higher level of agreement. This research provides a deeper understanding of the performance of the K-NN algorithm in the context of Covid-19 cases data in Indonesia, which can be used as a foundation for further improvements in modeling and understanding Covid-19 case data.*
Keywords: *K-Fold Cross Validation, K-NN, Covid-19*

PENDAHULUAN

Pandemi Covid-19 telah menjadi tantangan global dalam beberapa tahun terakhir. Virus ini telah mempengaruhi sebagian besar aspek kehidupan manusia, termasuk kesehatan, ekonomi, dan masyarakat, Indonesia merupakan salah satu negara terdampak. Untuk mengatasi pandemi ini dengan efektif, analisis data yang akurat dan efisien sangat penting. Salah satu alat yang digunakan untuk menganalisis data kasus Covid-19 adalah algoritma *machine learning*, algoritma *K-Nearest Neighbor (K-NN).* Algoritma K-NN telah digunakan secara luas dalam pemrosesan data dan pengambilan keputusan.

K-NN memiliki kesamaan dengan *Random Decision Forest (RDF)* dalam hal prediksi dan merupakan algoritma klasifikasi dengan tingkat kompleksitas yang rendah serta mampu melakukan perhitungan dengan cepat (Bird, dkk, 2020). Penerapan algoritma K-NN, telah banyak dilakukan peneliti-peneliti sebelumnya diantaranya yaitu dalam penelitian K-NN digunakan untuk mendeteksi penyakit kanker payudara (Binabar & Ivandari, 2018). Dalam penelitian lain, algoritma K-NN berhasil mendeteksi dini penyakit hepatitis C, dengan akurasi model 92% (Kusuma dan Astuti, 2022), algoritma (K-NN) berhasil dalam melakukan klasifikasi terhadap penyakit diabetes melitus, studi kasus warga Desa Jatitengah (Fasnuari, dkk., 2022) dan algoritma K-NN juga berhasil melakukan prediksi penyakit stroke dengan akurasi yang didapatkan sebesar 95% (Maskuri, 2022).

Pada data kasus Covid-19, K-NN dapat digunakan untuk mengklasifikasikan dan memprediksi tren kasus, pergerakan virus, dan dampak pandemi pada berbagai wilayah di Indonesia, namun untuk mengukur kinerja algoritma K-NN dalam menganalisis data Covid-19, diperlukan pendekatan evaluasi yang tepat. Sehingga dalam penelitiam ini digunakan penerapan metode evaluasi kinerja yaitu *K-Fold Cross Validation*, untuk menganalisis dan memahami sejauh mana algoritma K-NN efektif dalam klasifikasi data kasus Covid-19 di Indonesia.

K-Fold Cross Validation digunakan untuk mengukur kinerja algoritma K-NN secara lebih umum dengan membagi dataset menjadi data *training* dan data *testing*. Hal ini dapat membantu mengidentifikasi apakah model memiliki tingkat akurasi yang baik, serta membantu menghindari *overfitting* atau *underfitting*. Berdasarkan hal tersebut, penelitian ini bertujuan untuk memberikan kontribusi signifikan dalam penggunaan *machine learning*, khususnya algoritma *K-Nearest Neighbor* dalam analisis dan pemahaman data Covid-19, yang dapat mendukung masa endemi di Indonesia dan negara-negara lainnya.

METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder yang berasal dari publikasi satuan tugas penanganan covid-19 yang dapat diakses pada website https://covid19.go.id/ yaitu data pada 7 Agustus 2022 - 18 Juni 2023 di 34 provinsi Indonesia, dengan zona resiko rendah, sedang dan tinggi.

TABEL 1. Data Penelitian

No	Tanggal	Provinsi	Jumlah Positif (orang)	Zona Resiko
1	07-08-2022	Aceh	22	rendah
2	14-08-2022	Aceh	35	rendah
⋮	⋮	Aceh	⋮	⋮
46	18-06-2023	Aceh	3	rendah
47	07-08-2022	Bali	1183	tinggi
48	14-08-2022	Bali	983	tinggi
⋮	⋮	⋮	⋮	tinggi
92	18-06-2023	Bali	22	tinggi
93	07-08-2022	Bangka Belitung	36	sedang
94	14-08-2022	Bangka Belitung	66	sedang
⋮	⋮	⋮	⋮	⋮
138	18-06-2023	Bangka Belitung	5	sedang
⋮	⋮	⋮	⋮	⋮
1564	18-06-2023	Yogyakarta	39	tinggi

Adapun langkah-langkah yang dilakukan dalam penelitian ini adalah:

1. Pengumpulan data
   Mengumpulkan data terkait kasus Covid-19 di Indonesia selama tahun 2022-2023. Data ini dapat diperoleh melalui publikasi di website satuan tugas penanganan covid-19
2. *Preprocessing data*
   Melakukan *preprocessing data* seperti membersihkan data yang tidak valid dan mengisi nilai yang hilang *(missing value)*
3. Penerapan Algoritma K-NN
   Menerapkan algoritma K-Nearest Neighbor pada data Covid-19, selanjutnya menetukan parameter k yang optimal untuk kemudian melakukan klasifikasi terkait kasus Covid-19 pada tiap provinsi di Inonesia
4. Penerapan K-*Fold* Cross Validation
   Menerapkan metode *K-Fold Cross Validation* dengan bantuan software R studio untuk mengukur kinerja algoritma K-NN, dalam penelitian ini digunakan perbandingan *fold* = 10, 15, 20, 25 dan 30 yang akan dibagi menjadi *data training* dan *data testing.*
5. Evaluasi Kinerja
   Menghitung evaluasi kinerja perbandingan *Fold* = 10, 15, 20, 25 dan 30 seperti nilai akurasi dan kappa untuk setiap *iterasi K-Fold Cross Validation*, hal ini akan memberikan gambaran tentang seberapa baik algoritma K-NN berkinerja pada dataset Covid-19.

HASIL DAN PEMBAHASAN

Penerapan Algoritma *K-Nearest Neighbor (K-NN).*

Akurasi perhitungan algoritma K-NN juga meningkat ketika digunakan lebih banyak sampel dan data *training* (Muliono, 2020). Tahapan penerapan algoritma K-NN:

1. Menentukan nilai parameter k sebagai *neighborhood* yang paling dekat dari data latih terhadap data uji,
2. Menghitung jarak antara nilai vektor data uji dengan semua vektor data latih dengan teorema *euclidean distance* (Wu, X. dan Kumar, 2009) pada persamaan (1)

$$d(x,y) = \sum_{j=1}^{n}(x_j - y_j)^2 \quad (1)$$

keterangan:

d = jarak kedekatan

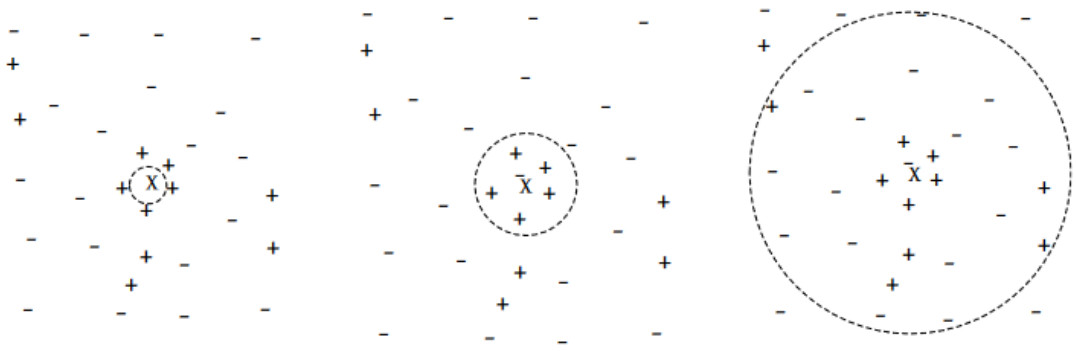
x = data *training*

y = data *testing*

n = jumlah atribut individu antara 1 sampai dengan n

j = atribut individu antara 1 sampai dengan n

3. Mengambil sejumlah nilai parameter *k* data latih terdekat



(a) (b) (c)

GAMBAR 1. Klasifikasi *K-Nearest Neighboors* dengan nilai *k* (tetangga) (a) kecil, (b) medium dan (c) besar. (Sumber: Wu, X. dan Kumar, 2009)

Ada beberapa faktor penting yang memengaruhi kinerja algoritma *K-Nearest Neighbors* (k-NN). Salah satunya adalah pilihan nilai *k*. Hal ini dapat dilihat pada Gambar 1, yang menunjukkan objek uji tanpa label "X" dan objek pelatihan yang termasuk ke dalam kelas " + " atau "−". Jika nilai *k* terlalu kecil, hasilnya dapat sangat sensitif terhadap titik-titik noise, sedangkan jika nilai *k* terlalu besar, maka lingkungan sekitarnya bisa mencakup terlalu banyak titik dari kelas lain. Kita dapat perkiraan nilai terbaik untuk k melalui metode validasi silang, *k* = 1 mampu bersaing dengan nilai k lainnya, terutama untuk data set kecil yang sering digunakan dalam penelitian atau *training* kelas. Tetapi jika jumlah sampel yang cukup besar, nilai k yang lebih besar cenderung lebih tahan terhadap gangguan dari titik-titik noise, adapun dalam penelitian ini dipilih nilai *k* = 5, 7, dan 9.

Evaluasi Kinerja

Pada penelitian ini akan dilakukan evaluasi kinerja pada klasifikasi K-NN menggunakan nilai akurasi dan kappa. Proses perhitungan nilai akurasi dapat dilakukan dengan menggunakan persamaan (2), sebagai berikut

$$akurasi = \frac{Jumlah\ klasifikasi\ benar}{jumlah\ data\ uji}\ x\ 100\% \qquad (2)$$

Koefisien Kappa sering digunakan untuk mengukur sejauh mana tingkat efektivitas metode yang digunakan. Kappa digunakan untuk mengukur tingkat kesesuaian antara sepasang variabel, yang sering digunakan sebagai metrik kesepakatan antar penilai, nilai ini berhubungan dengan data yang merupakan hasil penilaian, bukan pengukuran. Kappa membandingkan kemungkinan persetujuan dengan yang diharapkan jika pemeringkatannya independen. Nilai rentang terletak pada [− 1, 1] dengan 1 menunjukkan persetujuan penuh dan 0 berarti tidak ada kesepakatan atau independensi (Xia, 2020). Adapun pedoman nilai kappa yang dihasilkan sebagai berikut:

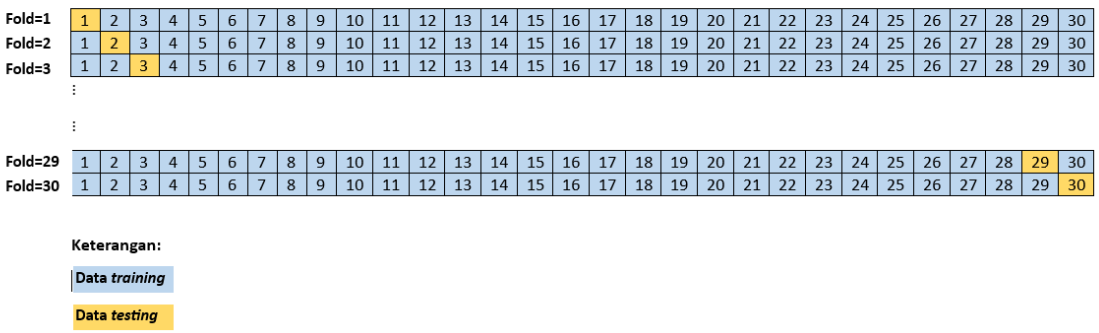
a. Nilai Kappa > 0,75 berarti ada kesesuaian yang baik *(excellent)* antara baris dan kolom.
b. Nilai Kappa antara 0,4 sampai 0,75 berarti ada kesesuaian yang cukup *(fair to good)* antara baris dengan kolom
c. Nilai Kappa <0,4 berarti ada keseuaian yang buruk *(poor)* antara baris dengan kolom

TABEL 2. Perbandingan Akurasi dengan Penerapan *K-Fold Cross Validation* pada metode K-NN

Fold	**k**	**Akurasi**	**Kappa**
10	5	0.6596	0.4694
	7	0.6404	0.4411
	9	0.6377	0.4321
15	5	0.6688	0.4829
	7	0.6616	0.4707
	9	0.6360	0.4286
20	5	0.6754	0.4944
	7	0.6517	0.4549
	9	0.6389	0.4319
25	5	0.6762	0.4967
	7	0.6531	0.4609
	9	0.6578	0.4656
30	5	0.6865	0.5123
	7	0.6601	0.4706
	9	0.6483	0.4518

Berdasarkan Tabel 2**,** dapat diketahui hasil klasifikasi yang dilakukan menggunakan K-NN, terlihat bawa *Fold* =30 dengan tetangga terdekat (k)=5 memiliki nilai akurasi terbesar dibandingkan nilai *Fold* lainnya yaitu 0,6865 dan Kappa= 0,5123. Berikut uraiannya:

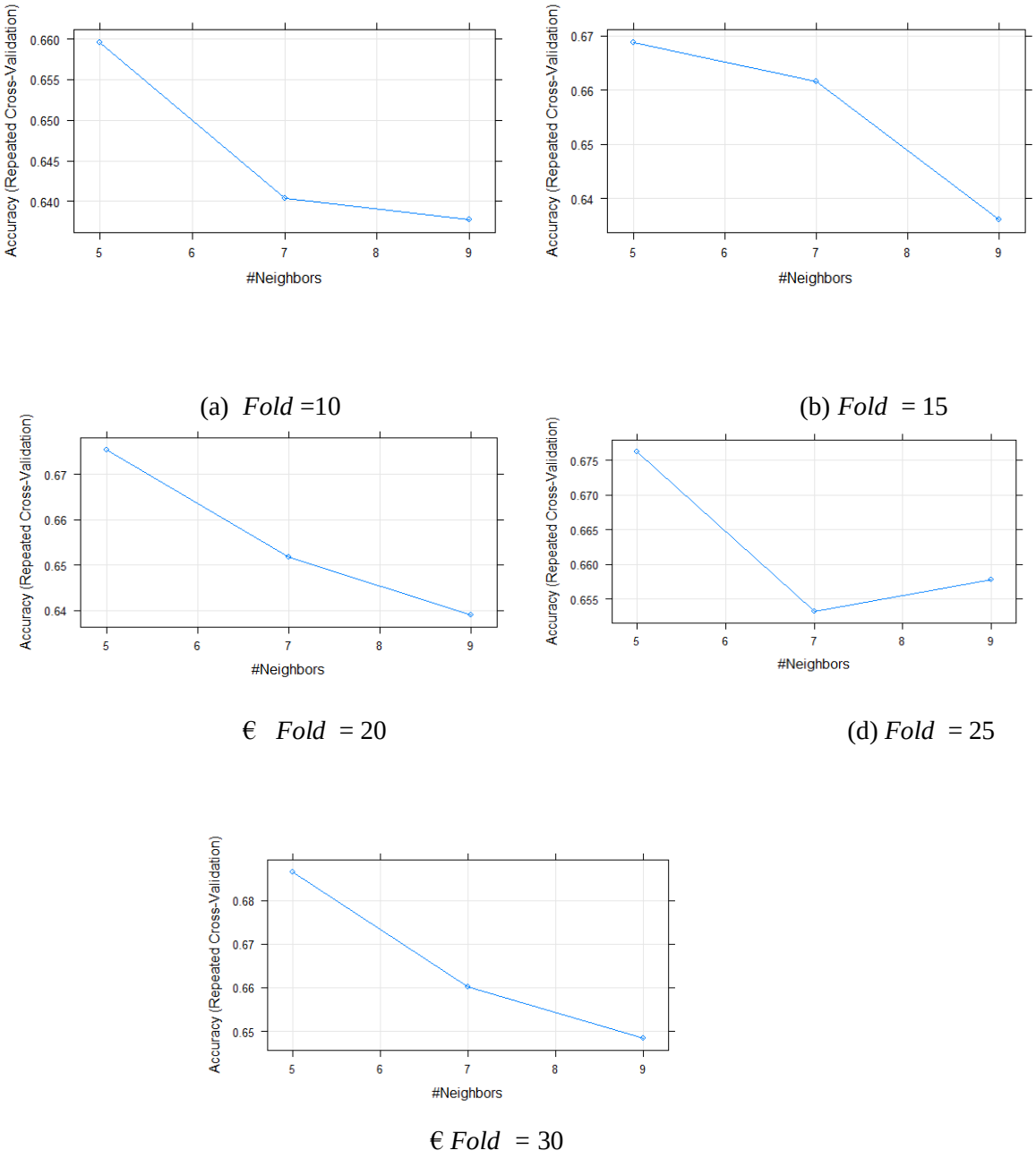
1. *Fold* = 30.
   Ini mengindikasikan bahwa pengujian yang dilakukan dengan menggunakan validasi silang *(cross-validation)* dengan 30 *Fold* yang dapat dilihat pada **GAMBAR 2**. Validasi silang adalah teknik yang digunakan untuk menguji kinerja model pada berbagai subset data yang berbeda. Dalam kasus ini, data dibagi menjadi 30 bagian yang berbeda, dan model diuji sebanyak 30 kali dengan menggunakan setiap bagian sebagai data *testing* satu per satu, sementara bagian lainnya digunakan sebagai data *training*. Hal ini membantu memastikan bahwa model tidak hanya bekerja dengan baik pada satu subset data tertentu dan memungkinkan pengukuran yang lebih stabil terhadap kinerja model.



GAMBAR 2. Simulasi data *training* dan *testing* pada *Fold*=30

2. $k = 5$

   Ini adalah parameter *k* dalam algoritma K-NN. Nilai *k* ini menentukan berapa banyak tetangga terdekat yang akan digunakan oleh model saat membuat prediksi, dalam penelitian ini model menggunakan 5 tetangga terdekat untuk memutuskan kelas prediksi untuk setiap data uji, hal ini juga terlihat pada Gambar 3 bahwa *k*=5 memiliki akurasi yang baik dibandingkan dengan nilai *k* lainnya dalam penelitian ini.



GAMBAR 3. Grafik Perbandingan Akurasi dengan Penerapan *K-Fold Cross Validation* pada metode K-NN

3. Akurasi = 0.6865:
   Ini adalah nilai akurasi model pada pengujian. Akurasi mengukur sejauh mana model berhasil memprediksi kelas dengan benar pada data pengujian, dalam penelitian ini, akurasi sebesar 0.6865 menunjukkan bahwa model K-NN mampu memprediksi dengan benar sekitar 68.65% dari data pengujian. Ini adalah metrik penting untuk mengevaluasi kinerja model.

4. Nilai kappa = 0.5123
   Kappa merupakan metrik statistik yang mengukur tingkat kesepakatan antara prediksi model dan hasil sebenarnya. Nilai kappa sebesar 0.5123 menunjukkan bahwa model K-NN dalam penelitian ini memiliki kesesuaian yang cukup *(fair to good)*. Nilai kappa yang positif menunjukkan bahwa model tidak hanya melakukan prediksi secara acak, tetapi memiliki tingkat akurasi yang lebih tinggi daripada yang diharapkan.

KESIMPULAN

Berdasarkan hasil yang diperoleh metode *K-Fold Cross Validation* untuk menganalisis kinerja algoritma *K-Nearest Neighbors* (K-NN) pada data kasus Covid-19 di Indonesia tahun 2022-2023. Hasil penelitian menunjukkan bahwa penggunaan *K-Fold Cross Validation* dengan 30 *Fold* memberikan akurasi yang lebih komprehensif tentang kemampuan model K-NN dalam mengklasifikasikan data kasus Covid-19. Hasil pengujian menunjukkan bahwa model K-NN dengan parameter k=5 dan fol=30 menghasilkan tingkat akurasi sebesar 68.65% dan tingkat kesepakatan (kappa) sebesar 0.5123. Ini menunjukkan bahwa model K-NN secara umum mampu memprediksi kasus Covid-19 dengan tingkat cukup baik. Penelitian ini memberikan kontribusi penting dalam konteks pemahaman dan analisis data kasus Covid-19 di Indonesia, yang dapat digunakan sebagai dasar untuk pengembangan lebih lanjut dalam meningkatkan akurasi prediksi dan pemahaman terhadap dinamika penyebaran wabah ini.

DAFTAR PUSTAKA

Binabar, S. W., & Ivandari, I. (2018). Optimasi Parameter K pada Algoritma *K-NN* untuk Deteksi Penyakit Kanker Payudara. *IC-Tech*, 13(1).

Bird, J.J., Barnes, C.M., Premebida, C., Ekárt, A. & Faria, D.R., (2020). Country-Level Pandemic Risk and Preparedness Classification Based on COVID-19 Data: A Machine Learning Approach. *Plos One*, 15(10). 1-20.

Fasnuari, H.A.D., Yuana, H., & Chulkamdi, M.T. (2022). Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Klasifikasi Penyakit Diabetes Melitus Studi Kasus : Warga Desa Jatitengah. *ANTIVIRUS: Jurnal Ilmiah Teknik Informatika*, 16(2). 133 – 142.

Kusuma, N.M.R.P., & Astuti, L.G. 2022.Implementasi Algoritma K-Nearest Neighbor (K-NN) dalam Deteksi Dini Penyakit Hepatitis C. *Jurnal Nasional Teknologi Informasi dan Aplikasinya*. 1(1).197-204

Maskuri, N.M., Harliana, H., Sukerti, K., & Bhakti, H. R. M. (2022). Penerapan Algoritma K-Nearest Neighbor (*K-NN*) untuk Prediksi Penyakit Stroke. *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, 4(1). 130–140.

R. Muliono, J. H. Lubis, & N. Khairina. 2020. Analysis K-Nearest Neighbor Algorithm for improving prediction student graduation time. *SinkrOn*, 4(2).

Santoso, S. 2005. *Mengatasi Berbagai Masalah Statistik*. Jakarta : PT Elex Media Komputindo.

Satuan Tugas Penanganan COVID-19. 2023. Analisis Data COVID-19 di Indonesia Updte 18 Juni 2023. [*cited* 2023 Juni-September]. Available from: https://covid19.go.id/

Xia, Y. L. 2020. Progress in Molecular Biology and Translational Science. *Science Diriect* 171: 309-491.

Wu, X. & Kumar, V. 2009. *The Top Ten Algorithms in Data Mining*. Minnesota: CRC Press.