

IMPLEMENTASI ALGORITMA CATBOOST UNTUK KLASIFIKASI KELOMPOK PENDAPATAN NEGARA (INCOME GRUP) BERDASARKAN INDIKATOR MALNUTRISI GLOBAL

Vivi Apriya Mardani^{1*}, Aprilisa Arum Sari², Ridwan Dwi Irawan³

¹ Teknik Informatika

Universitas Duta Bangsa Surakarta

^{1*} 220103040@mhs.udb.ac.id

² Teknik Informatika

Universitas Duta Bangsa Surakarta

² aprilisa_arumsari@udb.ac.id

³ Teknik Informatika

Universitas Duta Bangsa Surakarta

³ ridwan_dwiirawan@udb.ac.id

Abstrak— Masalah malnutrisi masih menjadi salah satu tantangan global yang berdampak terhadap kualitas sumber daya manusia dan pembangunan ekonomi suatu negara. Berbagai indikator malnutrisi, seperti stunting, wasting, underweight, overweight, undernourishment, dan food insecurity, memiliki keterkaitan dengan kondisi sosial ekonomi dan tingkat pendapatan suatu negara. Penelitian ini bertujuan untuk mengimplementasikan algoritma CatBoost dalam mengklasifikasikan kelompok pendapatan negara (income group) berdasarkan indikator malnutrisi global. Data penelitian diperoleh dari World Health Organization (WHO) dan World Bank yang terdiri atas empat kategori kelompok pendapatan, yaitu Low Income, Lower-Middle-Income, Upper-Middle-Income, dan High Income. Metode penelitian menggunakan pendekatan machine learning dengan pembagian data sebesar 70% sebagai data pelatihan dan 30% sebagai data pengujian. Model dibangun menggunakan algoritma CatBoost dengan parameter iteration sebesar 100, depth sebesar 4, dan learning rate sebesar 0,05. Hasil penelitian menunjukkan bahwa model CatBoost memperoleh tingkat akurasi sebesar 93,33%. Selain itu, hasil classification rate menunjukkan nilai precision, recall, dan f1-score yang tinggi pada Sebagian besar kelas. Analisis feature importance menunjukkan bahwa indikator food insecurity moderate severe dan wasting merupakan variabel yang paling berpengaruh dalam menentukan kelompok pendapatan negara. Dengan demikian, algoritma CatBoost mampu mengklasifikasikan kelompok pendapatan negara berdasarkan indikator malnutrisi global secara efektif dan akurat.

Kata kunci— CatBoost, klasifikasi, income group, malnutrisi global, machine learning.

Abstract— Malnutrition remains a global challenge that affects human resource quality and a country's economic development. Several malnutrition indicators, including stunting, wasting, underweight, overweight, undernourishment, and food insecurity, are associated with socioeconomic conditions and national income levels. This study aims to implement the CatBoost algorithm to classify country income groups based on global malnutrition indicators. The dataset was obtained from the World Health Organization (WHO) and the World Bank and consists of four income categories, namely Low Income, Lower-Middle Income, Upper-Middle Income, and High Income. This study employed a machine learning approach using a 70:30 split for training and testing data. The CatBoost model was developed using 100 iterations, a depth of 4, and a learning rate of 0.05. The experimental results showed that the proposed model achieved an accuracy of 93.33%. Furthermore, the classification report indicated high precision, recall, and F1-score values across most classes. Feature importance analysis revealed that food insecurity and wasting were the most influential variables in determining country income groups. Therefore, the CatBoost algorithm is effective and accurate for classifying country income groups based on global malnutrition indicators.

Keywords— CatBoost, classification, income group, global malnutrition, machine learning

I. PENDAHULUAN

Masalah malnutrisi masih menjadi salah satu tantangan kesehatan global yang berdampak terhadap kualitas sumber daya manusia, produktivitas, serta pembangunan ekonomi suatu negara. Menurut World Health Organization (WHO), pada tahun 2023 sekitar 22,3% balita di dunia mengalami stunting, 6,8% mengalami wasting, dan 5,6% mengalami overweight. Selain itu, hampir 45% kematian anak di bawah usia lima tahun berkaitan dengan kondisi malnutrisi. Kondisi tersebut menunjukkan bahwa permasalahan gizi tidak hanya berdampak pada sektor kesehatan, tetapi juga berpengaruh terhadap kualitas tenaga kerja, tingkat kemiskinan, dan pertumbuhan ekonomi suatu negara [1], [2]. Oleh karena itu, indikator malnutrisi menjadi salah satu parameter yang penting untuk menggambarkan kondisi sosial ekonomi suatu negara.

World Bank mengelompokkan negara ke dalam empat kategori pendapatan, yaitu *Low Income*, *Lower-Middle Income*, *Upper-Middle Income*, dan *High Income* berdasarkan Gross National Income (GNI) per kapita [3]. Meskipun demikian, hubungan antara indikator malnutrisi dan kelompok pendapatan negara tidak bersifat linier karena dipengaruhi oleh berbagai faktor, seperti akses terhadap layanan kesehatan, ketahanan pangan, tingkat pendidikan, serta kebijakan pemerintah. Dataset yang menggambarkan kondisi tersebut umumnya memiliki karakteristik multivariat, terdiri atas beberapa indikator yang saling berkaitan, serta berpotensi memiliki distribusi kelas yang tidak seimbang. Karakteristik tersebut menjadikan proses klasifikasi kelompok pendapatan negara sebagai permasalahan yang cukup kompleks sehingga memerlukan pendekatan machine learning yang mampu mempelajari pola data secara efektif.

Berbagai penelitian sebelumnya telah menerapkan algoritma machine learning untuk klasifikasi maupun prediksi pada bidang kesehatan dan sosial ekonomi. Namun, sebagian

besar penelitian masih berfokus pada prediksi status gizi atau klasifikasi penyakit, sedangkan penelitian yang menghubungkan indikator malnutrisi global dengan klasifikasi kelompok pendapatan negara masih relatif terbatas. Selain itu, evaluasi performa model sering kali hanya didasarkan pada satu kali pembagian data (*hold-out validation*), sehingga hasil akurasi yang diperoleh berpotensi dipengaruhi oleh komposisi data latih dan data uji. Oleh karena itu, pemilihan metode klasifikasi yang mampu menghasilkan performa tinggi sekaligus mampu menangani karakteristik data yang kompleks menjadi aspek penting dalam penelitian ini.

Salah satu algoritma yang memiliki kemampuan tersebut adalah CatBoost. Algoritma ini merupakan metode *gradient boosting* yang dirancang untuk menghasilkan model dengan tingkat akurasi yang tinggi, mengurangi risiko *overfitting*, serta mampu mempelajari hubungan nonlinier antarvariabel secara efisien [4], [5]. Dengan memanfaatkan indikator malnutrisi global yang diperoleh dari World Health Organization (WHO) dan kelompok pendapatan negara dari World Bank, penelitian ini bertujuan mengimplementasikan algoritma CatBoost untuk mengklasifikasikan kelompok pendapatan negara berdasarkan indikator malnutrisi global. Hasil penelitian diharapkan dapat memberikan kontribusi sebagai referensi dalam pemanfaatan machine learning pada analisis kesehatan masyarakat dan pembangunan ekonomi, sekaligus menjadi dasar pengembangan model klasifikasi yang lebih akurat dan andal.

II. METODOLOGI PENELITIAN

2.1. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan teknik data mining menggunakan algoritma CatBoost untuk mengklasifikasikan kelompok pendapatan negara (*income group*) berdasarkan indikator malnutrisi global. Algoritma CatBoost dipilih karena memiliki kemampuan yang baik dalam menangani data yang kompleks serta

menghasilkan performa klasifikasi yang tinggi. Penelitian ini menggunakan metodologi CRISP-DM (Cross Industry Standard Process for Data Mining) yang terdiri atas enam tahapan, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Proses implementasi dilakukan menggunakan bahasa pemrograman Python pada Google Colab dengan bantuan library Pandas, NumPy, Scikit-learn, CatBoost, Matplotlib, dan Seaborn.

2.2. Jenis dan Sumber Data

Penelitian ini menggunakan data sekunder yang diperoleh dari World Health Organization (WHO) dan World Bank. Variable yang digunakan meliputi Stunting, Wasting, Underweight, Overweight, Undernourishment, dan Food Insecurity Moderate Severe sebagai variabel independent, sedangkan income group (High Income, Upper-Middle Income, Lower-Middle Income, dan Low Income) digunakan sebagai variabel dependen.

2.3. Metode Pengumpulan Data

Metode pengumpulan data dilakukan melalui studi dokumentasi dan studi literatur. Studi dokumentasi dilakukan dengan mengunduh dataset dari WHO dan World Bank, sedangkan studi literatur dilakukan dengan mengkaji jurnal dan artikel ilmiah yang berkaitan dengan algoritma CatBoost, machine learning, dan indikator malnutrisi global.

2.4. Tahapan Penelitian

Penelitian ini menggunakan metodologi CRISP-DM yang terdiri atas enam tahapan sebagai berikut.

1. Business Understanding

Mengidentifikasi permasalahan penelitian serta menentukan tujuan untuk mengklasifikasikan kelompok pendapatan negara berdasarkan indikator malnutrisi global.

2. Data Understanding

Mengumpulkan dan memahami karakteristik dataset yang diperoleh dari

WHO dan World Bank serta melakukan analisis awal terhadap data.

3. Data Preparation

Melakukan pembersihan data, pemilihan variabel, serta pembagian dataset menjadi 70% data training dan 30% data testing.

4. Modeling

Membangun model klasifikasi menggunakan algoritma CatBoost dengan parameter yang telah ditentukan.

5. Evaluation

Mengevaluasi performa model menggunakan metrik Accuracy, Precision, Recall, F1-Score, serta Confusion Matrix.

6. Deployment

Menyajikan hasil klasifikasi dan analisis Feature Importance sebagai interpretasi terhadap faktor-faktor yang memengaruhi kelompok pendapatan negara.

2.5. Tools Penelitian

Penelitian ini menggunakan Google Colaboratory sebagai lingkungan pengembangan dengan bahasa pemrograman python serta library Pandas, NumPy, Scikit-learn, CatBoost, Matplotlib, dan Seaborn untuk proses pengolahan data, pemodelan, evaluasi, dan visualisasi hasil penelitian.

III. HASIL DAN PEMBAHASAN

3.1 Deskripsi Dataset

Dataset yang digunakan pada penelitian ini merupakan dataset indikator malnutrisi global yang terdiri dari beberapa variabel persentase kondisi gizi di berbagai negara. Variabel target yang digunakan adalah income group, yaitu kelompok pendapatan negara yang dibagi menjadi empat kategori, meliputi *High Income*, *Upper-Middle Income*, *Lower-Middle Income*, dan *Low Income*. Sedangkan variabel prediktor yang digunakan terdiri atas enam indikator malnutrisi, yaitu persentase stunting, wasting, underweight, overweight, undernourishment, dan food insecurity.

Masing-masing variabel independen dinyatakan dalam bentuk persentase sehingga dapat menggambarkan kondisi gizi dan ketahanan pangan pada setiap negara. Variabel *Stunting-pct*, *Wasting-pct*, dan *Underweight-pct* menunjukkan kondisi kekurangan gizi pada balita, sedangkan *Overweight-pct* menggambarkan masalah kelebihan gizi. Selain itu, *Undernourishment-pcr* dan *Food-insecurity-severe-pct* digunakan untuk mengukur tingkat kerawanan pangan dan kecukupan konsumsi pangan masyarakat. Variabel-variabel tersebut dipilih karena memiliki hubungan yang erat dengan kondisi sosial ekonomi dan tingkat pendapatan suatu negara sehingga relevan digunakan sebagai prediktor dalam proses klasifikasi kelompok pendapatan negara menggunakan algoritma CatBoost.

Tabel 1. Deskripsi Variabel Penelitian

Variabel	Keterangan
<i>Stunting_pct</i>	Persentase balita mengalami stunting
<i>Wasting_pct</i>	Persentase balita mengalami wasting
<i>Underweight_pct</i>	Persentase balita dengan berat badan kurang
<i>Overweight_pct</i>	Persentase balita dengan kelebihan berat badan
<i>Undernourishment_pct</i>	Persentase penduduk mengalami kekurangan gizi
<i>Food_insecurity_moderate_severe_pct</i>	Persentase penduduk mengalami kerawanan pangan
<i>Income_group</i>	Kelompok pendapatan negara

Penelitian ini menggunakan enam indikator malnutrisi sebagai variabel independen dan satu variabel target berupa kelompok pendapatan negara. Pemilihan variabel tersebut dilakukan karena indikator malnutrisi memiliki hubungan erat dengan kondisi sosial dan ekonomi suatu negara.

3.2 Pembagian Data Training dan Testing

Pada penelitian ini dilakukan pembagian data menggunakan metode Hold-Out Validation dengan proporsi 70% data training dan 30% data testing. Pembagian data dilakukan secara stratifikasi (*stratify*) agar proporsi setiap kelas pendapatan tetap seimbang pada data latih dan data uji

Tabel 2. Pembagian Dataset

Dataset	Jumlah Data	Persentase
Training	210,6	70%
Testing	91	30%
Total	301	100%

Sebagian besar data digunakan untuk proses pelatihan model CatBoost agar model mampu mempelajari pola hubungan antara indikator malnutrisi dan kelompok pendapatan negara. Sementara itu, data testing digunakan untuk mengevaluasi kemampuan model dalam melakukan prediksi terhadap data yang belum pernah dipelajari sebelumnya.

3.3 Implementasi Algoritma CatBoost

Model klasifikasi dibangun menggunakan algoritma CatBoost dengan parameter:

- Iteration = 100
- Depth = 4
- Learning Rate = 0,05
- Loss Function = MultiClass
- Random State = 42

Penggunaan parameter tersebut bertujuan untuk menghasilkan model yang mampu melakukan klasifikasi multi-kelas secara optimal tanpa mengalami overfitting.

Tabel 3. Parameter Algoritma CatBoost

Parameter	Nilai
Iteration	100
Depth	4
Learning Rate	0,05
Loss Function	MultiClass
Random State	42

Model CatBoost dibangun menggunakan 100 iterasi dengan kedalaman pohon keputusan sebesar 4 dan laju pembelajaran sebesar 0,05. Kombinasi parameter tersebut dipilih untuk memperoleh keseimbangan antara akurasi model dan kompleksitas proses pembelajaran.

3.4 Hasil Evaluasi Model

Evaluasi model dilakukan menggunakan metrik Accuracy, Precision, Recall, dan f1-Score.

Tabel 4. Hasil Akurasi Model

Metode	Akurasi
CatBoost	93,33%

Berdasarkan hasil pengujian, algoritma CatBoost memperoleh tingkat akurasi sebesar 93,33%. Nilai tersebut menunjukkan bahwa model mampu mengklasifikasikan kelompok pendapatan negara dengan tingkat ketepatan yang baik berdasarkan indikator malnutrisi global yang digunakan.

Tabel 5. Classification Report

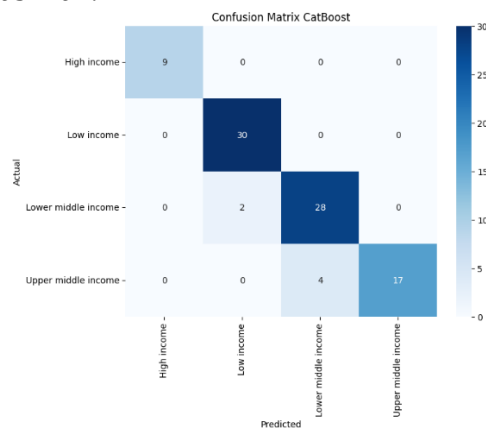
Kelas	Precision	Recall	F1-Score
High Income	1,00	1,00	1,00

Low Income	0,94	1,00	0,97
Lower-Middle Income	0,88	0,93	0,90
Upper-Middle Income	1,00	0,81	0,89

Nilai precision menunjukkan kemampuan model dalam memberikan prediksi yang tepat pada masing-masing kelompok pendapatan. Nilai recall menggambarkan kemampuan model dalam mengenai seluruh data pada suatu kelas, sedangkan F1-score menunjukkan keseimbangan antara precision dan recall. Semakin tinggi ketiga nilai tersebut, maka semakin baik performa model dalam melakukan klasifikasi.

3.5 Analisis Confusion Matrix

Menunjukkan confusion matrix hasil klasifikasi algoritma CatBoost. Nilai pada diagonal utama menunjukkan jumlah data yang berhasil diklasifikasikan secara benar pada masing-masing kelompok pendapatan. Sementara nilai di luar diagonal menunjukkan jumlah data yang mengalami kesalahan klasifikasi. Semakin besar nilai pada diagonal utama dan semakin kecil nilai di luar diagonal, maka semakin baik performa model yang dihasilkan.

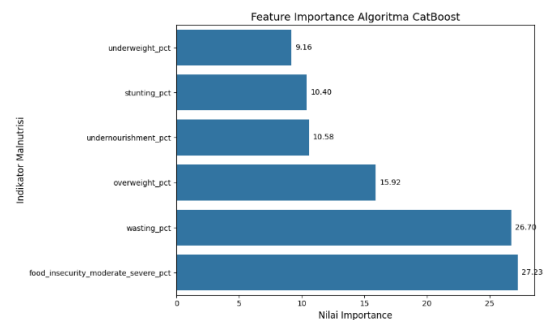


Gambar 1. Confusion Matrix Algoritma CatBoost

3.6 Grafik Feature Importance

Variabel dengan nilai importance tinggi merupakan indikator yang paling berpengaruh dalam menentukan kelompok pendapatan suatu negara. Sebaliknya, variabel dengan nilai importance terendah memiliki kontribusi yang relatif kecil terhadap proses klasifikasi. Hasil ini

menunjukkan bahwa kondisi malnutrisi dan krawanan pangan memiliki keterkaitan yang erat dengan tingkat pendapatan suatu negara.



Gambar 2. Grafik Feature Importance

IV. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma CatBoost berhasil diimplementasikan untuk mengklasifikasikan kelompok pendapatan negara (income group) berdasarkan indikator malnutrisi global yang meliputi stunting, wasting, underweight, overweight, undernourishment, dan food insecurity. Model yang dibangun menggunakan pembagian data sebesar 70% data pelatihan dan 30% data pengujian mampu menghasilkan pembagian data sebesar 93,33%, yang menunjukkan bahwa model memiliki performa klasifikasi yang sangat baik. Hasil evaluasi menggunakan classification report juga menunjukkan nilai precision, recall, dan f1-score yang tinggi pada Sebagian besar kelas, sehingga model mampu melakukan prediksi secara konsisten pada setiap kelompok pendapatan negara. Selain itu, analisis feature importance menunjukkan bahwa indikator food insecurity moderate severe dan wasting merupakan variabel yang paling berpengaruh dalam menentukan kelompok pendapatan suatu negara. Dengan demikian, algoritma CatBoost dapat digunakan sebagai metode yang efektif dalam mengidentifikasi pola hubungan antara indikator malnutrisi dan kelompok pendapatan negara serta berpotensi menjadi alat pendukung dalam analisis dan perumusan kebijakan pembangunan di bidang Kesehatan dan ekonomi.

UCAPAN TERIMA KASIH

Bagian ini memberikan apresiasi kepada perorangan maupun organisasi yang memberikan bantuan kepada penulis dan atau kepada pihak sponsor apabila ada.

REFERENSI

- [1] World Health Organization. (2023). *Levels and trends in child malnutrition: Key findings of the 2023 edition*. Geneva: World Health Organization.
- [2] UNICEF. (2023). *The state of the world's children 2023: For every child, nutrition*. New York: UNICEF.
- [3] World Bank. (2024). *World bank country and lending groups*. Washington, DC: World Bank.
- [4] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.
- [5] Hancock, J. T., & Khoshgoftaar, T. M. (2022). CatBoost for big data: An interdisciplinary review. *Journal of Big Data*, 9(1), 1–24.
- [6] Pahlevi, A., et al. (2024). Model klasifikasi risiko stunting pada balita menggunakan algoritma CatBoost classifier. *Bulletin of Computer Science Research*.
- [7] Handayani, R., et al. (2024). Machine learning klasifikasi status gizi balita menggunakan Random Forest. *Jurnal Teknologi Informasi*.
- [8] Sokolova, M., & Lapalme, G. (2022). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 59(2).
- [9] Zhang, Y., Liu, X., & Chen, J. (2022). Machine learning applications in healthcare data analysis. *Journal of Healthcare Engineering*, 2022.
- [10] Rahman, A., Hossain, M., & Karim, M. (2023). Classification techniques in data mining: A review. *International Journal of Data Science*, 5(2), 45–60.
- [11] World Health Organization, *World Health Statistics 2023: Monitoring health for the SDGs*, Geneva: World Health Organization, 2023.
- [12] FAO, IFAD, UNICEF, WFP and WHO, *The State of Food Security and Nutrition in the World 2023: Urbanization, agrifood systems transformation and healthy diets across the rural–urban continuum*, Rome: FAO, 2023.
- [13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 6638–6648, 2018.
- [14] S. Akter and M. R. Islam, "Global malnutrition and its impact on economic development: A machine learning perspective," *Global Health Action*, vol. 16, no. 1, p. 2198032, 2023.
- [15] N. T. Pham and H. V. Nguyen, "Application of CatBoost algorithm in handling complex categorical data for economic classification," *IEEE Access*, vol. 11, pp. 45210–45222, 2023.
- [16] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *Journal of Big Data*, vol. 7, no. 1, pp. 1–45, Dec. 2020.
- [17] M. S. Al-Sultany and A. H. Ali, "Machine learning approach for classifying country income levels based on socioeconomic indicators," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 234–241, 2023.
- [18] UNICEF, WHO, and World Bank Group, *Joint Child Malnutrition Estimates: Regional and Global Trends*, New York: UNICEF, 2024.
- [19] K. M. Sullivan and R. Sharma, "Feature importance analysis in health and economic forecasting models," *Data Mining and Knowledge Discovery*, vol. 37, no. 3, pp. 945–967, 2023.
- [20] (2002) The IEEE website. [Online]. Available: <http://www.ieee.org/>
- [21] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: <http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/>
- [22] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
- [23] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.
- [24] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
- [25] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
- [26] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.