ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Penerapan Data Mining dalam Mengklasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes

Diterima:
10 Juni 2024
Revisi:
10 Juli 2024
Terbit:
1 Agustus 2024

[1] Ahmad Fajar Abadi, [2] Nur Alamsyah, [3] Farhan Gagat Retnanto, [4] Erna Daniati, [5] Aidina Ristyawan
[1-5] Universitas Nusantara PGRI Kediri *[1]ahmadfajarabadi444@gmail.com, [2] nuralamsyah691@gmail.com, [3] farhangagat@gmail.com, [4] ernadaniati@unpkediri.ac.id, [5] aidinaristi@unpkediri.ac.id*

Abstrak— Penyakit stroke adalah kondisi medis mendadak yang disebabkan oleh gangguan aliran darah di otak, baik karena penyumbatan atau pecahnya pembuluh darah. Kondisi ini dapat menyebabkan sel-sel otak kekurangan oksigen, darah, dan nutrisi yang berakibat fatal. Di Indonesia, stroke menempati urutan pertama sebagai penyebab kematian, diikuti oleh diabetes dan hipertensi. Penelitian ini bertujuan untuk mengklasifikasi penyakit stroke menggunakan algoritma Naive Bayes dengan dataset dari *Kaggle.* Penelitian ini menggunakan metode kualitatif untuk analisis literatur dan metode kuantitatif untuk pengujian teori dengan data numerik. Data sekunder dari *Kaggle* dipreproses dengan menghilangkan data duplikat dan *missing value,* menghasilkan 4908 rekaman dari 5110 awal. Implementasi klasifikasi dilakukan menggunakan *RapidMiner Studio.* Hasil penelitian menunjukkan tingkat *acuraccy* 87.22%, dengan *precision* sebesar 14.93% dan *recall* 42.58%. Penelitian ini menunjukkan bahwa algoritma Naive Bayes dapat digunakan secara efektif untuk klasifikasi stroke dengan hasil yang cukup akurat.

Kata Kunci—stroke; *naïve bayes;* klasifikasi

Abstract— *Stroke is a sudden medical condition caused by a disruption of blood flow in the brain, either due to a blockage or rupture of blood vessels. This condition can lead to brain cells being deprived of oxygen, blood, and nutrients, which can be fatal. In Indonesia, stroke ranks as the leading cause of death, followed by diabetes and hypertension. This study aims to classify stroke using the Naive Bayes algorithm with a dataset from Kaggle. The research employs qualitative methods for literature analysis and quantitative methods for testing theories with numerical data. Secondary data from Kaggle was preprocessed by removing duplicates and missing values, resulting in 4908 records from an initial 5110. The classification implementation was carried out using RapidMiner Studio. The study results show an accuracy rate of 87.22%, with a precision of 14.93% and recall of 42.58%. This research demonstrates that the Naive Bayes algorithm can be effectively used for stroke classification with reasonably accurate results.*

Keywords—*stroke; naive bayes; classification*

Penulis Korespondensi:

Erna Daniati,
Sistem Informasi,
Universitas Nusantara PGRI Kediri,
Email: ernadaniati@unpkediri.ac.id
ID Orcid: [https://orcid.org/0009-0008-9471-4421]
Handphone: 081335242202

I. PENDAHULUAN

Stroke merupakan penyakit yang menyerang secara mendadak pada otak yang dapat mengganggu aliran darah yang diakibatkan oleh penyumbatan atau pecahnya pembuluh darah pada otak. Akibatnya sel – sel otak akan mengalami kekurangan oksigen, darah, dan nutrisi yang membuat penderita dapat meninggal dengan cepat[1]. Stroke menjadi salah satu penyakit yang banyak diderita oleh masyarakat Indonesia serta menjadi urutan pertama penyebab kematian paling tinggi dan disusul oleh diabetes dan hipertensi. Stroke termasuk jenis penyakit mematikan yang termasuk ke dalam 10 jenis penyakit mematikan di Indonesia. Berdasarkan data yang dikumpulkan dari sampel yang mewakili Indonesia, terdapat 41.590 kematian sepanjang tahun 2014. Kematian tersebut dilakukan dengan autopsi verbal, di mana sesuai dengan pedoman Badan Kesehatan Dunia (WHO) oleh dokter dan petugas terlatih[2].

Saat ini banyak orang yang belum mengenal tentang bagaimana penyakit stroke dan banyak yang tidak menyadari ketika gejala stroke sudah muncul dari awal. Banyak orang yang umumnya ragu untuk berkunjung ke rumah sakit untuk konsultasi tentang gejala yang dialami. Hal ini membuat perhatian yang

ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

membuat angka penderita penyakit stroke semakin tinggi dan terus menghantui kehidupan banyak orang. Ada beberapa faktor yang membuat angka penderita stroke bertambah yaitu gaya hidup masyarakat yang tidak terkontrol terhadap makanan cepat saji, stres, merokok, kurang olahraga, kerja berlebihan dan beberapa faktor lainnya[3].

Hal inilah yang mendorong banyak penelitian terhadap penyakit stroke, dengan menggunakan metode berbasis komputer. Dengan menggunakan algoritma tertentu, dapat dilakukan prediksi dengan mengelola dataset yang besar dengan hasil yang cepat dan akurat. Prediksi merupakan suatu proses untuk memperkirakan sesuatu secara sistematis berdasarkan informasi terdahulu dan yang tersedia sekarang[4]. Banyak penelitian yang dilakukan sebelumnya, salah satunya yang berjudul Klasifikasi Penyakit Stroke Menggunakan Metode *Naïve Bayes.* Penelitian tersebut menggunakan *brain stroke prediction dataset* dari *website kaggle*. Tujuan dari penelitian tersebut adalah mengklasifikasi penyakit stroke dengan metode *naive bayes* dari data *training* dan data *testing* yang berbeda. Data tersebut menggunakan kelas terkena stroke dan tidak terkena stroke dengan menggunakan 10 variabel bebas. Hasil akurasinya yaitu sebesar 80% yang diperoleh saat proporsi data *training* dan data *testing* 80:20[5].

ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

II. METODE

A. Jenis Penelitian

Penelitian ini menggunakan 2 jenis penelitian, yaitu penelitian kualitatif dan kuantitatif. Penelitian kualitatif digunakan untuk menganalisa dan memahami kajian literatur yang berkaitan dengan variabel atau objek yang digunakan dalam pengumpulan data[6]. Sedangkan penelitian kuantitatif digunakan untuk menyelidiki masalah berdasarkan pengujian sebuah teori yang terdiri dari beberapa variabel dan data – datanya berupa sesuatu yang dapat dihitung[7]

B. Metode Pengumpulan Data

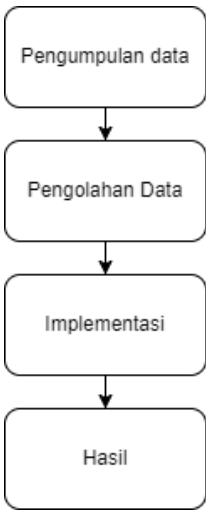
Pada penelitian ini, penulis menggunakan metode pengumpulan sekunder, di mana data yang digunakan diambil dari data peneliti terdahulu. Jadi penulis tidak perlu terjun langsung ke lapangan untuk mengambil datanya. Data yang dipakai merupakan data *Stroke Prediction Dataset* dari situs *website kaggle.com*, untuk data pendukung lainnya diambil dari buku atau jurnal publikasi dari penelitian sebelumnya.

C. *Preprocessing* Data

Preprocessing data merupakan tahapan untuk melakukan pembersihan data, dimana dilakukan pengecekan data apakah ada data duplikat atau data yang hilang. Jika ditemukan data duplikat dan data yang hilang bisa dilakukan input data dengan *median, mean,* atau pun bisa menghapus data tersebut[5]. Data dari penelitian ini diambil dari *website kaggle.com* dengan 11 atribut fitur dan 1 atribut label. Pada dataset yang digunakan, dilakukan *preprocessing* dengan cara menghilangkan data *missing value*. Data yang di dapat sebanyak 5110 *record.* Dari data tersebut ada sebanyak 202 yang dinyatakan sebagai *missing value*, dimana data tersebut terletak pada atribut BMI. Data yang dinyatakan *missing value* tersebut akhirnya dihilangkan agar valid, sehingga data yang diolah menjadi 4908 *record.*

D. Tahapan Penelitian

Penelitian ini terdiri dari beberapa tahapan. Tahapan pertama yaitu pengumpulan data, dimana data yang diambil dari *website kaggle.com.* Selanjutnya, tahap pengolahan data merupakan tahapan menganalisis data untuk memperbaiki atau menghapus data yang tidak lengkap, data duplikat dan data yang tidak akurat. Tahapan selanjutnya yaitu implementasi, implementasi dilakukan menggunakan *tools rapidminer studio.* Dari implementasi tersebut akan menghasilkan klasifikasi dari data yang digunakan. Berikut *flowchart* tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahapan penelitian

E. Teknik Klasifikasi

Teknik klasifikasi adalah suatu proses untuk memasukkan sautu objek ke dalam kelas tertentu dari jumlah kelas yang ada. Klasifikasi melakukan pembangunan model berdasarkan data latih yang ada, kemudian menggunakan model tersebut untuk mengklasifikasikan pada data yang baru. Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target yang memetakan setiap

ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

set atribut (fitur) ke satu jumlah label kelas yang tersedia [8]. Untuk menentukan kelas, klasifikasi memerlukan petunjuk menemukan sample yang cocok untuk dianalisis[9].

F. Data Mining

Data mining merupakan suatu proses mencari informasi dalam data yang berukuran besar dengan dipilih menggunakan teknik, metode, atau algoritma yang bervariasi[10]. Data mining digunakan untuk ekstrasi data yang berukuran besar menggunakan ilmu gabungan dari beberapa bidang ilmu seperti matematika, statistik, dan kecerdasan buatan. Salah satu cara untuk ekstrasi data yaitu dengan menggunakan metode klasifikasi[11].

G. Algoritma *Naive Bayes*

Naive bayes adalah sebuah metode yang digunakan untuk mengklasifikasi data statistik dengan cara memprediksi probabilitas keanggotaan suatu *class*. *Naïve Bayes* merupakan suatu kelas keputusan, dengan menggunakan probabilitas matematika dengan syarat bahwa nilai keputusan adalah benar [12]. Untuk memprediksi probabilitas dengan mudah, *naive bayes* memiliki rumus sebagai berikut[1].

$$P(X \vee Y) = \frac{P(X|Y).P(Y)}{P(X)}$$

P(Y|X) : Peluang terjadinya Y berdasarkan kondisi X

P(Y) : Peluang terjadinya Y (prior prob)

P(X|Y) : Peluang terjadinya X berdasarkan kondisi pada hipotesis Y

P(X) : Peluang terjadinya X

H. *Confusion Matrix*

Tabel 1. *Confusion matrix*

		Observed	
		True	*False*
Precdit Class	*True*	*True Positive* (TP)	*False Positive* (FP)
	False	*True Negative* (FN)	*True Negative* (TN)

Confusion matrix merupakan table klasifikasi yang menampilkan jumlah data *training* yang benar dan jumlah data training yang salah[13] *Confusion matrix* digunakan untuk menghitung jumlah nilai *accuracy, precission, dan recall*[14]. *Confusion matrix* terdiri dari *true positive, false positive, true negative* dan *false negative* untuk menghitung *precision*, *recall* dan *accracy Precision* merupakan tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. *Recall* merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Sedangkan *accuracy* merupakan tingkat kedekatan antara nilai prediksi dengan nilai aktual. Untuk menghitung *precision, recall* dan *accuracy* dapat menggunakan persamaan berikut[15]:

ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

A. Accuracy

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} X\ 100\,\% \quad (1)$$

B. Precision

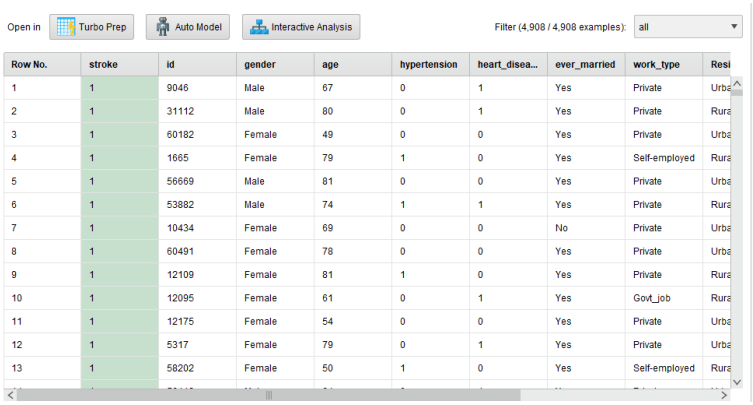
$$Precission = \frac{TP}{TP + FP} X\ 100\,\% \quad (2)$$

C. Recall

$$Recall = \frac{TP}{TP + FN} X\ 100\,\% \quad (3)$$

III. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data *training*, dataset yang digunakan diperoleh dari *website kaggle.com.* Analisis *Stroke Prediction Dataset* ini menggunakan algoritma naive bayes dan menggunakan *software rapidminer* sebagai alat hitungnya.


Open in Turbo Prep Auto Model Interactive Analysis Filter (4,908 / 4,908 examples): all

Row No.	stroke	id	gender	age	hypertension	heart_disea...	ever_married	work_type	Resi
1	1	9046	Male	67	0	1	Yes	Private	Urba
2	1	31112	Male	80	0	1	Yes	Private	Rura
3	1	60182	Female	49	0	0	Yes	Private	Urba
4	1	1665	Female	79	1	0	Yes	Self-employed	Rura
5	1	56669	Male	81	0	0	Yes	Private	Urba
6	1	53882	Male	74	1	1	Yes	Private	Rura
7	1	10434	Female	69	0	0	No	Private	Urba
8	1	60491	Female	78	0	0	Yes	Private	Urba
9	1	12109	Female	81	1	0	Yes	Private	Rura
10	1	12095	Female	61	0	1	Yes	Govt_job	Rura
11	1	12175	Female	54	0	0	Yes	Private	Urba
12	1	5317	Female	79	0	1	Yes	Private	Urba
13	1	58202	Female	50	1	0	Yes	Self-employed	Rura

Gambar 2. *Data Training*

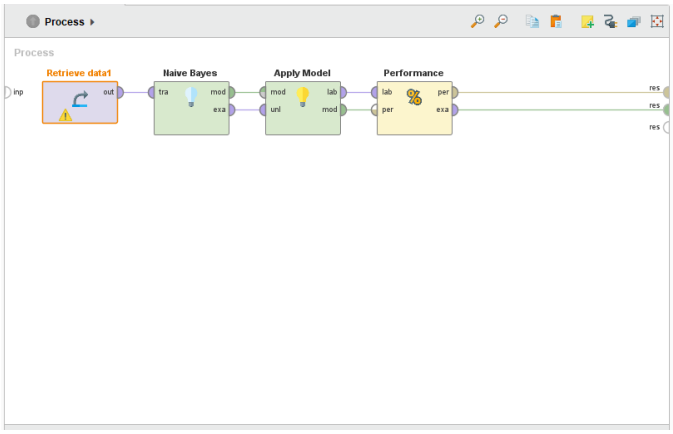
Pada gambar 2 merupakan proses *training* data. Pada data tersebut terdapat *class* 1 untuk "terkena stroke" dan *class* 0 untuk "tidak terkena stroke"

Gambar 3. Reprocessing data

Pada gambar 3 merupakan proses *reprocessing* data untuk menghilangkan data missing *value.* Dari proses ini ditemukan 202 data *missing value.*

ISSN: 2580-3336 (Print) / 2549-7952 (Online)

Url: https://proceeding.unpkediri.ac.id/index.php/inotek/



Gambar 4. Pengecekan data

Pada gambar 4 adalah proses menjalankan untuk mengetahui hasil dari dataset yang digunakan.

accuracy: 87.22%

	true 1	true 0	class precision
pred. 1	89	507	14.93%
pred. 0	120	4192	97.22%
class recall	42.58%	89.21%	

Gambar 5. Hasil *accuracy*

Pada gambar 5 tingkat *accuracy* nya 87.22%. Pada prediksi terkena stroke terdapat 89 data orang yang "terkena stroke" dan 507 data orang yang "tidak terkena stroke". Jumlah *precision* "terkena stroke" adalah 14.93% dan jumlah *recall* adalah 42.58%. Kemudian pada prediksi tidak terkena stroke terdapat 120 data orang yang "terkena stroke" dan 4192 data orang yang "tidak terkena stroke". Jumlah *precision* "tidak terkena stroke" adalah 97.22% dan jumlah *recall* 89.21%.

Selain di uji menggunakan *software rapidminer studio*, hasil *accuracy, precision,* dan *recall* bisa di hitung menggunakan rumus yang telah dijabarkan di atas. Berikut cara menghitung nya.

a. *Accuracy*

$$¿\frac{(TP+TN)}{¿¿}$$

$$¿\frac{89+4192}{(89+507+120+4192)}\,X\,100\,\%$$

$$¿\frac{4281}{4908}\,X\,100\,\%$$

$$¿0{,}8722\,X\,100\,\%$$

$$¿87{,}22\,\%$$

b. *Precision*

$$¿\frac{TP}{TP+FP}\,100\,\%$$

ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

$$¿\frac{89}{89+507}100\,\%$$

$$¿\frac{89}{596}100\,\%$$

$$¿14{,}93\,\%$$

c. *Recall*

$$¿\frac{TP}{TP+FP}100\,\%$$

$$¿\frac{89}{89+120}100\,\%$$

$$¿\frac{89}{209}100\,\%$$

$$¿0{,}4258\,x\,100\,\%$$

$$¿42{,}58\,\%$$

IV. KESIMPULAN

Berdasarkan hasil dari penelitian yang telah dilakukan, maka dapat disimpulkan bahwa algoritma yang digunakan untuk mengelola data menggunakan naive bayes. Untuk *tools* yang digunakan untuk perhitungan menggunakan software rapidminer. Dari dataset terdapat 5110 *record*, kemudian dilakukan *preprocessing* data ditemukan 202 data missing value, maka data yang diolah menjadi 4908 *record*. Kemudian hasil *accuracy* nya 87.22%, *precision* nya 14.93% dan *recall* nya 42.58%.

UCAPAN TERIMAKASIH

Dengan ini, kami sebagai penulis menyampaikan rasa terima kasih yang mendalam kepada semua pihak yang telah memberikan dukungan terhadap penelitian kami. Kami berterima kasih kepada semua yang telah berkontribusi dalam penelitian ini. Terima kasih juga kami sampaikan kepada para peneliti yang telah bekerja keras mengumpulkan data dan menganalisis hasil.Tanpa adanya dukungan dan kerjasama dari berbagai pihak, penelitian ini tidak akan dapat terlaksana. Sebagai penulis, kami juga ingin mengucapkan terima kasih kepada institusi kami atas dukungan dan fasilitas yang telah disediakan selama proses penelitian berlangsung. Kami berharap penelitian ini dapat memberikan manfaat yang signifikan, terutama dalam pemahaman tentang klasifikasi penyakit stroke.Melalui penelitian ini, kami berharap bahwa temuan-temuan yang diperoleh dapat memberikan kontribusi yang berarti dalam mengembangkan pengetahuan mengenai klasifikasi penyakit stroke yang akurat.

DAFTAR PUSTAKA

[1] Y. Azhar, A. Khoiriyah Firdausy, and P. J. Amelia, “SINTECH Journal | 191 Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke”, [Online]. Available: https://doi.org/10.31598

[2] J. Teknika and R. Estian Pambudi, "Teknika 16 (02): 221-226," *IJCCS*, vol. x, No.x, pp. 1–5.

[3] F. Karim, G. W. Nurcahyo, and S. Sumijan, "Sistem Pakar dalam Mengidentifikasi Gejala Stroke Menggunakan Metode Naive Bayes," *Jurnal Sistim Informasi dan Teknologi*, pp. 221–226, Aug. 2021, doi: 10.37034/jsisfotek.v3i4.69.

[4] J. Homepage, F. Akbar, H. Wira Saputra, A. Karel Maulaya, and M. Fikri Hidayat, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Implementation of Decision Tree Algorithm C4.5 and Support Vector Regression for Stroke Disease Prediction Implementasi Algoritma Decision Tree C4.5 dan Support Vector Regression untuk Prediksi Penyakit Stroke," vol. 2, pp. 61–67, 2022.

[5] N. Yolanda Paramitha *et al.*, "Klasifikasi Penyakit Stroke Menggunakan Metode Naïve Bayes," 2023. [Online]. Available: https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-

[6] Moh. Imron, J. Junal, and E. Masnawati, "Wacana Rubrik Kriminal di Media Daring Jawa Pos Radar Madura," *Stilistika: Jurnal Pendidikan Bahasa dan Sastra*, vol. 15, no. 1, p. 99, Jan. 2022, doi: 10.30651/st.v15i1.10597.

[7] Mm. Ali, T. Hariyati, M. Yudestia Pratiwi, and S. Afifah Sekolah Tinggi Agama Islam Ibnu Rusyd Kotabumi, "Metodologi Penelitian Kuantitatif Dan Penerapan Nya Dalam Penelitian."

[8] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, p. 437, Apr. 2020, doi: 10.30865/mib.v4i2.2080.

[9] N. B. Putri and A. W. Wijayanto, "Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing," *Komputika : Jurnal Sistem Komputer*, vol. 11, no. 1, pp. 59–66, Jan. 2022, doi: 10.34010/komputika.v11i1.4350.

[10] A. Algoritma, K. Pada, S. Rapidminer, and W. Ainurrohmah, "Akurasi Algoritma Klasifikasi pada Software Rapidminer dan Weka," *Prosiding Seminar Nasional Matematika*, vol. 4, pp. 493–499, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[11] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," 2021.

[12] "6552-21016-1-PB".

[13] N. A'ayunnisa, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indonesian Journal of Data and Science (IJODAS)*, vol. 3, no. 3, pp. 115–121, 2022.

[14] Suryani and Mustakim, "Estimasi Keberhasilan Siswa dalam Pemodelan Data Berbasis Learning Menggunakan Algoritma Support Vector Machine," *Bulletin of Informatics and Data Science*, vol. 1, no. 2, 2022, [Online]. Available: https://ejurnal.pdsi.or.id/index.php/bids/index

[15] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, "Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako," *JURIKOM (Jurnal Riset Komputer)*, vol. 8, no. 6, p. 219, Dec. 2021, doi: 10.30865/jurikom.v8i6.3655.