

Prediksi Harga Komoditas Pangan di Indonesia dengan Model *ExtraTrees* dan *K-Nearest Neighbors Regression* - Datavidia

Tristan Khayru Abiyudha^{#1}, Salman Faiz Hidayat^{#2}, Daniel Winston Mandela Tulung^{#3}

[#]Department of Computer Science and Electronics, Universitas Gadjah Mada

¹tristankhayruabiyudha@mail.ugm.ac.id

²salmanfaizhidayat@mail.ugm.ac.id

³danielwinstonmandela@mail.ugm.ac.id

Gedung C, Lantai 4, Sekip Utara, Bulaksumur, Sendowo, Sinduadi, Kec. Mlati, Kabupaten Sleman,
Daerah Istimewa Yogyakarta 55281, Indonesia

Abstrak— Fluktuasi harga komoditas pangan merupakan tantangan yang signifikan dalam sektor ekonomi dan perdagangan. Model prediksi harga yang akurat dapat membantu pemangku kepentingan, seperti pemerintah dan pelaku industri, dalam pengambilan keputusan strategis. Penelitian ini bertujuan untuk mengembangkan model machine learning berbasis Extra Trees Regressor dan K-Nearest Neighbors (KNN) guna memprediksi harga 13 jenis komoditas pangan di Indonesia. Dataset yang digunakan mencakup periode Januari 2022 hingga September 2024 untuk data latih dan Oktober hingga Desember 2024 untuk data uji.

Dalam proses pemodelan, dilakukan pra-pemrosesan data yang mencakup konversi format tanggal dan penanganan missing values menggunakan interpolasi linear. Model dievaluasi menggunakan Mean Absolute Percentage Error (MAPE), dengan hasil menunjukkan bahwa Extra Trees Regressor memiliki performa terbaik dibandingkan KNN dalam sebagian besar wilayah. Hasil prediksi menunjukkan bahwa komoditas dengan volatilitas tinggi, seperti Cabai Merah Keriting dan Cabai Rawit Merah, memiliki tingkat kesalahan prediksi yang lebih tinggi dibandingkan komoditas dengan harga lebih stabil, seperti Gula Konsumsi dan Tepung Terigu Curah.

Ke depan, penelitian ini dapat dikembangkan dengan menambahkan variabel makroekonomi, seperti inflasi dan nilai tukar, serta menggunakan model deep learning berbasis time-series, seperti LSTM, untuk menangkap pola harga dalam jangka panjang.

Kata Kunci— Machine Learning, Prediksi Harga, Extra Trees Regressor, K-Nearest Neighbors, MAPE, Time-Series

I. PENDAHULUAN

Harga bahan pangan di Indonesia seringkali mengalami volatilitas yang tinggi akibat berbagai faktor, termasuk inflasi, nilai tukar mata uang, dan kondisi pasar global. Data dari Badan Pusat Statistik menunjukkan bahwa kelompok makanan, minuman, dan tembakau mengalami kenaikan harga sebesar 6,36% pada Februari 2024 [1]. Kenaikan ini berdampak signifikan pada pengeluaran rumah tangga serta rantai pasokan industri pangan. Oleh karena itu, kemampuan untuk meramalkan harga bahan pangan menjadi sangat

penting dalam pengambilan keputusan strategis oleh pelaku industri dan pemerintah.

Kompetisi Datavidia 9.0 menantang peserta untuk membangun model prediksi harga pangan berbasis data time-series. Tantangan utama dalam peramalan ini meliputi:

1. Mengelola *missing values* dan *noise* dalam data. Data deret waktu seringkali tidak lengkap atau mengandung noise yang dapat mempengaruhi akurasi model prediksi.
2. Menentukan fitur yang paling berpengaruh dalam memprediksi harga. Identifikasi variabel atau fitur yang memiliki dampak signifikan terhadap fluktuasi harga bahan pangan sangat penting untuk meningkatkan performa model.
3. Memilih model machine learning yang mampu menangkap pola *time-series* dengan akurat. Pemilihan model yang tepat, seperti *Random Forest*, *Gradient Boosting*, atau model *ensemble*, menjadi kunci dalam menghasilkan prediksi yang akurat.

Penelitian ini bertujuan untuk:

1. Mengembangkan model *machine learning* untuk peramalan harga bahan pangan.
2. Mengevaluasi performa model menggunakan *Mean Absolute Percentage Error* (MAPE). MAPE merupakan metrik yang banyak digunakan dalam peramalan deret waktu karena kemampuannya menangani data dengan skala yang berbeda [2].
3. Mengidentifikasi faktor yang berkontribusi terhadap perubahan harga bahan pangan.

Faktor-faktor yang mempengaruhi volatilitas harga pangan di Indonesia sangat kompleks dan beragam. Menurut penelitian sebelumnya, faktor-faktor seperti harga dunia, pendapatan per kapita, nilai tukar, produksi domestik, dan kondisi iklim

memiliki pengaruh signifikan terhadap perubahan harga komoditas pangan seperti beras, kedelai, dan jagung [3]. Selain itu, volatilitas harga minyak dunia dan suku bunga riil juga ditemukan berperan dalam mempengaruhi harga pangan [4]. Pemahaman mendalam mengenai faktor-faktor ini sangat penting dalam pengembangan model prediksi yang akurat.

Dalam konteks peramalan harga pangan, pengelolaan *missing values* dan *noise* dalam data menjadi tantangan tersendiri. Data *time-series* seringkali mengalami ketidaksempurnaan yang dapat mempengaruhi akurasi model prediksi. Oleh karena itu, teknik pra-pemrosesan data yang tepat diperlukan untuk mengatasi masalah ini. Selain itu, pemilihan fitur yang relevan dan signifikan dalam memprediksi harga pangan juga menjadi aspek krusial. Fitur-fitur seperti harga komoditas dunia, nilai tukar, dan indikator ekonomi lainnya perlu dipertimbangkan dalam pengembangan model.

Pemilihan model *machine learning* yang tepat juga menjadi kunci dalam peramalan harga pangan. Model yang mampu menangkap pola *time-series* dengan baik, seperti *Random Forest*, *Gradient Boosting*, atau model *ensemble*, perlu dieksplorasi dan dievaluasi kinerjanya [5]. Evaluasi performa model menggunakan metrik seperti *Mean Absolute Percentage Error* (MAPE) akan memberikan gambaran mengenai akurasi prediksi yang dihasilkan.

Dengan memahami dan mengatasi tantangan-tantangan tersebut, diharapkan penelitian ini dapat memberikan kontribusi signifikan dalam pengembangan model prediksi harga pangan yang akurat, sehingga dapat mendukung pengambilan keputusan yang lebih baik oleh pelaku industri dan pemerintah.

II. METODE ANALISIS

A. Pengumpulan dan Pra-pemrosesan Data

A.1. Penggabungan dan Penanganan Data

Dalam rangka memudahkan proses analisis eksploratif, seluruh data latih dari masing-masing komoditas digabungkan menjadi satu dataframe yang lebih besar. Penggabungan ini memungkinkan analisis lebih menyeluruh terhadap tren harga, distribusi data, serta pola yang muncul di berbagai komoditas secara bersamaan. Dengan memiliki satu dataframe utama, analisis statistik deskriptif, visualisasi tren harga, serta pemeriksaan nilai yang hilang atau anomali dapat dilakukan dengan lebih efisien.

Namun, meskipun data digabungkan untuk analisis awal, proses pemodelan tetap dilakukan secara terpisah untuk setiap komoditas. Hal ini karena setiap komoditas memiliki karakteristik harga yang berbeda, baik dalam pola musiman, volatilitas, maupun faktor-faktor yang mempengaruhi perubahan harga. Oleh karena itu, model yang digunakan untuk setiap komoditas disesuaikan dengan karakteristik uniknya masing-masing, tanpa menggunakan data dari komoditas lain.

A.2. Deskripsi Dataset

Dataset yang digunakan dalam kompetisi ini mencakup data harga komoditas dari 1 Januari 2022 hingga 30 September 2024 untuk data latih (*training set*), sedangkan data uji (*test set*) mencakup periode hingga Desember 2024. Data yang dikumpulkan terdiri dari harga 13 jenis komoditas pangan, yaitu:

- Daging Ayam Ras
- Daging Sapi Murni
- Telur Ayam Ras
- Cabai Rawit Merah
- Gula Konsumsi
- Beras Premium
- Tepung Terigu Curah
- Beras Medium
- Minyak Goreng Curah
- Bawang Merah
- Minyak Goreng Kemasan Sederhana
- Bawang Putih Bonggol
- Cabai Merah Keriting

Dataset menunjukkan adanya *missing values* yang signifikan dalam beberapa komoditas, dengan jumlah terbanyak pada Minyak Goreng Curah (5.437 nilai hilang), Cabai Rawit Merah (2.035 nilai hilang), serta Minyak Goreng Kemasan Sederhana (1.289 nilai hilang).

A.3. Analisis Variabilitas Harga

Beberapa komoditas menunjukkan tingkat fluktuasi harga yang tinggi, yang tercermin dalam nilai standar deviasi (std) mereka:

- Komoditas dengan fluktuasi tinggi:
 - Cabai Merah Keriting (std: 16,040)
 - Cabai Rawit Merah (std: 19,227)

Fluktuasi harga pada komoditas ini kemungkinan besar dipengaruhi oleh faktor musiman, permintaan dan penawaran, serta kondisi cuaca.

- Komoditas dengan harga stabil:
 - Gula Konsumsi (std: 1,737)
 - Tepung Terigu Curah (std: 1,136)
 - Stabilitas harga pada komoditas ini kemungkinan disebabkan oleh rantai pasokan yang lebih terkontrol.
- Perbedaan harga komoditas berdasarkan nilai rata-rata (mean):
 - Paling mahal: Daging Sapi Murni (*mean* 136,376) → Produk premium dengan harga tertinggi.

- Paling murah: Tepung Terigu Curah (*mean* 10,761) dan Beras Medium (*mean* 12,399) → Produk kebutuhan pokok dengan produksi besar-besaran.
- Rentang harga ekstrem antara nilai maksimum dan minimum:
 - Cabai Merah Keriting: 127,910 (maks) vs. 16,380 (min) → Kemungkinan terjadi lonjakan harga akibat kelangkaan pasokan.
 - Minyak Goreng Kemasan Sederhana: 35,970 (maks) vs. 14,000 (min) → Berpotensi dipengaruhi oleh kebijakan harga dan fluktuasi minyak sawit global.
- Perbedaan antara nilai *median* dan *mean*:
 - Komoditas seperti Cabai Merah Keriting dan Cabai Rawit Merah memiliki *mean* lebih tinggi dari *median*, menunjukkan adanya lonjakan harga sesekali yang menyebabkan nilai rata-rata lebih besar dibanding nilai tengah distribusi harga

A.4. Penanganan Data Hilang

Dalam dataset yang digunakan, terdapat missing values pada beberapa komoditas seperti Minyak Goreng Curah, Cabai Rawit Merah, dan Minyak Goreng Kemasan Sederhana. Untuk menangani masalah ini, diterapkan dua tahap *preprocessing* utama, yaitu konversi format tanggal dan interpolasi linear untuk mengisi nilai yang hilang.

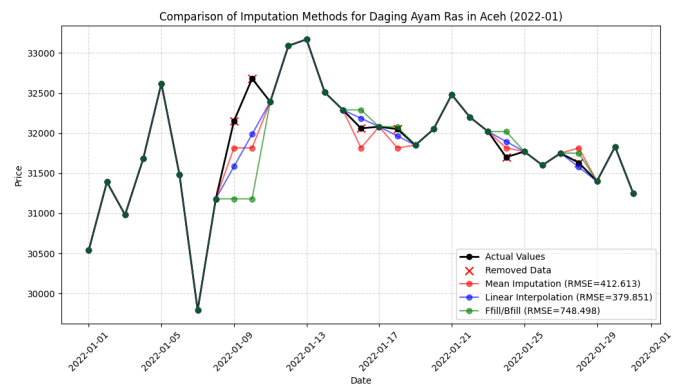
Langkah pertama dalam *preprocessing* adalah mengubah format tanggal pada dataset. Kolom 'Date', yang awalnya berbentuk *string*, dikonversi ke format *datetime* agar dapat digunakan dalam analisis *time-series* dan pemodelan machine learning. Proses ini memastikan bahwa data dapat diurutkan dengan benar berdasarkan waktu, serta memungkinkan penerapan interpolasi linear secara lebih akurat.

Setelah konversi tanggal dilakukan, beberapa metode diuji untuk menangani missing values dalam dataset:

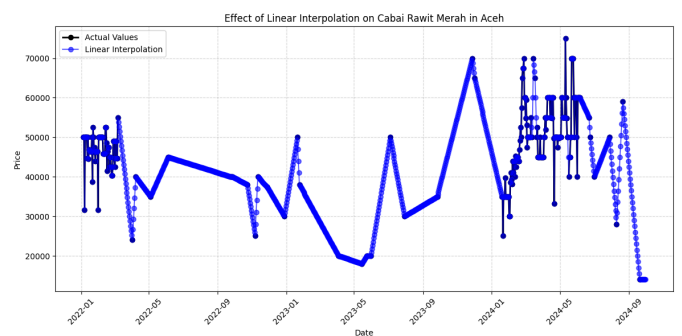
Metode Penanganan	Kelebihan	Kekurangan
Forward Fill (ffill) & Backward Fill (bfill)	Cocok untuk data dengan pola stabil	Distorsi jika tren harga tidak stabil
Mean Imputation	Mudah diterapkan	Tidak akurat untuk pola harga yang dinamis
Interpolasi Linear (dipilih)	Mampu menghasilkan estimasi yang lebih mendekati tren asli	Bergantung pada pola perubahan data

Berdasarkan evaluasi yang dilakukan, Interpolasi Linear dipilih karena lebih stabil dibandingkan metode lain serta mampu mempertahankan pola tren harga asli dengan baik [6]. Metode ini memperkirakan nilai yang hilang dengan menghubungkan titik-titik harga terdekat dalam *time-series*, memungkinkan rekonstruksi data yang lebih akurat.

Visualisasi perbandingan metode penanganan data hilang dapat dilihat pada Gambar 1, sedangkan hasil interpolasi linear yang diterapkan pada dataset diperlihatkan dalam Gambar 2.



Gambar 1. Perbandingan Metode Penanganan Data Hilang



Gambar 2. Hasil Interpolasi Linear untuk Data dengan Missing Values

B. Pemilihan dan Pelatihan Model

Dengan pendekatan naif, karena variasi data dan jumlah data yang banyak, maka diperlukan sebuah model yang mumpuni untuk memproses dan belajar dari data yang banyak. Akan tetapi, karena data memiliki format *wide*, maka akan sulit untuk memberikan label target 'harga' dan akan menimbulkan kompleksitas dalam menjaga konsistensi bentuk data. Selain itu, dataset *test* yang hanya memiliki nilai 0 sebagai *placeholder*, pembuatan fitur *time series* seperti *lag*, *rolling mean*, and *seasonality* tidaklah praktis karena fitur tidak akan memiliki nilai yang berarti untuk model dalam memprediksi nilai pada dataset *test*. Nilai-nilai dari fitur *time series* tersebut hanya akan memiliki nilai nol.

Maka dari itu, satu pendekatan awal yang dilakukan adalah dengan *training* model dan membuat prediksi untuk setiap

komoditas. Ini terbukti efektif karena menghasilkan nilai MAPE pada *public leaderboard* sekitar 0.05389-0.05385.

Kemudian, untuk mengoptimasi komputasi, pendekatan diubah dengan memecah lingkup menjadi lebih kecil. Dari pada melatih model dan memprediksi untuk setiap komoditas, model dilatih dan memprediksi untuk setiap komoditas per provinsi. Ini dilakukan agar model-model yang dilatih bisa fokus pada setiap provinsi untuk satu komoditas, dan juga mengurangi jumlah fitur yang diambil untuk dilatih. Secara keseluruhan, akan ada 442 model yang dilatih dan membuat prediksi untuk setiap komoditas per provinsi, mengingat terdapat 13 komoditas dan 34 provinsi. Dan secara komputasi, ini masih tergolong efisien, karena model *machine learning* ringan dibandingkan dengan *deep learning*.

Untuk mencari model yang terbaik, dilakukan percobaan dengan 6 model machine learning sebagai *benchmark*, dengan *parameter* berikut:

- ExtraTreesRegressor (n_estimators=100, random_state=42),
- GradientBoostingRegressor (n_estimators=100, learning_rate=0.05, random_state=42),
- KNeighborsRegressor (n_neighbors=7),
- RandomForestRegressor (n_estimators=100, random_state=42),
- XGBRegressor (n_estimators=100, learning_rate=0.05),
- CatBoostRegressor (n_estimators=100, learning_rate=0.05)

Kemudian, setiap model akan dilatih pada setiap komoditas per provinsi. Model dengan nilai MAPE terendah akan dipilih untuk memprediksi komoditas pada provinsi tersebut.

Setelah melatih 6 model pada setiap komoditas per provinsi, ditemukan bahwa model *ExtraTreesRegressor* (ET) dan KNN model adalah model yang paling banyak digunakan. Perlu diingat bahwa menilai model berdasarkan nilai MAPE tidak akan menunjukkan kapabilitas model sebenarnya, karena jumlah penggunaan yang beragam, pada komoditas dan provinsi yang berbeda pula.

ExtraTreesRegressor memiliki cara kerja yang mirip dengan *RandomForestRegressor* (RF), dimana beberapa “pohon” dibuat untuk mempelajari model pola pada data. Pohon-pohon ini dilatih secara *ensemble*, dan hasil prediksi setiap pohon akan dirata-ratakan. Perbedaannya dengan model RF adalah metode *splitting* atau pemisahan *range* nilai dari sebuah fitur pada data. RF melakukan *splitting* secara optimal, sedangkan ET melakukannya secara acak. Meskipun acak, *splitting* dengan metode tersebut dapat memberikan keberagaman prediksi dari setiap model, dan juga mengurangi resiko *overfitting*, fenomena dimana model terlalu “menghafal” data pada training, dan tidak bisa memberikan prediksi general terhadap data baru. *Splitting* acak juga mempercepat proses *training*, sehingga pendekatan dengan 442 model yang diajukan masih memungkinkan secara komputasi.

Model kedua yang digunakan adalah KNN atau *K-Nearest Neighbor*. Model ini memiliki logika yang cukup sederhana, dimana model akan mempertimbangkan *k* nilai tetangga dari

sebuah *data point* yang akan diprediksi, dan membuat prediksi berdasarkan nilai-nilai tetangga tersebut. Jumlah tetangga yang dipertimbangkan ditetapkan oleh parameter model KNN yaitu *n_neighbors*. Model KNN berguna dalam kasus *time series* seperti prediksi harga pangan, karena harga pangan seringkali mengikuti pola, trend, dan pola musiman. Sehingga, KNN dapat memprediksi nilai-nilai yang mirip dengan nilai-nilai data di sekitarnya, dan data yang diprediksi akan mengikuti pola yang ada.

Selanjutnya, meskipun CatBoost, XGBoost, dan Gradient Boosting dikenal sebagai model yang kuat dalam menangani berbagai jenis data, model-model ini memiliki beberapa keterbatasan dalam konteks prediksi harga pangan berbasis *time series*.

Pertama, *Gradient Boosting Regressor* (GBR) memiliki kinerja yang baik dalam menangkap pola non-linear, tetapi cenderung lebih lambat dibandingkan ET dalam melakukan training karena proses boosting yang bertahap. Dalam eksperimen ini, GBR tidak selalu unggul dalam hal akurasi dibandingkan ET, sehingga ET lebih sering terpilih karena memberikan keseimbangan antara akurasi dan efisiensi komputasi.

Kedua, *XGBoost Regressor* (XGBR) merupakan varian dari GBR yang lebih optimal dalam hal efisiensi komputasi. Namun, dalam data ini, XGBoost kurang optimal karena tidak dapat menangkap pola musiman dengan baik. Model berbasis pohon seperti ET lebih fleksibel dalam menangani perbedaan pola antar provinsi, sedangkan XGBoost lebih cocok untuk dataset dengan pola yang lebih seragam.

Ketiga, *CatBoost Regressor* (CBR) memiliki keunggulan dalam menangani fitur kategori tanpa perlu dilakukan *encoding* secara eksplisit. Namun, karena dataset ini hanya memiliki sedikit fitur kategori, keunggulan utama CatBoost tidak termanfaatkan sepenuhnya. Selain itu, CatBoost memiliki waktu training yang lebih panjang dibandingkan ET, membuatnya kurang efisien dalam skenario di mana 442 model perlu dilatih.

Untuk mengoptimasi model ET dan KNN, dilakukan *tuning* pada parameter *n_estimators* untuk ET dan *n_neighbors* untuk KNN. Untuk *n_estimators* sendiri, atau jumlah iterasi pelatihan, dilakukan percobaan dari 1-400. Jumlah iterasi dipilih jika menghasilkan nilai MAPE terendah. Kemudian, untuk *n_neighbors*, dilakukan percobaan nilai dari 1-31, dengan asumsi bahwa model bisa mempelajari data pada bulan yang sama untuk membuat prediksi.

Secara keseluruhan, pemilihan model dalam eksperimen ini didasarkan pada keseimbangan antara akurasi, efisiensi komputasi, dan kemampuan menangkap pola harga di berbagai provinsi. Ditemukan bahwa ET unggul dalam menangani variasi data yang kompleks dengan kecepatan yang lebih baik dibandingkan model lain, sementara, ditemukan bahwa KNN, walaupun model yang sangat simpel, sangat efektif dalam menemukan harga prediksi saat pola dari data lokal memiliki pengaruh yang kuat. Model lain yang sering digunakan seperti GBR, XGB, dan CBR memiliki potensi namun dalam konteks dataset ini mereka tidak menunjukkan performa yang lebih unggul dibandingkan ET dan KNN.

C. Metrik Evaluasi

Dalam eksperimen ini, metrik evaluasi yang digunakan untuk menilai performa model adalah *Mean Absolute Percentage Error* (MAPE). MAPE merupakan salah satu metrik yang paling umum digunakan dalam peramalan deret waktu (*time-series forecasting*) karena memberikan nilai kesalahan dalam bentuk persentase, sehingga memudahkan interpretasi hasil prediksi di berbagai skala data.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%$$

Keunggulan utama MAPE adalah kemudahannya dalam interpretasi dan fleksibilitas dalam membandingkan performa model di berbagai wilayah dan komoditas tanpa dipengaruhi oleh skala data. Namun, MAPE memiliki kelemahan, seperti sensitivitas terhadap nilai aktual yang kecil dan ketidakmampuan menangani nilai nol, yang dapat menyebabkan error dalam perhitungan. Untuk mengatasi hal ini, penelitian ini memastikan adanya teknik interpolasi dan filtering data guna menghindari pembagian dengan nol.

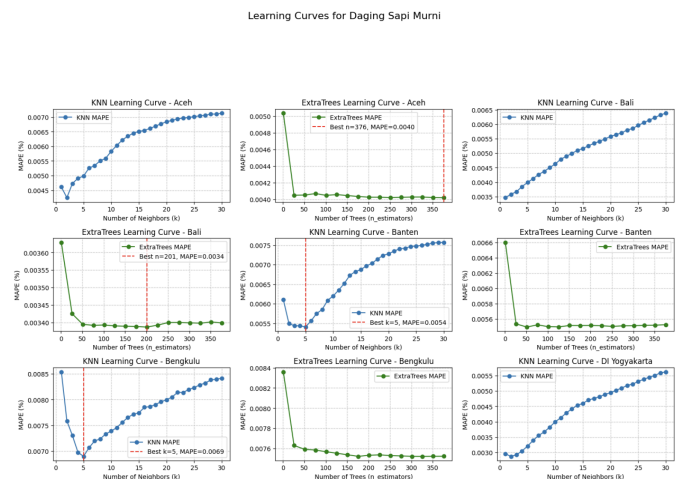
Pemilihan MAPE dalam penelitian ini didasarkan pada relevansinya dalam peramalan deret waktu serta kemampuannya memberikan gambaran yang jelas tentang akurasi model bagi pemangku kepentingan. Dengan metrik ini, performa model dapat dievaluasi secara objektif, memastikan hasil prediksi harga komoditas yang lebih akurat dan dapat diandalkan.

III. HASIL DAN PEMBAHASAN

Pada akhirnya, model ET dan KNN terbukti menjadi model terbaik untuk memprediksi harga pangan dengan data yang diberikan. Di dalam *public leaderboard Kaggle*, metode yang diajukan memiliki nilai MAPE 0.05265 sesuai dengan Figur terlampir. Jika dibandingkan dengan *benchmark* 6 model diawal pada, *benchmark* hanya menghasilkan nilai MAPE kisaran 0.053. *Learning Curve* untuk *benchmark* dicontohkan pada Gambar. 3.

Gambar 3. Contoh *benchmark* pada komoditas 'Daging Sapi Murni' di Kepulauan Riau (Lebih lengkap pada notebook: <https://www.kaggle.com/code/salmanfaizhidayat/multi-model-approach>).

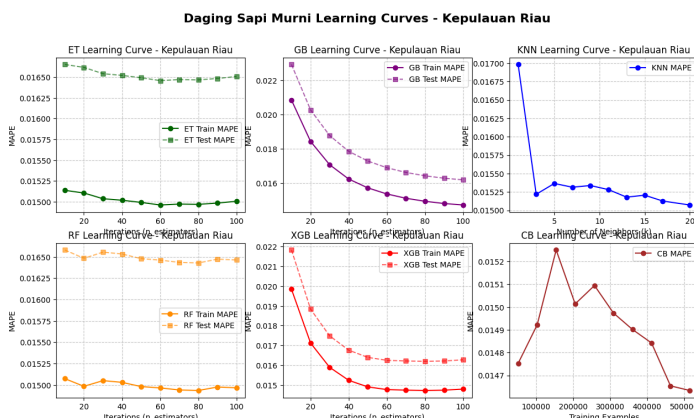
Secara distribusi, model ET lebih banyak digunakan dibanding KNN, sebanyak 252 kali dibandingkan dengan 190 kali (dengan nilai *n_neighbors* yang bervariasi). Nilai *n_neighbors* beragam, akan tetapi, tidak semua nilai dari 1-31 terpakai, hanya 1-16 kecuali 15. Kemudian untuk *n_estimators*, diuji dari minimum 1 dengan maksimum 400. Adanya fitur pencarian *n_estimators* dan *n_neighbors* juga menambah kompleksitas waktu. Garis merah pada grafik menunjukkan nilai *n_estimators* dan *n_neighbors* yang optimal pada Gambar. 4.



Gambar 4. Contoh *training* model ET dan KNN dengan *n_estimators* dan *n_neighbors* yang dinamik pada komoditas Daging Sapi Murni (Lebih lengkap pada notebook: <https://www.kaggle.com/code/salmanfaizhidayat/best-submission-haqtuahopp>).

Meskipun begitu, terdapat banyak hal yang bisa diperbaiki pada metode yang diajukan:

- *Web Scraping*: Data yang diberikan memiliki banyak nilai yang hilang, seperti data cabai rawit merah di Aceh yang memiliki nilai yang hilang 78%. Dengan mencari data di internet dari sumber lain melalui *web scraping*, data yang hilang bisa diisi dengan data eksternal tersebut.
- *Feature Engineering*: Karena masalah yang diberikan adalah *time series*, maka pembuatan fitur *time series* seperti *lag*, *rolling mean*, dan *trend* bisa memberikan manfaat. Akan tetapi, diperlukan data lebih untuk membuat fitur-fitur tersebut dengan *web scraping*.
- Sistem Pengisian Nilai Hilang yang Lebih Baik: Selain *linear interpolation*, dengan *web scraping* yang cukup, dapat dilakukan teknik *oversampling* seperti SMOTE dan ADASYN.
- Analisis Data yang Lebih Menyeluruh: Eksplorasi data yang lebih menyeluruh terhadap data *training*, komoditas global, Google Trends, dan lain-lain, dapat memberikan informasi tentang langkah yang tepat untuk memproses



data dan *modeling*, dan pada akhirnya meningkatkan akurasi.

IV. KESIMPULAN

Melalui penggunaan model *ensemble*, *Extra Trees* (ET) dan *K-Nearest Neighbours* (KNN) dengan nomor klaster atau *n_neighbors* yang bervariasi, penelitian ini berhasil memprediksi harga komoditas pangan di Indonesia. Hasil eksperimen menunjukkan bahwa ET digunakan sebanyak 252 kali dan menjadi model yang paling sering dipilih untuk berbagai wilayah dan jenis komoditas. Hal ini menunjukkan bahwa ET unggul dalam menangkap pola tren harga komoditas dibandingkan dengan model berbasis tetangga terdekat seperti KNN.

Meskipun demikian, model KNN tetap relevan, terutama untuk komoditas dan wilayah tertentu dengan pola harga yang lebih volatil. KNN_5 (47 kali), KNN_6 (31 kali), dan KNN_4 (34 kali) merupakan varian yang paling sering digunakan, mengindikasikan bahwa pemilihan jumlah tetangga optimal berpengaruh terhadap performa model dalam memprediksi harga komoditas.

Temuan ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa metode *ensemble* seperti *Extra Trees* mampu memberikan estimasi yang lebih akurat pada data deret waktu yang kompleks dan memiliki noise tinggi [5], sementara KNN dapat menjadi alternatif yang efektif ketika pola harga lebih dipengaruhi oleh hubungan lokal dalam data [6]. Oleh karena itu, kombinasi model yang fleksibel sesuai dengan karakteristik dataset sangat penting dalam membangun sistem prediksi harga komoditas yang robust.

Dengan hasil ini, diharapkan penelitian ini dapat berkontribusi dalam pengembangan sistem prediksi harga pangan yang lebih akurat di Indonesia, serta membantu pemangku kebijakan dalam mengantisipasi fluktuasi harga yang dapat mempengaruhi stabilitas ekonomi dan ketahanan pangan [7][8].

DAFTAR PUSTAKA

- [1] Badan Pusat Statistik (BPS), "Inflasi Tahun ke Tahun (y-on-y) pada Bulan Februari 2024 adalah 2.75%," 2024. [Online]. Available: <https://www.bps.go.id/id/pressrelease/2024/03/01/2301/>.
- [2] J. Makridakis, S. C. Wheelwright, and R. J. Hyndman, *Forecasting: Methods and Applications*, 3rd ed. New York, NY, USA: John Wiley & Sons, 1998.
- [3] A. R. D. L. I., "Understanding Short-Term and Long-Term Price Fluctuations of Main Staple Food Commodities in Aceh Province, Indonesia: An ARDL Investigation," *Ekonomikalia Journal of Economics*, vol. 1, no. 1, pp. 27-39, 2023.
- [4] R. Grabowski and S. Self, "Staple food price instability and structural change: Indonesian experience," *Journal of Asian Economics*, vol. 47, pp. 35-44, 2016.
- [5] M. A. Erdianto, "Perancangan Model Peramalan Jangka Pendek Harga Komoditas Pertanian di Indonesia Menggunakan Machine Learning," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 3, no. 4, pp. 338-346, Feb. 2023. [Online]. Available: <https://ejournals.com/klik>.

[6] F. Kamalov dan H. Sulieman, "Time series signal recovery methods: comparative study," *arXiv preprint arXiv:2110.12631*, 2021. [Online]. Tersedia: <https://arxiv.org/abs/2110.12631>

[7] P. Jha and M. G. Rao, "Price Stabilization Policies in Indian Agriculture," *Economic and Political Weekly*, vol. 37, no. 12, pp. 1129-1136, 2002.

[8] S. Myers, "On the Costs of Food Price Fluctuations in Low-Income Countries," *Food Policy*, vol. 31, no. 4, pp. 288-301, 2006.