LAPORAN HASIL PENELITIAN
BIG DATA CHALLENGE
"ANALISIS STATUS GIZI SISWA BERBASIS DATA DENGAN ALGORITMA EXTREME GRADIENT BOOSTING (XGBOOST)"

Disusun oleh:

Muhammad Alvaro Khikman	B2A022061
Septian Malik Putra	B2A022055
Suci Izzati	B2A023025

PROGRAM STUDI S1 STATISTIKA
FAKULTAS SAINS DAN TEKNOLOGI PERTANIAN
UNIVERSITAS MUHAMMADIYAH SEMARANG

TAHUN 2024/2025

1. PENDAHULUAN

Gizi merupakan salah satu faktor kunci yang menentukan kualitas hidup dan kesehatan seseorang. Dalam beberapa tahun terakhir, isu gizi, baik gizi kurang maupun gizi lebih, telah menjadi perhatian global. Menurut laporan Organisasi Kesehatan Dunia (*REDUCING STUNTING IN CHILDREN Equity considerations for achieving the Global Nutrition Targets 2025*, no date), malnutrisi pada anak dan remaja dapat berdampak serius terhadap perkembangan fisik, kognitif, serta produktivitas di masa depan. Seorang anak dengan gizi yang kurang akan mudah mengantuk dan kurang semangat sehingga dapat mempengaruhi proses belajar dan cara berfikir anak serta dapat menyebabkan penurunan daya ingat dan daya tahan anak(*CHILDHOOD STUNTING: Challenges and opportunities*, no date). Di Indonesia, masalah gizi masih menjadi tantangan besar. Data dari Kementerian Kesehatan (2023) menunjukkan bahwa prevalensi stunting (pendek akibat kekurangan gizi kronis) mencapai 21,6%, sementara kasus obesitas pada anak juga mengalami peningkatan, terutama di daerah perkotaan.

Perubahan pola makan, kurangnya aktivitas fisik, serta gaya hidup yang tidak sehat turut berkontribusi pada masalah gizi ganda (double burden of malnutrition). UNICEF (2022) menyatakan bahwa ketidakseimbangan gizi tidak hanya memengaruhi kesehatan individu, tetapi juga menghambat pertumbuhan ekonomi dan kualitas sumber daya manusia suatu negara. Selain itu, laporan Global Nutrition Report (2023) menegaskan bahwa intervensi gizi yang tepat, seperti fortifikasi pangan dan edukasi gizi, dapat secara signifikan mengurangi prevalensi malnutrisi. Anak usia sekolah berada dalam fase tumbuh dan berkembang sehingga cukup berisiko terkait masalah gizi yang akan berdampak pada menurunnya kemampuan kognitif dan terjadinya gangguan integrasi sensorik dan gangguan atensi. Oleh karna itu pada penelitian ini penulis berfokus pada pengklasifikasian status gizi siswa dengan menerapkan algoritma XGBoost. Adapun harapannya agar dapat memberikan kontribusi yang signifikan dalam bidang data mining agar dapat menyelesaikan permasalahan dengan mengembangkan penelitian ini khususnya dalam pengatasan tantangan klaasifikasi yang belum optimal.

2. DASAR TEORI

2.1 Penelitian Terdahulu

Pada penelitian ini menggunakan beberapa referensi penelitian terdahulu sebagai rujukan. Penelitian tersebut ditunjukkan pada **Tabel 1** berikut ini.

Tabel 1 Penelitian Terdahulu

No	Judul Penelitian	Hasil Penelitian
1	Analisis Klasifikasi Spam Email Menggunakan Metode Extreme Gradient Boosting (XGBoost) (Gladyza *et al.,* 2025)	Hasil evaluasi Klasifikasi spam email menggunakan metode Extreme Gradient Boostingmenunjukkan bahwa model yang diusulkan memiliki akurasi sebesar 95,3%, precision 95,1%, recall 95,6%, dan F1-score 95,2%. Analisis confusion matrix mengungkapkan bahwa model berhasil mengklasifikasikan 326 email spam dengan benar (True Positive) dan 323 email non-spam dengan benar (True Negative), sementara tingkat kesalahan yang tercatat relatif kecil, yaitu 17 email non-spam salah diklasifikasikan sebagai spam (False Positive) dan 15 email spam salah diklasifikasikan sebagai non-spam (False Negative).
2	Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit (Yulianti, Soesanto and Sukmawaty, 2022)	Berdasarkan hasil penelitian pada pembahasan didaptkan kesimpulan bahwa Hasil klasifikasi menggunakan metode XGBoost dengan parameter yang default pada dataset nasabah pengguna kartu kredit menghasilkan model yang dikatakan cukup baik yaitu akurasi model sebesar 80,02%, untuk presisi sebesar 85,32%, recall sebesar 94,86% dan dapat dikategorikan sebagai good classification. Untuk percobaan kedua menggunakan teknik optimasi yaitu proses hyperparameter tuning menggunakan 7 hyperparameter dengan memvalidasi data, maka didapatkan hasil hyperparameter tuning

		yang diperoleh akurasi model sebesar 83,42%, presisi sebesar 85,36%, recall sebesar 95,28% dan hasil klasifikasi termasuk kategori good classification.
3	Classification of Drinking Water Source Suitability in West Java Using XGBoost and Cluster Analysis Based on SHAP Values (Sari *et al.*, 2024)	Analisis yang telah dilakukan memperoleh model XGBoost dengan kombinasi hyperparameter, yaitu N estimator= 394, max depth= 5, learning rate= 0.0174, subsample= 0.7075, dan colsample bytree= 0.8855 serta hanya menggunakan 12 peubah merupakan model paling baikdalam melakukan pengklasifikasian sumber air minum layak di Jawa Baratdengannilai akurasi dan F1-scoresebesar77.43% dan 80.17%sehingga dapat dilakukan klasterisasi berdasarkan nilai SHAP. Nilai SHAP terbagi atas 4 klaster dengan kontribusi yang berbeda terhadap klasifikasi kelayakan sumber air minum.Klaster 2 dan 3 merupakan klaster mayoritas sertaklaster 1 dan 4 merupakan klaster minioritas.

Berdasarkan tinjauan dari beberapa penelitian terdahulu dapat disimpulkan bahwa metode algoritma XGBoost ini dikembangkan untuk mengatasi masalah dalam prediksi dan klasifikasi data yang kompleks. XGBoost menggunakan teknik ensemble learning, yaitu menggabungkan beberapa model machine learning untuk meningkatkan akurasi prediksi. Keunggulan XGBoost adalah kemampuannya dalam mengatasi overfitting dan memproses data yang sangat besar dengan cepat.

2.2 Extreme Gradient Boosting (XGBoost)

XGBoost adalah algoritma machine learning digunakan untuk memprediksi nilai kategori ataupun numerik dalam sebuah data. XGBoost ialah kategori ensemble learning yakni menggabungkan lebih dari satu model machine learining agar dapat meningkatkan hasil akurasi prediksi nilai kategori atau numerik dalam suatu data (Mustapha, Hasan and Olatunji, no date). Algoritma Extreme Gradient Boosting (XGBoost) menggunakan teknik gradient boosting agar dapat meningkatkan akurasi

prediksi atau peramalan dalam mengatasi masalah overfitting (Islam, Sholahuddin and Abdullah, 2021). Algoritma Extreme Gradient Boosting (XGBoost) bekerja dengan cara menggabungkan beberapa model machine learning sederhana dan menjadikannya satu model yang lebih kompleks dan akurat. Oleh karna hal tersebut Algoritma Extreme Gradient Boosting (XGBoost) sangat cocok diterapkan dalam penelitian ini.

2.3 Contoh Kasus yang Menerapkan XGBoost

Contoh penggunaan XGBoost adalah pada kasus prediksi jenis cuaca berdasarkan data sensor. Dalam kasus ini, XGBoost berfungsi untuk memprediksi jenis cuaca berdasarkan beberapa faktor berdasarkan data sensor suhu seperti cerah, berawan, hujan, badai dan sebagainya. XHBoost yang dilengkapi dengan fitur-fitur seperti regulasi, parallel processing, dan handliing missing values dapat memberikan hasil prediksi cuaca lebih akurat dan dapat membantu dalam pengambilan keputusan terkait prediksi cuaca kedepan.

2.4 Rumus atau Formula XGBoost

Algoritma XGBoost ini digunakan untuk memprediksi nilai target berdasarkan fitur-fitur yang diberikan. XGBoost menggunakan teknik ensemble learning, yaitu menggabungkan beberapa model pembelajaran mesin untuk meningkatkan akurasi prediksi.

1. Model Prediksi

   XGBoost membangun model sebagai jumlah dari K pohon keputusan (tree ensemble model):

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \qquad \text{.....(1)}$$

   Dimana:

- $\hat{y}i$: adalah prediksi untuk sample i,
- fk: adalah fungsi pohon keputusan pada iterasi ke- k,
- F: ruang semua pohon keputusan

2. Fungsi Loss

$$L = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \qquad \text{.....(2)}$$

   Dimana:

- l(yi ,$\hat{yi}$): fungsi loss (misalnya squared error untuk regresi atau log loss untuk klasifikasi)
- Ω(fk): **regularisasi kompleksitas model** untuk menghindari verfitting.

  Regularisasi Ω(fk) dihitung sebagai:

$$\Omega(f) = \gamma T + 12\lambda \sum_j w_{2j} \quad \text{.....(3)}$$

  Dimana:

- T T : Jumlah daun dalam pohon keputusan
- wj wj : nilai skor pada daun j
- γ γ dan λ λ : parameter regularisasi untuk mengontrol kompleksitas model

  Jadi Rumus atau formula yang digunakan dalam XGBoost adalah:

Obj = L + Ω

  Dimana:

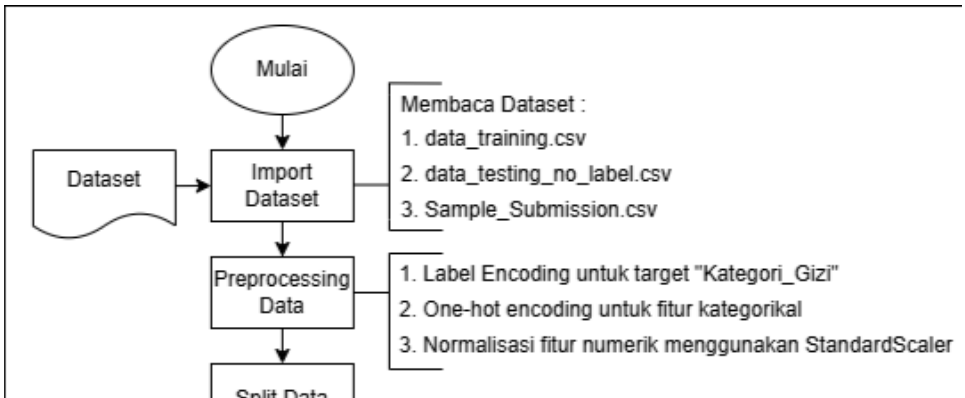
- **Obj**: fungsi objektif yang ingin dioptimalkan
- **L**: fungsi kerugian (loss function) yang mengukur kesalahan prediksi
- **Ω**: fungsi regularisasi yang mengukur kompleksitas model

Untuk mengoptimalkan fungsi objektif, XGBoost menggunakan teknik gradient boosting. Teknik ini menggabungkan beberapa model pembelajaran mesin yang lemah (weak learner) menjadi satu model yang kuat (strong learner). Setiap model lemah dihasilkan dengan meminimalkan fungsi kerugian pada residual (selisih antara nilai target dan prediksi) dari model sebelumnya.

3. METODOLOGI PENELITIAN

3.1 Diagram Alur Implementasi

Penelitian ini menggunakan bahasa pemrograman Python dengan menerapkan library TensorFlow & Keras , Pandas, NumPy, dan Scikit-learn serta XGBoost sebagai model pembelajaran mesin. Gambar 1 di bawah ini merupakan diagram alur implementasi dari penelitian ini, Diagram Alur Implementasi sebagai berikut:



\

Gambar 1 Diagram Alur Implementasi

Proses dimulai dengan mengimpor dataset, yang terdiri dari data_training.csv, data_testing_no_label.csv, dan Sample_Submission.csv. Selanjutnya, dilakukan tahap preprocessing data, yang mencakup label encoding untuk target "Kategori_Gizi", one-hot encoding untuk fitur kategorikal, serta normalisasi fitur numerik menggunakan StandardScaler. Setelah preprocessing selesai, dataset dibagi menjadi data training dan data testing. Data training selanjutnya diproses dengan pemisahan fitur (X) dan target (Y) serta pembagian antara training set sebesar 80% dan validation set sebesar 20%. Model XGBoost diinisialisasi dan dilatih menggunakan data training.

Pada tahap evaluasi model, dilakukan prediksi pada validation set serta perhitungan akurasi model. Setelah model terlatih, dilakukan prediksi pada data testing, dan hasil prediksi disimpan dalam file CSV. Proses penelitian berakhir setelah seluruh hasil prediksi tersimpan dan dapat digunakan untuk pengumpulan pada platform Kaggle.

4. HASIL PENELITIAN

Pada bagian ini, hasil penelitian mengenai analisis status gizi siswa menggunakan algoritma Extreme Gradient Boosting (XGBoost) akan dipaparkan. Proses penelitian dilakukan melalui beberapa tahap, mulai dari pengumpulan data, preprocessing, pelatihan model, hingga evaluasi performa model.

4.1 Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini berasal dari hasil pengukuran antropometri siswa, yang mencakup variabel seperti:

- ID_Siswa
- Usia
- Berat_Badan
- Tinggi_Badan
- IMT
- Asupan_Kalori
- Aktivitas_Fisik
- Kategori_Gizi
- Tingkat_Kesulitan
- Jenis_Kelamin
- Frekuensi_Makan
- Konsumsi_Sayur_Buah
- Durasi_Tidur (Menit)
- Fast_Food_Per_Minggu
- Riwayat_Penyakit

Dataset yang digunakan untuk pelatihan model terdiri dari 2.000 data siswa, dengan proporsi 80% sebagai data latih dan 20% sebagai data validasi.

4.2 Hasil Pelatihan Model

Model XGBoost dilatih dengan konfigurasi pada Tabel 2 sebagai berikut:

Tabel 2 Hasil Pelatihan Model

Parameter	Nilai
objective	multi:softmax
num_class	3
eval_metric	mlogloss
use_label_encoder	False
random_state	42

Setelah diterapkan pada dataset, model menunjukkan performa pada Tabel 3 sebagai berikut:

Tabel 3 Performa Model

Metrik Evaluasi	Nilai	Nilai di platform Kaggle
Accuracy	0.9969	1.0000

Precision	0.9969	-
Recall	0.9969	-
F1 Score	0.9969	-

Model XGBoost yang digunakan menunjukkan performa yang sangat baik dengan nilai accuracy, precision, recall, dan F1-score masing-masing sebesar 0.9969.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dalam penelitian ini, dilakukan klasifikasi multi-kelas untuk mengkategorikan status gizi siswa menggunakan algoritma XGBoost. Model yang dikembangkan menunjukkan performa yang sangat tinggi dengan nilai Accuracy, Precision, Recall, dan F1-Score masing-masing sebesar 0.9969. Hasil ini mengindikasikan bahwa model memiliki tingkat prediksi yang hampir sempurna dalam mengklasifikasikan status gizi siswa. Beberapa poin penting yang dapat disimpulkan dari hasil evaluasi ini nilai akurasi tinggi (0.9969) yang berarti model berhasil mengklasifikasikan hampir seluruh data dengan benar, menunjukkan kemampuan generalisasi yang sangat baik. Nilai presisi yang tinggi (0.9969) menunjukkan bahwa model jarang memberikan prediksi positif yang salah, yang berarti jumlah false positive sangat minim. Nilai Recall yang tinggi (0.9969) menandakan bahwa model mampu menangkap hampir seluruh instance positif dengan sangat baik, mengurangi kemungkinan false negative. Nilai F1-Score yang seimbang dengan presisi dan recall menegaskan bahwa model tidak hanya akurat, tetapi juga memiliki keseimbangan yang optimal dalam menangani kesalahan klasifikasi.

5.2 Saran

Model XGBoost yang dikembangkan dalam penelitian ini menunjukkan kinerja yang sangat tinggi dalam klasifikasi multi-kelas status gizi siswa. Dengan nilai evaluasi yang hampir sempurna di semua metrik, model ini memiliki potensi besar untuk diterapkan dalam sistem kesehatan dan pendidikan guna meningkatkan pemantauan gizi siswa secara lebih efisien dan akurat. Namun, meskipun hasil evaluasi menunjukkan performa yang hampir sempurna, perlu dilakukan analisis lebih lanjut untuk memastikan tidak adanya overfitting. Validasi menggunakan dataset independen atau metode seperti cross-validation dapat membantu mengonfirmasi bahwa model tetap andal dalam berbagai kondisi data yang berbeda.

6. DAFTAR PUSTAKA

CHILDHOOD STUNTING: Challenges and opportunities (no date).

Gladyza, L. *et al.* (2025) *Analisis Klasifikasi Spam Email Menggunakan Metode Extreme Gradient Boosting (XGBoost)*. Available at: http://j-ptiik.ub.ac.id.

Islam, S.F.N., Sholahuddin, A. and Abdullah, A.S. (2021) 'Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah', in *Journal of Physics: Conference Series*. IOP Publishing Ltd. Available at: https://doi.org/10.1088/1742-6596/1722/1/012016.

Mustapha, I.B., Hasan, S. and Olatunji, S.O. (no date) 'Effective Email Spam Detection System using Extreme Gradient Boosting'. Available at: https://doi.org/10.48550/arXiv.2012.14430.

REDUCING STUNTING IN CHILDREN Equity considerations for achieving the Global Nutrition Targets 2025 (no date).

Sari, A.P. *et al.* (2024) 'Classification of Drinking Water Source Suitability in West Java Using XGBoost and Cluster Analysis Based on SHAP Values', *Indonesian Journal of Statistics and Its Applications*, 8(2), pp. 202–214. Available at: https://doi.org/10.29244/ijsa.v8i2p202-214.

Yulianti, E.H., Soesanto, O. and Sukmawaty, Y. (2022) 'Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit', *JOMTA Journal of Mathematics: Theory and Applications*, 4(1).