Laporan Analisis Kompetisi Datavidia Arkavidia 9.0

Christama Ezra Yudianto
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
ezranewbie17@gmail.com

Jahzeel Erason Nicodemus
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
jahzeel.erason@gmail.com

Abdillah Fatah Suryanto
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
adi27noice@gmail.com

***Abstract*—Ketidakstabilan harga bahan pangan adalah tantangan besar dalam ketahanan pangan dan telah menyebabkan kesulitan pada banyak pihak pengelola produksi. Salah satu pendekatan untuk mengatasi masalah ini adalah dengan menggunakan model machine learning yang telah dikembangkan dengan dataset dan model yang tepat. Analisis ini menggunakan berbagai metode model, seperti ARIMA, Propher, XGBoost, LSTM, dan LGBMRegressor untuk mencari mana metode model yang paling efektif untuk masalah ini. Data utama yang digunakan adalah harga bahan pangan yang didapat dari Badan Pangan Nasional (Bapanans).**

Hasil analisis menunjukkan bahwa model prediksi dengan metode ARIMA mengahsilkan nilai kesalahan yang paling rendah serta visualisasi scatter plot yang memperlihatkan kedekatan antara harga aktual dan prediksi, menandakan keakuratan model dalam menghadapi ketidakstabilan harga. Berdasarkan hasil tersebut, model machine learning dapat dipakai sebagai alat bantu dalam memprediksi harga terdhadap fluktuasi harga bahan pangan di pasar.

Keywords—Machine learning, Model, Dataset, Grafik, Harga komoditas, Prediksi

I. PENDAHULUAN

Ketidakstabilan harga bahan pangan telah berdampak terhadap kestabilan ekonomi, kamandirian pangan, dan ketahanan pangan suatu negara. Ketidakstablian tersebut dipengaruhi oleh beberapa faktor seperti dinamika produksi dan permintaan, perubahan nilai tukar antar mata uang, dan kondisi ekonomi global. Hal ini menyulitkan banyak pihak dalam rantai pengelola produksi. Di tengah meningkatnya pasar pangan global, salah satu tantangannya berada pada akurasi prediksi harga komoditas pangan. Akurasi tersebut sangat diperlukan karena akan membantu dalam langkah langkah preventif pada ketidakstabilan ekonomi.

Salah satu bidang yang mendalami permasalahan akurasi prediksi harga komoditas pangan adalah data science. Data science dapat didefinisikan sebagai berikut: "Data science adalah paradigma yang interdisipliner dan meresap, di mana berbagai teori dan model digabungkan untuk mengubah data menjadi pengetahuan (dan nilai)." [1]. Data science mencari informasi dan mengambil wawasan stategis dari sebuah dataset yang kompleks. "Percobaan dan analisis terhadap dataset besar tidak hanya berfungsi untuk memvalidasi teori dan model yang sudah ada, tetapi juga untuk penemuan pola pola baru yang muncul dari data secara berbasis data, yang dapat membantu para ilmuwan dalam merancang teori dan model yang lebih baik, sehingga menghasilkan pemahaman yang lebih mendalam mengenai kompleksitas fenomena sosial, ekonomi, biologis, teknologi, budaya, dan alam"[1]. Proses tersebut merupakan langkah dalam mengembangkan model machine learning yang tidak hanya menghafal data, tetapi juga mampu memprediksi data yang baru secara akurat. "Machine learning adalah ilmu (dan seni) pemrograman komputer agar mereka dapat belajar dari data" [2].

Analsis ini merupakan penerapan data science dengan melibatkan data yang didapat dari permasalahan harga komoditas pangan. Data tersebut akan dibersihkan dan diperbaiki kualitasnya sehingga dapat ditemukan aspek musiman, tren, dan indikator lainnya tentang harga komoditas pangan dari dataset. Dengan menggunakan pemahaman terhadap dataset, pembangunan model dapat dilaksankan untuk membuat sebuah model machine learning yang dapat menghadapi ketidakstabilan harga komoditas pasar.

II. METODE ANALISIS

A. *Pengumpulan dan Pemahaman Data*

Dataset yang digunakan berasal dari dataset yang disediakan oleh komite lomba Arkavidia 9.0 pada cabang Datavidia untuk memastikan akurasi dan relevansi prediksi harga, dengan sumber data sebagai berikut:

- Harga Bahan Pangan: Disiapkan oleh Badan Pangan Nasional (Bapanas).
    - Isi data: Berisi data pergerakan harga komoditas di 34 provinsi Indonesia.
    - Folder train: Berisi data historis untuk model pelatihan.
    - Folder test: Berisi data untuk evaluasi model.
    - sample_submission.csv: Menunjukkan Format akhir pengumpulan prediksi.
- Global Commodity Price: Didapatkan dari Investing.com.
    - Isi data: Harga komoditas global seperti minyak mentah, gas alam, batu bara, minyak sawit, dan gula.
- Google Trends: Didapatkan dari laman Google Trends.
    - Isi data: Pola pencarian terkait komoditas pangan sebagai indikator sentimen pasar yang berkaitan dengan harga komoditas.
- Nilai Tukar Mata Uang: Didapatkan dari Yahoo Finance.
    - Isi data: Pergerakan nilai tukar pada dataset yang dapat memengaruhi harga impor bahan pangan.

B. *Exploratory Data Analysis (EDA) dan Business Question*

- Tujuan EDA:

- Memahami distribusi harga komoditas di berbagai provinsi.
- Mengidentifikasi tren musiman dan anomali dalam data.
- Menemukan korelasi antara harga komoditas dengan faktor eksternal seperti nilai tukar.

- Langkah-langkah EDA:
  - Memeriksa Struktur Dataset.
  - Mengecek Data yang Hilang (Missing Values).
  - Analisis Statistik Harga Komoditas (Visualisasi Data).
  - Tren Harga dalam Rentang Waktu.
  - Deteksi Outlier dan Analisis Clustering.

C. Data Processing

- Handling Missing Values:
  - Menggunakan interpolasi linier untuk mengisi harga yang hilang berdasarkan tren waktu.
  - Jika terlalu banyak data hilang, menggantinya dengan median harga per provinsi.
- Normalisasi Value:
  - Label Encoding untuk variabel kategori seperti provinsi dan komoditas.
  - MinMax Scaling agar harga lebih stabil dalam model berbasis regresi dan neural network.

D. Analisis Pemilihan Model & Eksperimen Algoritma

- Model Peramalan Deret Waktu ARIMA (1,1,1)

AutoRegressive Integrated Moving Average (ARIMA) adalah metode peramalan deret waktu yang digunakan untuk menganalisis dan memprediksi data berdasarkan nilai historisnya[3]. Model ini menggabungkan tiga komponen utama: AutoRegressive (AR), yang menggunakan hubungan linear antara nilai saat ini dan nilai sebelumnya; Integrated (I), yang mewakili proses diferensiasi untuk menghilangkan tren dan menjadikan data stasioner; serta Moving Average (MA), yang memperhitungkan kesalahan prediksi dari langkah-langkah sebelumnya untuk meningkatkan akurasi perkiraan.

- Model Machine Learning XGBoost

$$\hat{y} = \sum_{k=1}^{K} \alpha_k h_k(x)$$

Extreme Gradient Boosting (XGBoost) adalah algoritma berbasis pohon keputusan yang dioptimalkan dengan teknik boosting untuk meningkatkan akurasi prediksi[4]. Keunggulannya meliputi penggunaan regularisasi L1 dan L2 untuk mencegah overfitting, efisiensi komputasi melalui paralelisasi dan cache optimization, serta kemampuannya dalam menangani dataset besar. Dalam peramalan deret waktu, XGBoost memanfaatkan fitur berbasis lag dan tren dan dapat dikombinasikan dengan ARIMA untuk meningkatkan akurasi prediksi.

- Model Machine Learning Random Regressor

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$

Model Machine Learning Random Regressor merujuk pada penggunaan algoritma berbasis Random Forest untuk tugas regresi. Random Forest adalah ensemble learning method yang terdiri dari banyak pohon keputusan yang dilatih secara acak untuk menghindari overfitting dan meningkatkan akurasi prediksi[5]. Dalam regresi, model ini bekerja dengan menghitung rata-rata dari hasil prediksi setiap pohon keputusan dalam ensemble.

- Model Machine Learning CatBoostRegressor

$$\hat{y} = \sum_{k=1}^{K} \alpha_k h_k(x)$$

CatBoostRegressor adalah algoritma pembelajaran mesin berbasis gradient boosting yang dioptimalkan untuk menangani fitur kategorikal dengan cara yang lebih efisien dibandingkan dengan model boosting lainnya[6]. CatBoost (Categorical Boosting) secara otomatis menangani variabel kategorikal tanpa perlu pra-pemrosesan seperti encoding, yang membuatnya sangat efisien dan mudah digunakan pada data dengan banyak kategori. Algoritma ini juga dirancang untuk mengurangi overfitting dan bekerja dengan baik pada data dengan ukuran besar dan kompleksitas tinggi.

- Model Machine Learning Gradient Boosting Machines (GBM)

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \alpha_m \cdot h_m(x_i)$$

Gradient Boosting Machines (GBM) adalah algoritma ensemble yang membangun pohon keputusan secara iteratif untuk mengurangi kesalahan model sebelumnya. Pada setiap iterasi, model baru fokus pada memperbaiki kesalahan atau residual error dengan menggunakan fungsi kerugian untuk mengukur perbaikan[7]. Parameter learning rate mengontrol seberapa besar pengaruh model baru terhadap hasil akhir, sehingga model berkembang dengan lebih hati-hati untuk mencapai prediksi yang lebih akurat.

- Model Machine Learning LightGBM (Light Gradient Boosting Machine)

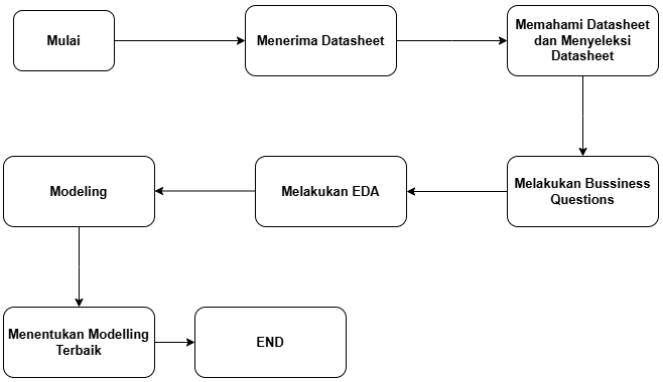
$$\text{Gain}(S) = \frac{1}{|S|}\sum_{i\in S}\hat{y}_i^2 - \left(\sum_{i\in S} y_i\right)^2$$

LightGBM (Light Gradient Boosting Machine) adalah algoritma pembelajaran mesin berbasis metode Gradient Boosting yang dikembangkan untuk meningkatkan efisiensi dan skalabilitas dalam pemrosesan data besar[8]. LightGBM menggunakan pohon keputusan berbasis leaf-wise yang lebih efisien dibandingkan dengan metode level-wise pada GBM tradisional. Hal ini memungkinkan model untuk meminimalkan loss function lebih cepat dan dengan lebih sedikit iterasi. LightGBM juga mendukung fitur-fitur seperti pemrosesan data tidak terstruktur.

E. *Justifikasi Pemilihan Model*

- Mengapa ARIMA? ARIMA dipilih karena kemampuannya dalam menangani data deret waktu yang memiliki pola musiman dan tren yang jelas.
- Mengapa XGBoost? XGBoost digunakan karena fleksibilitasnya dalam menangani hubungan nonlinier, serta kemampuannya dalam menangani data dengan jumlah besar secara efisien.
- Mengapa LSTM? LSTM dipilih karena kemampuannya dalam menangkap pola kompleks dalam data sekuensial serta mempertahankan informasi jangka panjang.
- Mengapa Hybrid ARIMA + XGBoost? Kombinasi ini memungkinkan penggabungan kekuatan ARIMA dalam memodelkan tren linier dengan XGBoost yang mampu menangkap pola nonlinier secara lebih efektif.
- Mengapa Prophet? Prophet dipilih karena kemudahan dalam interpretasi model serta kemampuannya menangani missing values dengan baik.

F. *Evaluasi Model*



Untuk mengukur performa model, digunakan metrik evaluasi Mean Absolute Percentage Error (MAPE) yang mengukur rata-rata kesalahan absolut dalam bentuk persentase. dan sesuai ketentuan penyisihan.

III. HASIL DAN PEMBAHASAN

1. Pembahasan

A. *Bussiness Question pada Dataset*

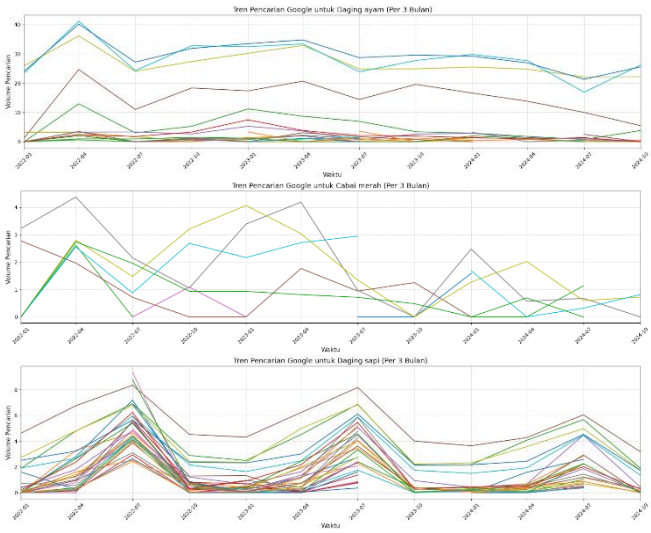
Analisis ini mengajukan beberapa pertanyaan bisnis untuk melakukan eksplorasi data awal:

- Bagaimana grafik tren yang terjadi pada dataset Google Trends?
- Apakah dataset Google Trends dapat membantu prediksi pada dataset harga bahan pangan?
- Selain dataset Google Trends, dataset apa lagi yang dapat digunakan?
- Bagaimana feature serta target pada dataset yang akan dilakukan train data?

B. *Eksplorasi Data untuk menjawab Bussiness Question dan Eliminasi beberapa Dataset*

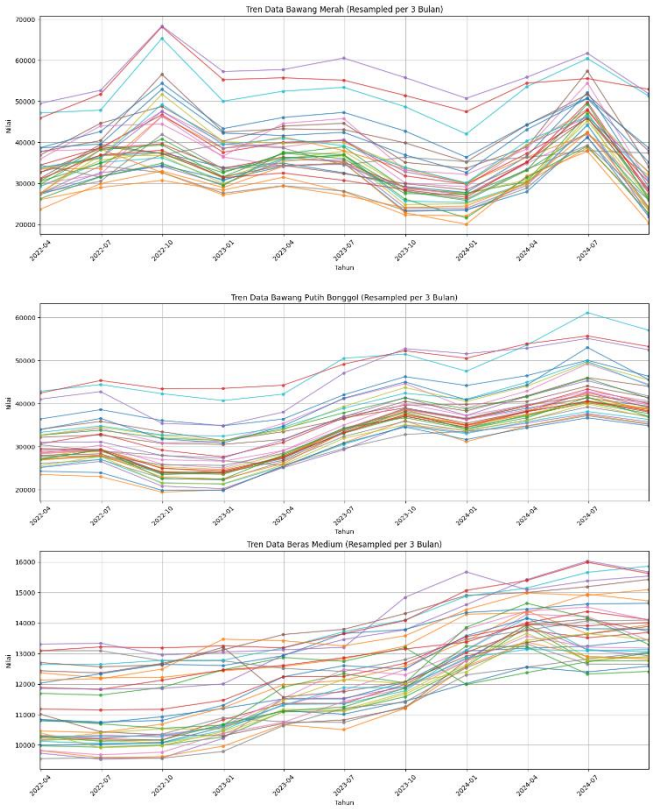
Pada tahap awal, analisis ini hanya memilih 2 dataset untuk dilakukan eksplorasi analisis data. Mengapa analisis ini hanya menggunakan 2 dataset tersebut? Saat melakukan pemahaman data, ternyata terdapat beberapa dataset yang terlihat kurang berpengaruh terhadap prediksi harga bahan pokok utama, seperti Global Commodity Price dan Mata Uang. Oleh karena itu, analisis ini hanya melakukan eksplorasi data pada 2 dataset, yaitu Google Trends dan dataset utama Harga Bahan Pokok.

- Pada dataset pertama, analisis ini melakukan visualisasi data dan volume pencarian dalam periode setiap 3 bulan, seperti berikut (beberapa grafik lainnya dapat dilihat di notebook analisis ini):

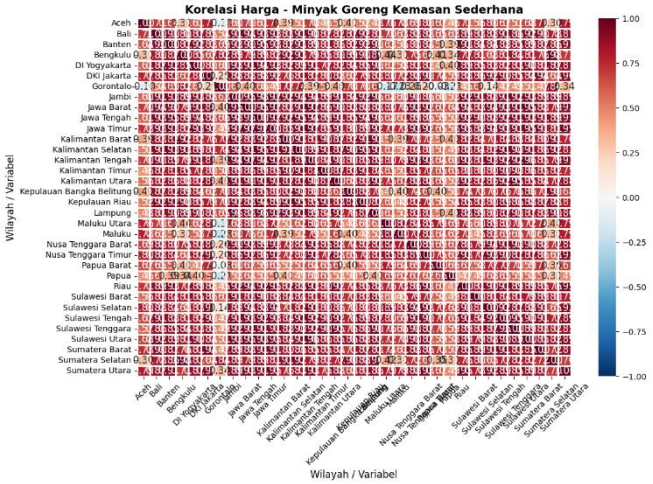


Analisis ini menemukan terdapat beberapa plotting grafik yang terputus. Hal ini disebabkan dataset Google Trends memiliki nilai kosong yang cukup besar pada beberapa wilayah provinsi.

- Dikarenakan Google Trends memiliki banyak nilai kosong, analisis ini berfokus pada dataset Harga Bahan Pokok. Seperti pada visualisasi di atas, analisis ini juga melakukan visualisasi serupa, yaitu memvisualisasikan data dan volume pencarian dalam periode setiap 3 bulan, seperti berikut:



Berdasarkan hasil visualisasi tren yang i-resample setiap 3 bulan, ditemukan beberapa fluktuasi signifikan dalam volume pencarian atau produksi. Volume tinggi terjadi di awal tahun 2022, menurun hingga pertengahan tahun, lalu meningkat kembali menjelang akhir tahun. Dengan diketahui tren nilai uang terhadap waktu, analisis mengenai salah satu korelasi harga bahan pokok antarwilayah provinsi dibuat sebagai berikut:

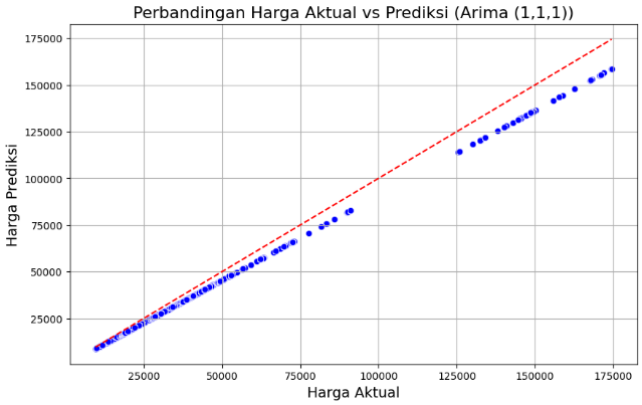


Pada grafik *heatmap* tersebut, korelasi antarwilayah cukup kuat. Selain pada grafik *heatmap* korelasi harga minyak goreng kemasan sederhana, korelasi yang cukup kuat juga banyak terjadi pada korelasi harga bahan pangan lainnya.

Maka, jawaban Business Question adalah Grafik tren pada dataset Google Trends mempunyai fluktuasi volume pencarian berdasarkan waktu. Setelah menemukan plotting grafik yang terputus serta nilai kosong di beberapa wilayah pada dataset Google Trends, analisis ini memutuskan untuk berfokus pada dataset Harga Bahan Pokok dan untuk proses train data, feature-nya adalah provinsi/wilayah, sedangkan target-nya adalah harga per waktu. Untuk dataset lainnya, analisis ini hanya memilih 2 dataset untuk dieksplorasi karena dataset lainnya, yaitu Global Commodity Price dan Mata Uang dinilai kurang berpengaruh.

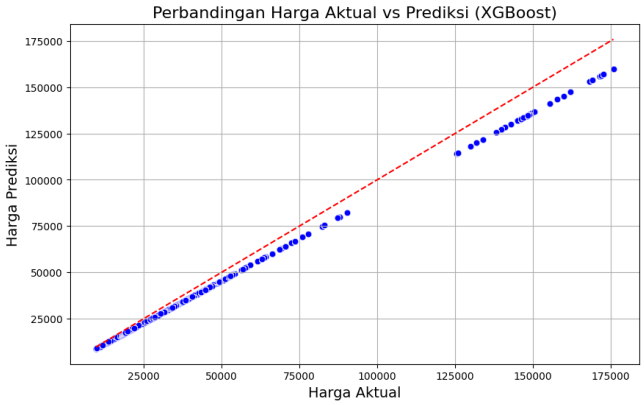
C. Modeling

- Metode Time Series Arima(1,1,1)



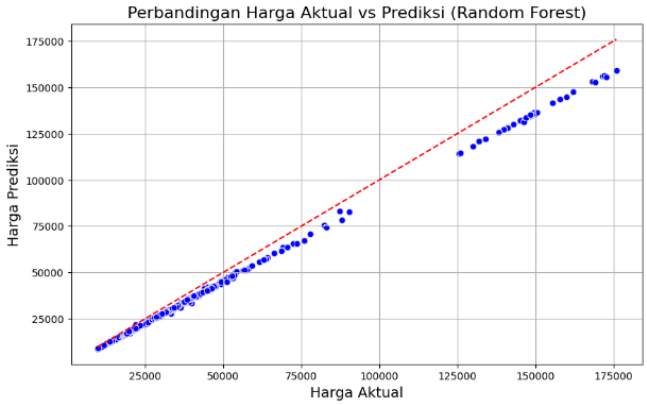
Metode Time Series ARIMA(1,1,1) menggabungkan AR(1), I(1), dan MA(1) untuk prediksi data deret waktu. Namun, perbandingan harga aktual dan prediksi menunjukkan pola berlawanan

- Metode Machine Learning XgBoost



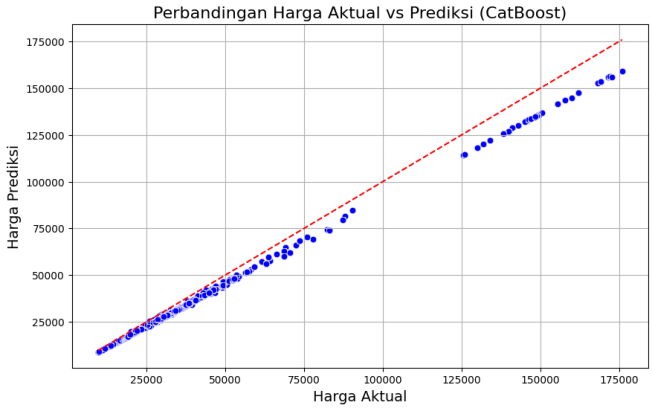
Metode Machine Learning XGBoost, yang menggunakan teknik boosting untuk meningkatkan kinerja prediksi.

- Metode Machine Learining Random Forest Regressor



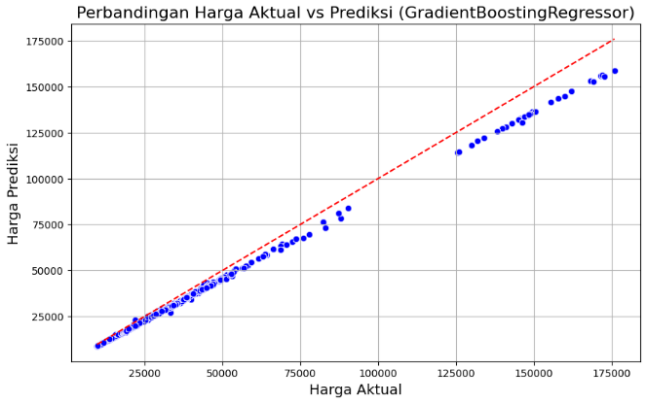
Metode Machine Learning Random Forest Regressor, yang menggunakan ensemble learning dengan membangun banyak pohon Keputusan

- Metode Machine Learning CatBoost Regressor



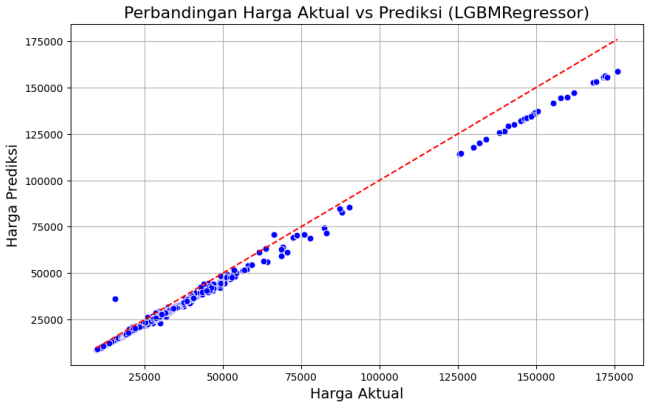
Metode Machine Learning CatBoost, yang menggunakan teknik gradient boosting dan dirancang untuk menangani data kategorikal.

- Metode Machine Learning GradientBoostingRegressor



Gradient Boosting Regressor adalah algoritma machine learning yang menggunakan teknik boosting untuk membangun model prediktif dengan menggabungkan serangkaian pohon Keputusan secara berurutan. Setiap pohon mencoba memperbaiki kesalahan dari pohon sebelumnya, sehingga meningkatkan akurasi prediksi.

- Metode Machine Learning LGBMRegressor



LGBMRegressor adalah algoritma machine learning yang menggunakan teknik gradient boosting dan dirancang untuk efisiensi serta kecepatan dalam menangani data besar. Algoritma ini bekerja dengan membangun pohon keputusan (decision trees) secara asimetris, yang membuatnya lebih cepat dan hemat memori dibandingkan metode boosting lainnya.

2. Hasil

TABLE 1. TABEL MAPE MASING-MASING MODEL

Model	**MAPE (Train)**
ARIMA (1,1,1)	0,05305
XGBoost	0,05396
Random Forest Regressor	0,05327
CatBoostRegressor	0,05620
Gradient Boosting Machines	0,05465
LightGBM	0,05881

Berdasarkan nilai MAPE (Mean Absolute Percentage Error) yang diberikan, model dengan nilai MAPE terendah adalah yang terbaik karena menunjukkan tingkat kesalahan prediksi yang lebih kecil. Model terbaik yang dipilih adalah **ARIMA (1,1,1)** dengan nilai MAPE terendah, yaitu **0,05305**. Meskipun perbedaan nilai MAPE antar model tidak terlalu signifikan, ARIMA (1,1,1) mempunyai performa yang sedikit lebih unggul dibandingkan model lainnya dalam hal akurasi prediksi.

IV. KESIMPULAN

Analisis ini menunjukkan bahwa integrasi dataset yang mempunyai deret waktu bisa menjadi bahan untuk menghasilkan sistem prediksi yang adaptif dan responsif dengan menggunakan model ARIMA, XGBoost, Random Forest Regressor, CatBoostRegressor, Gradient Boosting Machines, atau LightGBM. Secara keseluruhan, machine learning adalah Solusi yang efektif untuk memprediksi harga komoditas. Model yang dikembangkan mampu menyediakan prediksi yang mendekati nilai aslinya dan juga mempunyai fleksibilitas dan ketahanan dalam menghadapi ketidakstabilan pasar pada komoditas yang berbeda. Hal tersebut bisa menjadi alat yang berpotensi bermanfaat dalam pengambilan keputusan ekonomi di tengah dinamika pasar yang kompleks.

Analisis ini menyarankan pembuatan model lebih lanjut dengan dataset dari sumber lainnya untuk memperkuat validitas prediksi.

REFERENSI

[1] Grossi, F. Giannotti, D. Pedreschi, P. Manghi, P. Pagano, dan M. Assante, "Data science: a game changer for science and innovation," International Journal of Data Science and Analytics, vol. 11, pp. 263 278, Apr. 2021. [Online]. Tersedia: https://link.springer.com/article/10.1007/s41060-020-00240-2.

[2] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed., Sebastopol, CA, USA: O'Reilly Media, 2019.

[3] Q. D. H. B. Sitepu, S. Sutarman, and M. A. P. Siregar, "Metode Autoregressive Integrated Moving Average (ARIMA) dalam memprediksi jumlah penumpang kereta api Kota Binjai," *Int. J. Math. Appl. Sci. (IJMAS)*, vol. 18, no. 2, pp. [Online]. Tersedia: https://science.e-journal.my.id/ijma/article/download/50/53/.

[4] R. M. Syafei and D. A. Efrilianda, "Machine learning model using extreme gradient boosting (XGBoost) feature importance and light gradient boosting machine (LightGBM) to improve accurate prediction of bankruptcy," *Recursive J. of Informatics*, vol. 1, no. 2, p. 64, 2023. [Online]. Tersedia: https://journal.unnes.ac.id/sju/index.php/rji.

[5] B. Kriswantara and R. Sadikin, "Machine learning used car price prediction with random forest regressor model," *JISICOM*, vol. 6, no. 1, 2023. [Online]. Tersedia: https://doi.org/10.52362/jisicom.v6i1.752.

[6] V. Kumar, N. Kedam, K. V. Sharma, K. M. Khedher, and A. E. Alluqmani, "A comparison of machine learning models for predicting rainfall in urban metropolitan cities," *Sustainability*, vol. 15, no. 18, p. 13724, 2023. [Online]. Tersedia: https://doi.org/10.3390/su151813724.

[7] S. D. Nurrohmah, "Analisis klasifikasi menggunakan metode Gradient Boosting Machine (GBM) dan Light Gradient Boosting Machine (LGBM)," M.Sc. thesis, Universitas Gadjah Mada, 2023. [Online]. Tersedia: https://etd.repository.ugm.ac.id/home/detail_pencarian_downloadfiles/1048176.

[8] R. M. Syafei and D. A. Efrilianda, "Machine learning model using extreme gradient boosting (XGBoost) feature importance and light gradient boosting machine (LightGBM) to improve accurate prediction of bankruptcy," *Recursive J. of Informatics*, vol. 1, no. 2, pp. 12–20, 2023. [Online]. Tersedia: https://doi.org/10.15294/rji.v1i2.71229.