



OPTIMASI ALGORITMA NAÏVE BAYES MENGGUNAKAN FITUR SELEKSI BACKWARD ELIMINATION UNTUK KLASIFIKASI PREVALENSI STUNTING

Muhajir Yunus^{1)*}, Muhammad Kunta Biddinika¹⁾, Abdul Fadlil¹⁾

¹ Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Email: muhajiryunus@gmail.com

Abstrak

Stunting adalah masalah kekurangan gizi kronis yang ditandai dengan tinggi badan anak di bawah normal untuk usianya. Anak yang mengalami stunting memiliki risiko lebih tinggi terhadap berbagai penyakit kronis dan masalah kesehatan lainnya dan cenderung memiliki *intelligence quotient* yang lebih rendah dan performa yang buruk di sekolah karena sebanyak 90% jumlah sel otak tercipta sejak dalam kandungan hingga anak berumur 24 bulan. Tujuan dari penelitian ini adalah untuk mengklasifikasi prevalensi stunting pada anak usia di bawah 5 tahun dengan mengimplementasikan metode *naïve bayes* menggunakan fitur seleksi *backward elimination* berdasarkan data perhitungan z-score dengan data sampel berjumlah 224 *record*, yang terdiri dari 4 atribut dan 1 label yaitu jenis kelamin, usia, berat badan, tinggi badan dan status gizi. Dari hasil penelitian yang telah dilakukan diperoleh nilai akurasi tertinggi sebesar 92,54% sedangkan hasil dari pengujian model tanpa menggunakan seleksi fitur mendapatkan akurasi sebesar 53,50%. Penelitian ini menggunakan data training dan testing dengan ratio sebesar 70%:30%.

Kata kunci: backward elimination; data mining; machine learning; naïve bayes; stunting.

NAÏVE BAYES ALGORITHM OPTIMIZATION USING BACKWARD ELIMINATION SELECTION FEATURE FOR STUNTING PREVALENS CLASSIFICATION

Abstract

Stunting is a chronic nutritional deficiency problem characterized by a child's below-normal height for their age. Children who experience stunting have a higher risk of various chronic diseases and other health problems. They tend to have a lower intelligence quotient and perform poorly in school because around 90% of brain cells are developed from conception to the age of 24 months. The aim of this research is to classify the prevalence of stunting in children under 5 years old by implementing the naïve Bayes method using backward elimination feature selection based on z-score calculation data. The sample data consists of 224 records, including 4 attributes (gender, age, weight, height) and 1 label (nutritional status). The research yielded the highest accuracy value of 92.54% when using feature selection, whereas testing the model without feature selection resulted in an accuracy of 53.50%. The research employed a training and testing data ratio of 70%:30%.

Keywords: backward elimination; data mining; machine learning; naïve bayes; stunting.

Submitted: 27 April 2023

Reviewed: 1 Mei 2023

Accepted: 6 Juni 2023

Published: 12 Juli 2023

PENDAHULUAN

Stunting adalah salah satu bentuk malnutrisi yang terjadi pada anak-anak yang disebabkan oleh kurangnya asupan gizi yang seimbang sejak masa kehamilan hingga usia 2 tahun (Ponum et al. 2020). Masalah stunting pada anak meliputi faktor biologis, ekonomi, sosial dan lingkungan (Mutiarasari et al. 2021). Beberapa faktor yang mempengaruhi stunting antara lain kurangnya akses terhadap makanan bergizi, air bersih dan sanitasi yang buruk, serta kurangnya pengetahuan tentang gizi dan perawatan Kesehatan (Imam et al. 2021). Stunting dapat menyebabkan masalah kesehatan jangka Panjang dan mempengaruhi perkembangan intelektual dan sosial-emosional anak (Fentiana et al. 2022). Tahun 2014, World Health Organization (WHO) menyatakan terdapat 162 juta anak di bawah usia 5 tahun yang menderita stunting. WHO juga memprediksi pada tahun 2025 akan ada tambahan 125 juta penderita stunting jika anak tidak tertangani (World Health Organization (WHO) 2014). Stunting memiliki efek jangka Panjang dan berpotensi menjadi penyakit degenerative atau penyakit keturunan seperti diabetes (Anita, Purwati, and Desmarnita 2021).

Teknologi informasi menyediakan kemudahan bagi setiap orang untuk mengakses dan memperoleh informasi dari berbagai sumber (Bai et al. 2020). Teknologi seperti internet dan perangkat mobile membuat informasi dapat diakses dengan cepat dan mudah, tanpa batasan waktu dan tempat (Demir, Döven, and Sezen 2019). Namun hal ini juga menimbulkan tantangan baru, seperti memastikan keabsahan dan keakuratan informasi atau data yang diperoleh. Data semakin tumbuh dan berkembang cepat, namun tidak sebanding manfaatnya. Data harus diolah mejadi pengetahuan agar memberikan manfaat yang besar (Sima et al. 2020).

Data mining adalah disiplin ilmu yang mempelajari metode untuk mengekstraksi informasi dari data (Fernandes et al. 2019). Data yang sedemikian besar tentunya memiliki informasi yang tersembunyi di dalamnya, namun kemampuan manusia terbatas dalam menganalisis data atau menggali pengetahuan dari data tersebut (Wahono, 2016). Pengetahuan tersebut tentunya sangat berguna untuk mendukung pengambilan kebijakan atau keputusan. Selain itu, kemampuan komputasi yang semakin canggih dan terjangkau, serta persaingan bisnis yang semakin kompetitif merupakan faktor-faktor lainnya mengapa Data Mining semakin berperan dalam mendukung pengambilan keputusan (Asaad and Abdulhakim, 2021).

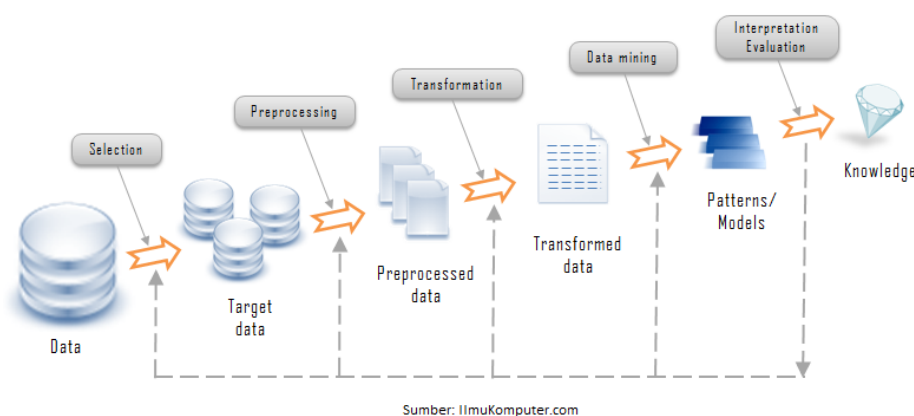
Naïve Bayes adalah sebuah algoritma klasifikasi probabilistik yang menggunakan teorema bayes untuk meperkirakan kemungkinan suatu kelas berdasarkan suatu input. Algoritma Naïve Bayes umumnya digunakan untuk mengklasifikasikan teks, seperti email spam atau dokumen yang menggunakan fitur seperti kata-kata yang terkandung dalam teks tersebut (Chen et al. 2020). Namun kelemahan algoritma Naïve Bayes yang perlu dipertimbangkan jika data yang tersedia sangat sedikit dan terdapat fitur yang tidak terwakili dalam data latih dapat mempengaruhi tingkat akurasi. *Backward elimination* adalah salah satu teknik dalam pemodelan statistik untuk memilih variabel yang paling signifikan atau yang paling penting dalam suatu model untuk meningkatkan akurasi dari proses klasifikasi dengan cara kerja sistem pemilihan mundur (Deng et al. 2019).

Beberapa penelitian terkait dengan masalah stunting seperti pada penelitian yang dilakukan oleh (Chilyabanyama et al. 2022) menunjukkan bahwa algoritma *machine learning* random forest yang memiliki akurasi prediksi tertinggi untuk stunting di dibandingkan dengan model algoritma yang lain. Penelitian yang dilakukan oleh (Islam et al. 2022) menunjukkan klasifikasi menggunakan metode random forest mendapatkan akurasi 81,4% dan 0,837 AUC untuk kekurangan berat badan dan akurasi 82,4% dan 0,853 AUC untuk kelebihan berat badan.

Penelitian lain yang dilakukan oleh (Bitew, Sparks, and Nyarko 2022) Algoritma *xgbTree* adalah algoritma machine learning yang cukup unggul untuk memprediksi kekurangan gizi masa kanak-kanak di Ethiopia di dibandingkan dengan metode machine learning lainnya. Maka dari penelitian terdahulu, penulis tertarik untuk meneliti implementasi gaussian untuk klasifikasi kejadian stunting pada anak usia di bawah lima tahun menggunakan algoritma *Naïve Bayes*.

METODE

Tahapan yang digunakan dalam penelitian ini menggunakan proses *knowledge discovery in databases* (KDD) seperti yang di ilustrasikan pada gambar dibawah ini.



Gambar 1. Tahapan Penelitian

Berdasarkan Gambar 1 menjelaskan bahwa tahapan penelitian ini di bagi menjadi beberapa tahap. Pertama peneliti mengumpulkan data bertujuan untuk mengidentifikasi kebutuhan entitas dan atribut data sesuai tujuan teknis data science. Dataset yang digunakan adalah data survey prevalensi stunting di daerah Kabupaten Gorontalo yang terdiri dari 4 atribut dan 1 label yaitu Jenis Kelamin, Usia, Berat Badan, Tinggi Badan dan Status Gizi, dengan rincian 122 data Normal dan 122 data Stunting.

Tabel 1. Dataset Klasifikasi Prevalensi Stunting

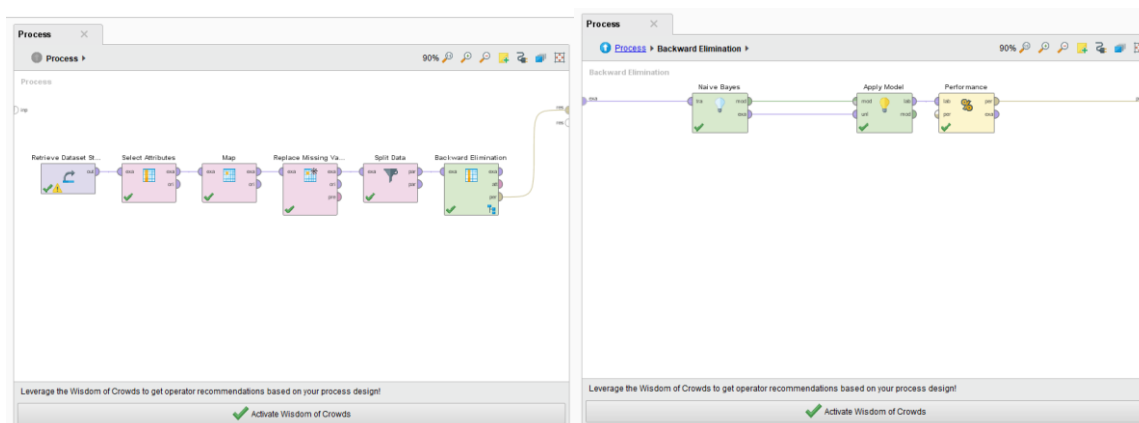
No.	Gender	Age	Weight	Height	Nutrition
1	Perempuan	26	10	84	Normal
2	Laki-Laki	18	8,5	76	Normal
3	Perempuan	38	10.1	89	Stunting
4	Laki-Laki	21	10.5	78	Stunting
....
224	Laki-Laki	17	9	76	Stunting

Tahap kedua Preprocessing. Tahapan ini bertujuan untuk melakukan pembersihan data yang kotor yaitu mencoba untuk mengisi nilai yang hilang, menghaluskan *noise* saat mengidentifikasi adanya *outlier* pada data dan memperbaiki inkonsistensi dalam data. Tahap ketiga transformasi data. Tahap ini memberikan inisialisasi terhadap data yang memiliki nilai akan disesuaikan tipe datanya seperti mengubah jenis kelamin Laki-Laki menjadi “L” dan Perempuan menjadi “P” kemudian untuk tipe datanya dirubah menjadi *binomial* karena hanya mempunyai dua kategori yaitu Laki-laki dan Perempuan, begitu juga dengan label/kelas target yaitu normal dan stunting. Kemudian untuk data kontinu tipe datanya dirubah menjadi *polynomial* karena mempunyai nilai lebih dari dua.

Row No.	Nutrition (binomial) label	Gender (binomial) regular	Age (Integer) regular	Weight (polynomial) regular	Height (polynomial) regular
103	Normal	Perempuan	4	11	92
104	Normal	Perempuan	4	10.3	89
105	Normal	Laki-Laki	3	9	81
106	Normal	Perempuan	11	7.5	71
107	Normal	Perempuan	45	11	93
108	Normal	Perempuan	30	10	84
109	Normal	Perempuan	32	10.3	85
110	Normal	Perempuan	30	9	84
111	Normal	Perempuan	10	7.4	68
112	Normal	Perempuan	38	10.1	89
113	Stunting	Laki-Laki	28	9	81
114	Stunting	Perempuan	44	11.2	90
115	Stunting	Laki-Laki	30	10.5	82
116	Stunting	Perempuan	39	11	87
117	Stunting	Perempuan	39	10.5	89
118	Stunting	Laki-Laki	22	8.8	80
119	Stunting	Laki-Laki	22	8.8	80

Gambar 2. Transformasi Data

Tahap ketiga adalah pembuatan model. Pada tahap ini peneliti mengimplementasikan algoritma *Naïve Bayes* sebagai metode untuk mengklasifikasikan data. Peneliti menggunakan beberapa operator pada RapidMiner untuk menganalisa data seperti “Select Attributes” operator ini memilih atribut mana dari sebuah *dataset* yang harus dipertahankan dan atribut mana yang harus dihapus. Ini digunakan dalam kasus ketika tidak semua atribut dari sebuah *dataset* diperlukan. Kemudian “Map” operator ini memetakan ke nilai tertentu dari atribut tersebut ke nilai baru sama halnya dengan melakukan transformasi data. Kemudian “Replace Missing Values” operator ini mengganti nilai yang hilang dalam contoh atribut yang dipilih dengan penggantian tertentu. “Split Data” operator ini menghasilkan jumlah himpunan bagian yang diinginkan. Penelitian ini membagi data training dan testing dengan ratio 70% : 30%. Kemudian Backward “Elimination” operator ini bertujuan untuk memilih variabel yang paling berpengaruh terhadap variabel target atau variabel dependent. “Apply Model” operator ini menerapkan model yang sudah dipelajari atau dilatih pada *dataset*. Kemudian “Performance” operator ini digunakan untuk mengevaluasi kinerja dari model.



Gambar 3. Pemodelan Naïve Bayes menggunakan *Backward Elimination*

Persamaan teorema Bayes didasarkan pada rumus umum berikut (Singh et al. 2019):

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \dots \dots \dots (1)$$

Keterangan

X : Data dengan class yang belum diketahui

H : Hipotesis data X merupakan suatu class spesifik

P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probability)

- P(H) : Probabilitas hipotesis H (prior probability)
- P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
- P(X) : Probabilitas X

Ide dasar di balik aturan Bayes adalah bahwa hasil dari suatu hipotesis atau peristiwa (H) dapat diprediksi berdasarkan beberapa bukti yang diamati (X). Secara umum, naive bayes untuk atribut tipe kategorikal mudah dihitung. Namun, ada perlakuan khusus untuk fitur numerik (kontinu) sebelum diintegrasikan ke dalam Naive Bayes, yaitu melalui penggunaan *probability density function*. (Nagesh Singh Chauhan 2020).

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \dots\dots\dots(2)$$

$$\sigma = \left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2 \right]^{0.5} \dots\dots\dots(3)$$

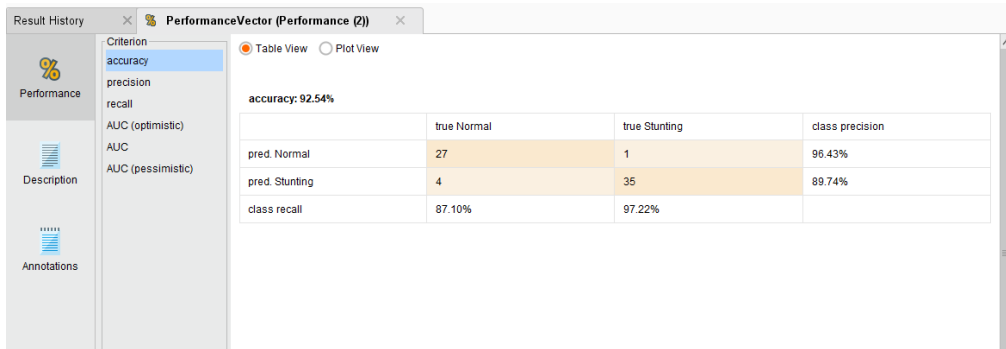
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \dots\dots\dots(4)$$

Keterangan

- μ : Mean
- σ : Standard deviation
- $f(x)$: Normal distribution

HASIL DAN PEMBAHASAN

Berdasarkan hasil evaluasi model algoritma naive bayes menggunakan seleksi fitur *Backward Elimination*, diperoleh akurasi tertinggi sebesar 92.54% seperti pada gambar berikut.



Gambar 4. Performa Model Naïve Bayes menggunakan *Backward Elimination*

Hasil optimasi model yang diperoleh dari penerapan seleksi fitur terhadap data prevalensi stunting, dapat dilihat dari hasil perbandingan pada skenario pengujian yang telah dilakukan seperti pada Tabel 2.

Tabel 2. Hasil Perbandingan Skenario Pengujian

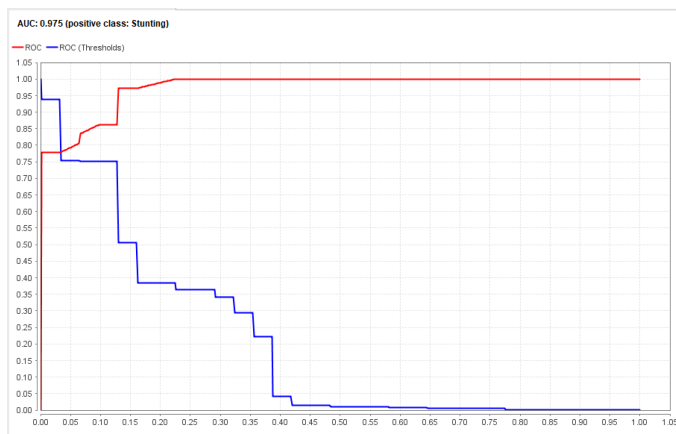
	Skenario 1	Skenario 2
Algoritma	NB + BE	NB
Akurasi %	92,54	53,50
Precision %	89,74	51,65
Recall %	97,22	61,84
AUC	0,975	0,523

Keterangan

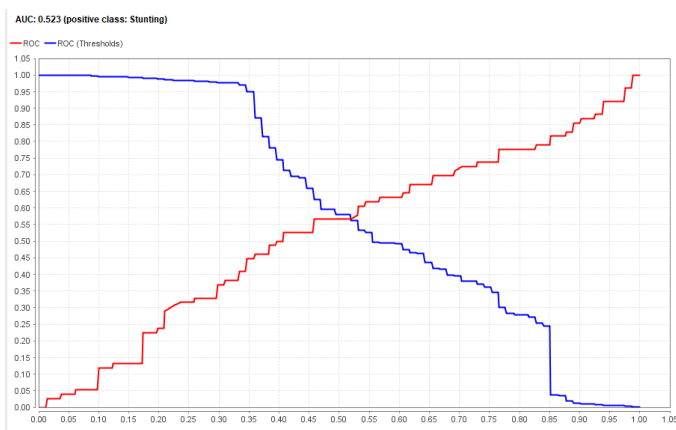
NB : Naïve Bayes

BE : *Backward Elimination*

Berdasarkan hasil pengujian menggunakan skenario 1 yaitu model Naïve Bayes menggunakan seleksi fitur *backward elimintaion* memiliki akurasi tertinggi sebesar 92.54%. Sedangkan akurasi terendah terdapat pada skenario 2 yaitu model Naïve Bayes tanpa seleksi fitur *backward elimination* dengan nilai akurasi 53.50%. Grafik area under curve (AUC) perbandingan skenario 1 dan 2 diilustrasikan pada gambar di bawah ini.



Gambar 5. Grafik AUC Skenario 1



Gambar 6. Grafik AUC Skenario 2

KESIMPULAN DAN SARAN

Hasil penelitian pada klasifikasi prevalensi stunting dengan mengimplementasikan algoritma Naïve Bayes menggunakan seleksi fitur *Backward Elimination* menghasilkan akurasi yang cukup signifikan. Penggunaan seleksi fitur dalam klasifikasi prevalensi stunting berhasil meningkatkan akurasi sebesar 39,04%.

Saran dari penelitian ini diharapkan untuk penelitian selanjutnya dapat menambahkan jumlah atribut dan *record* pada data.

DAFTAR PUSTAKA

- Anita, A., P. Purwati, and Ulty Desmarnita. 2021. "Book of Stunting Risk Detection and Monitoring Health (DRSMK) and Stunting Prevention Behavior in Children the First 1000 Days of Life." *Jurnal Aisyah: Jurnal Ilmu Kesehatan* 6(1), 279-284. 10.30604/jika.v6i1.605.
- Asaad, R. R., & Abdulhakim, R. M. (2021). The Concept of Data Mining and Knowledge Extraction Techniques. *Qubahan Academic Journal*, 1(2), 17-20. <https://doi.org/10.48161/qaj.v1n2a43>
- Bai, C., Dallasega, P., Orzes, G., & Sarkis, J. (2020). Industry 4.0 technologies assessment: A sustainability perspective. *International journal of production economics*, 229, 107776. <https://doi.org/10.1016/j.ijpe.2020.107776>
- Bitew, F. H., Sparks, C. S., & Nyarko, S. H. (2022). Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia. *Public health nutrition*, 25(2), 269-280. 10.1017/S1368980021004262.
- Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, 192, 105361. 10.1016/j.knosys.2019.105361.
- Chilyabanyama, O. N., Chilengi, R., Simuyandi, M., Chisenga, C. C., Chirwa, M., Hamusonde, K., ... & Bosomprah, S. (2022). Performance of machine learning classifiers in classifying stunting among under-five children in Zambia. *Children*, 9(7), 1-11. <https://doi.org/10.3390/children9071082>
- Demir, K. A., Döven, G., & Sezen, B. (2019). Industry 5.0 and human-robot co-working. *Procedia computer science*, 158, 688-695.
- Deng, X., Li, Y., Weng, J., & Zhang, J. (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78(5), 3797-3816. 10.1007/s11042-018-6083-5.
- Fentiana, N., Achadi, E. L., Besral, B., Kamiza, A., & Sudiarti, T. (2022). A Stunting Prevention Risk Factors Pathway Model for Indonesian Districts/Cities with a Stunting Prevalence of $\geq 30\%$. *Kesmas: Jurnal Kesehatan Masyarakat Nasional (National Public Health Journal)*, 17(3), 175-183. <http://dx.doi.org/10.21109/kesmas.v17i3.5954>
- Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public-school students in the capital of Brazil. *Journal of business research*, 94, 335-343. <https://doi.org/10.1016/j.jbusres.2018.02.012>
- Imam, A., Hassan-Hanga, F., Sallahdeen, A., & Farouk, Z. L. (2021). A cross-sectional study of prevalence and risk factors for stunting among under-fives attending acute malnutrition treatment programmes in north-western Nigeria: Should these programmes

be adapted to also manage stunting?. *International Health*, 13(3), 262-271. <https://doi.org/10.1093/inthealth/ihaa043>

- Islam, Md Merajul, Md Jahanur Rahman, Md Moidul Islam, Dulal Chandra Roy, N. A. M. Faisa. Ahmed, Sadiq Hussain, Md Amanullah, Md Menhazul Abedin, and Md Maniruzzaman. 2022. "Application of Machine Learning Based Algorithm for Prediction of Malnutrition among Women in Bangladesh." *International Journal of Cognitive Computing in Engineering* 3:46–57. doi: 10.1016/j.ijcce.2022.02.002.
- Mutiarasari, D., Miranti, M., Fitriana, Y., Pakaya, D., Sari, P., Bohari, B., ... & Hadju, V. (2021). A determinant analysis of stunting prevalence on under 5-year-old children to establish stunting management policy. *Open Access Macedonian Journal of Medical Sciences*, 9(B), 79-84. <https://doi.org/10.3889/oamjms.2021.5622>
- Nagesh Singh Chauhan. 2020. "Naïve Bayes Algorithm: Everything You Need to Know." Retrieved (<https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html>).
- Ponum, M., Khan, S., Hasan, O., Mahmood, M. T., Abbas, A., Iftikhar, M., & Arshad, R. (2020). Stunting diagnostic and awareness: Impact assessment study of sociodemographic factors of stunting among school-going children of Pakistan. *Bmc Pediatrics*, 20, 1-9.
- Sima, V., Gheorghe, I. G., Subić, J., & Nancu, D. (2020). Influences of the industry 4.0 revolution on the human capital development and consumer behavior: A systematic review. *Sustainability*, 12(10), 4035. <https://doi.org/10.3390/su12104035>
- Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019, April). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 593-596).
- Wahono, R. S. (2016) "Literature Review: Pengantar Dan Metode." Retrieved (<https://romisatriawahono.net/2016/05/07/literature-review-pengantar-dan-metode/>).
- World Health Organization (WHO). 2014. *Global Nutrition Targets 2025*. Vol. 122. World Health Organization.

How to cite:

Yunus, M., Biddinika, M. K., & Fadlil, A. (2023). Optimasi Algoritma Naïve Bayes Menggunakan Fitur Seleksi Backward Elimination untuk Klasifikasi Prevalensi Stunting. *DECODE: Jurnal Pendidikan Teknologi Informasi*, 3(2), 278-285. <http://dx.doi.org/10.51454/decode.v3i2.188>