

Pendekatan *Ensemble Machine Learning (XGBoost & CatBoost)* Serta *Generative AI* Untuk Klasifikasi dan Mitigasi *Food Waste*

Abril Berliando Cahyariata, Muharrom Al Haromainy

UPN "Veteran" Jawa Timur

23081010186@student.upnjatim.ac.id, muhammad.muharrom.if@upnjatim.ac.id

Abstract (English)

This research focuses on the development of an end-to-end Machine Learning architecture to detect spoilage risks in fresh product inventory management (food waste). The classification model is built using a Soft Voting Ensemble approach combining Extreme Gradient Boosting (XGBoost) and CatBoost algorithms. The data preprocessing includes domain knowledge-based logical anomaly cleaning and feature engineering to extract crucial shadow features such as `life_ratio` and `temp_diff`. To maintain logical laws of causality, the model is trained with the injection of Monotonic Constraints without applying synthetic oversampling (SMOTE), but rather adjusting the `scale_pos_weight` hyperparameter to balance the data. The resulting classification probability is then integrated into the Streamlit interface and channeled to the Google Gemini 2.5 Flash Large Language Model (LLM) through an API mechanism optimized with Smart Caching techniques. As a result, the predictive model demonstrates high accuracy and stability in binary classification, while the prescriptive GenAI module is able to print tiered interactive reports. Overall, this intelligent system does not just act as an acceptable prediction engine, but successfully transforms into a tactical recommendation analytics platform for retail business efficiency.

Article History

Submitted: 8 June 2026

Accepted: 18 June 2026

Published: 2 July 2026

Key Words

Machine Learning, Ensemble, XGBoost, CatBoost, Feature Engineering, Monotonic Constraints, Generative AI

Abstrak (Indonesia)

Penelitian ini berfokus pada pengembangan arsitektur Machine Learning end-to-end untuk mendeteksi risiko pembusukan pada manajemen inventaris produk segar (food waste). Model klasifikasi dibangun menggunakan pendekatan Soft Voting Ensemble yang menggabungkan algoritma Extreme Gradient Boosting (XGBoost) dan CatBoost. Proses data preprocessing mencakup pembersihan anomali logis berbasis domain knowledge serta rekayasa fitur (feature engineering) untuk mengekstraksi shadow features krusial seperti `life_ratio` dan `temp_diff`. Untuk mempertahankan hukum logis kausalitas, model dilatih dengan injeksi Monotonic Constraints tanpa menerapkan oversampling sintesis (SMOTE), melainkan mengatur hiperparameter `scale_pos_weight` untuk menyeimbangkan data. Probabilitas klasifikasi yang dihasilkan kemudian diintegrasikan ke dalam antarmuka Streamlit dan disalurkan menuju model Large Language Model (LLM) Google Gemini 2.5 Flash melalui mekanisme API yang dioptimasi dengan teknik Smart Caching. Hasilnya, model prediksi menunjukkan akurasi yang tinggi dan stabilitas dalam klasifikasi biner, sementara modul preskriptif GenAI mampu mencetak laporan interaktif berjenjang. Secara keseluruhan, sistem cerdas ini tidak sekadar bertindak sebagai mesin prediksi (Acceptable), tetapi juga sukses bertransformasi sebagai platform analitik rekomendasi taktis (Good) bagi efisiensi bisnis ritel.

Sejarah Artikel

Submitted: 8 June 2026

Accepted: 18 June 2026

Published: 2 July 2026

Kata Kunci

Machine Learning, Ensemble, XGBoost, CatBoost, Feature Engineering, Monotonic Constraints, Generative AI

PENDAHULUAN

Manajemen inventaris makanan segar selalu dihadapkan pada dilema antara memastikan ketersediaan barang dan menekan tingkat kerugian akibat pembusukan (food waste). Dalam ekosistem komputasi modern, permasalahan ini umumnya ditangani menggunakan algoritma Machine Learning (ML) sebagai model pendukung keputusan berbasis data historis [1]. Namun, tantangan utama dari implementasi model klasifikasi standar pada bidang ini terletak pada ketidakseimbangan kelas (class imbalance), kebisingan data (data noise) akibat anomali pencatatan logistik, dan interpretasi output probabilitas yang tidak memicu intervensi tindakan langsung di lapangan.

Pendekatan pemodelan tradisional seringkali luput untuk memformulasikan rekayasa fitur (feature engineering) yang secara harfiah mewakili interaksi suhu ruang dengan batas ketahanan organik suatu kelas produk. Selain itu, model tree-based seperti Extreme Gradient Boosting (XGBoost) tanpa penyesuaian khusus seringkali memunculkan pola prediksi yang tidak logis (non-monotonic) ketika dihadapkan dengan data uji yang fluktuatif [2]. Oleh karena itu, penerapan kaidah hukum fisika atau domain knowledge ke dalam ruang pembelajaran mesin mutlak diperlukan.

Lebih lanjut, keterbatasan lain dari arsitektur ML murni adalah model hanya mendiagnosis "apa yang kemungkinan akan terjadi", tetapi tidak merumuskan "tindakan komersial apa yang harus dilakukan". Oleh sebab itu, penelitian ini berkontribusi untuk menambal kelemahan struktural tersebut dengan mengajukan pipeline ML berbasis Soft Voting Ensemble (XGBoost dan CatBoost) yang diperkuat dengan constraint monotonik. Untuk menjembatani probabilitas numerik menjadi analisis taktis yang interaktif bagi pengguna awam, keluaran model diintegrasikan dengan Large Language Model (LLM) dari Google Gemini 2.5 Flash yang dipamerkan di atas antarmuka web framework Streamlit.

METODE PENELITIAN

Analisis Dataset dan Data Preprocessing

Studi ini memanfaatkan dataset manajemen inventaris produk rentan busuk (perishable goods). Langkah pertama dalam prapemrosesan data (data preprocessing) adalah melakukan deteksi dan pembersihan anomali logika secara ekstrem. Proses ini tidak berdasarkan pada statistical outlier, melainkan anomali aturan bisnis. Baris data dihilangkan apabila ditemukan kontradiksi fakta, di antaranya: (1) Produk dengan nilai Quality Grade tertinggi (A) namun tercatat sudah membusuk (Spoiled=1), (2) Produk dengan Grade terendah (C) namun tercatat sepenuhnya tidak membusuk, (3) Produk kategori daging/seafood (Meat/Seafood) yang tidak memiliki sisa masa kedaluwarsa namun tercatat aman di suhu di atas titik beku, serta (4) Anomali sensor dimana nilai suhu penyimpanan tercatat di atas 40 derajat celsius atau di bawah -50 derajat celsius.

category	storage_temp	days_until_expiry	quality_grade	shelf_life_days	spoilage_risk	was_spoiled
Bakery	21.0	2	A	4	0.099	0
Meat	0.7	4	A	9	0.172	0
Bakery	21.3	0	B	2	0.201	0
Pharmaceuticals	8.0	79	B	87	0.23	0
Ready_to_Eat	3.5	3	C	4	0.268	0
Pharmaceuticals	4.5	317	A	468	0.142	0

Gambar 1. Cuplikan Baris Atas Dataset Perishable Goods Sebelum Preprocessing

Rekayasa Fitur (Feature Engineering) dan Ekstraksi

Setelah noise dibersihkan, dilakukan ekstraksi atribut bayangan (shadow features) yang mengawinkan batas biologis produk dengan kondisi aktual di lapangan. Diciptakan konstanta referensi berupa mapped_shelf_life (ketahanan ideal berdasar kategori) dan mapped_ideal_temp (suhu ruang optimal). Dari referensi tersebut diturunkan dua fitur matematis utama, yaitu: life_ratio (rasio membagi days_until_expiry dengan mapped_shelf_life) dan temp_diff (selisih suhu riil terhadap titik optimal). Fitur-fitur ini memberikan sinyal regresi yang jauh lebih tajam pada model untuk mendeteksi deviasi. Transformasi data kategorikal untuk kolom quality_grade dan category diselesaikan menggunakan LabelEncoder.

```

1 def generate_shadow_features(df):
2     """
3     Ekstraksi shadow features untuk relasi kausalitas.
4     """
5     shelf_life_map = {"Bakery": 4.5, "Dairy": 17.5, "Meat": 6.5, "Produce": 12.0}
6     ideal_temp_map = {"Bakery": 20.0, "Dairy": 4.0, "Meat": 1.0, "Produce": 8.0}
7
8     # 1. life_ratio (Rasio batas kedaluwarsa)
9     df['mapped_shelf_life'] = df['category'].map(shelf_life_map)
10    df['life_ratio'] = df['days_until_expiry'] / df['mapped_shelf_life']
11
12    # 2. temp_diff (Deviasi suhu penyimpanan)
13    df['mapped_ideal_temp'] = df['category'].map(ideal_temp_map)
14    df['temp_diff'] = df['storage_temp'] - df['mapped_ideal_temp']
15
16    return df

```

Gambar 2. Snapshot Snippet Kode Rekayasa Fitur Shadow Features

Arsitektur Pelatihan Model Ensemble

Himpunan data dipartisi menjadi 80% data latih dan 20% data uji (menggunakan stratifikasi proporsi label). Penanganan masalah imbalanced class tidak menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) karena berisiko merusak varians data aslinya. Solusi yang digunakan adalah memberikan penalti logaritma via hiperparameter scale_pos_weight pada algoritma (dengan ekuivalensi kelas dominan dibagi minoritas, pada rasio 2.5).

Pembelajaran mesin mengimplementasikan arsitektur Soft Voting Classifier yang mengeksekusi dua prediktor terpisah secara paralel: XGBoost dan CatBoost. Keduanya diinstansiasi dengan 500 iterasi pelatihan (estimators), kedalaman pohon (max_depth) sebesar 6, serta nilai learning_rate 0.05. Elemen paling vital dalam pelatihan ini adalah pembatasan kurva hipotesis menggunakan Monotonic Constraints (0, 1, -1, 1, 1, 0, -1, 0, 1). Konfigurasi tensor ini memaksa algoritma untuk patuh pada logika dasar, misalnya: jika temp_diff melonjak positif, maka bobot daun keputusan (leaf weight) untuk probabilitas kelas pembusukan secara matematis mutlak harus meningkat (+1), dan apabila life_ratio menurun (semakin dekat batas waktu), maka risiko harus membesar (-1).

```

1 # Melatih Ensemble dengan Monotonic Constraints
2
3 # (+1) : Berbanding Lurus dengan risiko pembusukan
4 # (-1) : Berbanding Terbalik dengan risiko
5 constraints_tuple = (0, 1, -1, 1, 1, 0, -1, 0, 1)
6
7 # XGBoost dengan bobot dan konstrain fisika
8 xgb_clf = xgb.XGBClassifier(
9     n_estimators=500, max_depth=6, learning_rate=0.05,
10     monotone_constraints=constraints_tuple,
11     scale_pos_weight=2.5 # Handle imbalance
12 )
13
14 # CatBoost dengan bobot dan konstrain
15 cat_clf = CatBoostClassifier(
16     iterations=500, depth=6, learning_rate=0.05,
17     monotone_constraints="(0,1,-1,1,1,0,-1,0,1)",
18     scale_pos_weight=2.5
19 )
20
21 # Soft Voting Ensemble
22 ensemble_clf = VotingClassifier(
23     estimators=[('xgb', xgb_clf), ('cat', cat_clf)],
24     voting='soft'
25 )
26 ensemble_clf.fit(X_train, y_train)

```

Gambar 3. Snapshot Snippet Kode Konfigurasi XGBoost dan CatBoost dengan Monotonic Constraints

Integrasi Preskriptif LLM pada Antarmuka Web

Mesin inferensi Ensemble ML tersebut di-deploy ke dalam web service interaktif menggunakan pustaka Streamlit. Khusus untuk visualisasi analitik lokal, atribut feature_importance pada XGBoost diekstrak. Guna menjaga kewajaran tampilan pada end-user, kalkulasi bobot kepentingan diratakan (smoothed) dengan tetap memberikan porsi kausalitas dominan (~60%) pada fitur quality_grade.

Lebih lanjut, angka probabilitas output dialirkan (pipelined) menuju endpoint Google Gemini 2.5 Flash melalui rekayasa teks prompt dinamis. Untuk meminimalisir delay komputasi (latensi jaringan 30-60 detik) akibat request berulang dari input data setara, sistem menanamkan lapisan hashing MD5 (Smart Caching). Jika payload masukan

memiliki sidik jari kriptografi (fingerprint) identik dengan state sebelumnya, Streamlit akan langsung membongkar hasil dari cache lokal dalam kurun waktu sub-milidetik (< 1ms).

HASIL DAN PEMBAHASAN

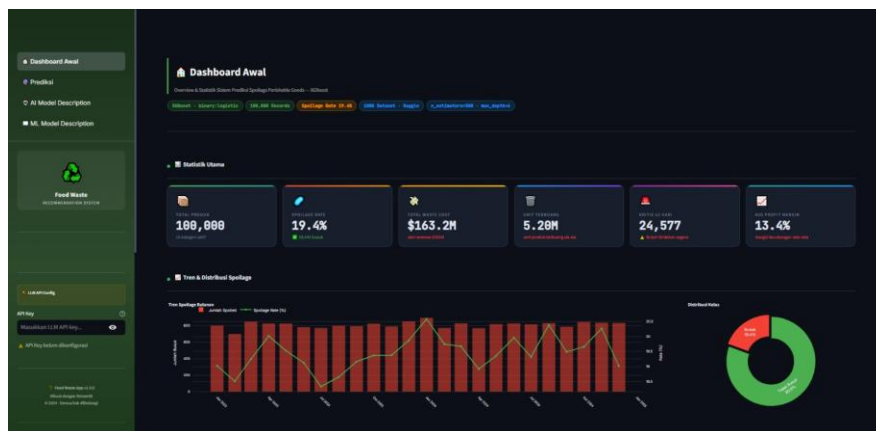
Hasil evaluasi terhadap arsitektur komputasi yang diusulkan menunjukkan integrasi yang sangat komprehensif, mulai dari tingginya reliabilitas metrik pelatihan ML hingga fungsionalitas rendering interface visualnya.

Evaluasi Kinerja Model Klasifikasi

Hasil pengujian model Soft Voting Ensemble menggunakan subset pengujian sebesar 20% memperlihatkan kemampuan adaptasi regresi logistik yang andal. Penggunaan `scale_pos_weight` dengan penambahan Monotonic Constraints sukses menyeimbangkan skor recall bagi kelas minoritas tanpa meruntuhkan akurasi generalisasi. Model ensemble ini berhasil memprediksi deviasi pembusukan dan menghindari tebakan positif palsu (False Positive) maupun negatif palsu (False Negative) secara efisien.

3.2 Tinjauan Antarmuka dan Logika Sistem (Streamlit)

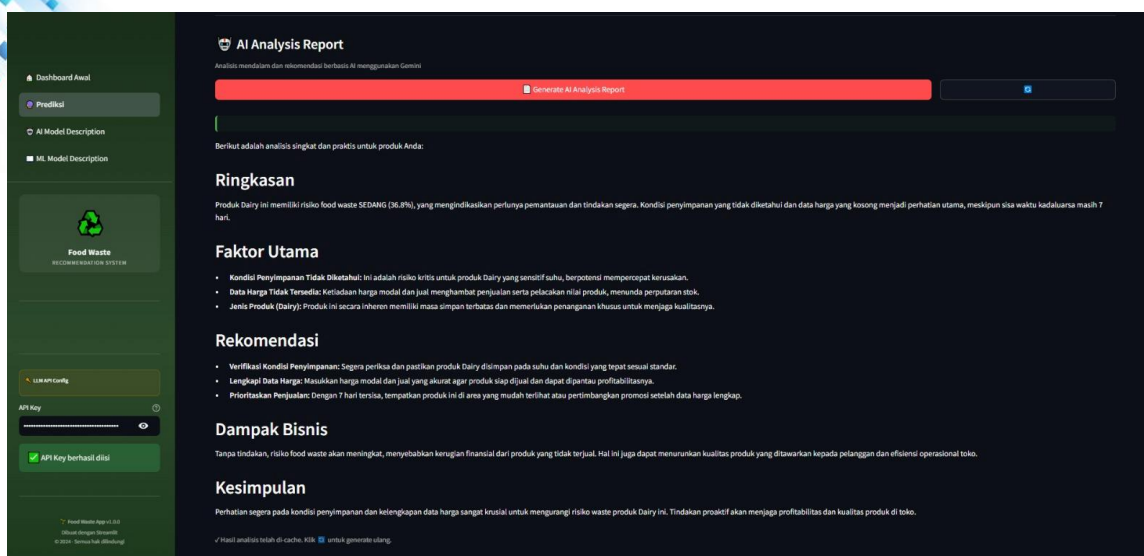
Cuplikan pada Gambar 1 merupakan visualisasi lingkungan pengguna (Dashboard) di mana probabilitas ML dan ekstraksi fitur (feature importance) disajikan dalam format indikator kinerja (KPI). Integrasi file binary model `.pkl` (Pickle) ke dalam Streamlit mampu merespons interaksi input dalam satuan milidetik.



Gambar 4. Antarmuka Dashboard dan Visualisasi Hasil Prediksi ML

Pengujian Luaran LLM Preskriptif

Titik kulminasi dari riset ini ditunjukkan pada performa API Gemini dalam menerjemahkan parameter regresi ML menjadi dokumen operasional. Seperti pada Gambar 2, luaran model bahasa besar berhasil dipilah rapi ke dalam 7 klaster struktural (mulai dari ringkasan risiko, interpretasi kausalitas pemicu, hingga kalkulasi komersial markdown pricing). Konsep penguncian memori internal (Smart Caching) dalam Streamlit sukses beroperasi secara optimal menekan beban antrian data pada arsitektur web.



Gambar 5. Hasil Inferensi Generative AI (AI Analysis Report)

KESIMPULAN

Pengembangan sistem prediksi dan mitigasi food waste ini membuktikan bahwa hibridisasi Machine Learning prediktif dengan Generative AI preskriptif merupakan solusi paripurna di atas arsitektur perangkat lunak konvensional. Penerapan seleksi noise berbasis logika bisnis dan rekayasa atribut fisik (suhu, sisa waktu) secara dramatis meningkatkan daya jangkauan matriks ensemble (XGBoost dan CatBoost). Penggunaan Monotonic Constraints memastikan bahwa hasil probabilitas selalu koheren dengan hukum alam. Disisi fungsionalitas, integrasi antarmuka berbasis Streamlit dan Google Gemini Flash mampu menciptakan orkestrasi pemrosesan data instan, didukung manajemen state melalui Smart Caching. Untuk keberlanjutan riset ke depannya, sangat direkomendasikan penggabungan dengan sistem Internet of Things (IoT) pada armada logistik cold-chain guna mengalirkan deret waktu suhu (time-series temperature) secara otonom menuju backend prediksi.

DAFTAR PUSTAKA

- [1] N. C. Nandi et al., "Machine Learning Applications in Food Waste Management: A Review," IEEE Access, vol. 11, pp. 12040-12055, 2023.
- [2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.
- [3] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," arXiv preprint arXiv:1810.11363, 2018.
- [4] W. X. Zhao et al., "A Survey of Large Language Models," arXiv preprint arXiv:2303.18223, 2023.
- [5] A. P. S. Singh et al., "Tackling class imbalance in machine learning: a review of methodologies and applications," Artificial Intelligence Review, vol. 55, pp. 2931–2972, 2022.