Implementasi Regresi Linier untuk Prediksi FoodPriceIndex: Perbandingan Model dan Pendekatan Pengisian Data

Oleh:

Ekmal Reyhan Tarihoran	1305223079
Hafizh Putra Ardhana	1305220049
Farrell Habibie Putra Haris	1305220068

FAKULTAS INFORMATIKA

UNIVERSITAS TELKOM

2024

ABSTRAK

Indeks harga pangan (FoodPriceIndex) adalah indikator penting yang mencerminkan dinamika harga pangan secara global. Pemodelan prediktif terhadap indeks ini menjadi krusial untuk membantu pemangku kebijakan dalam memahami tren harga dan merancang strategi yang relevan. Penelitian ini berfokus pada penggunaan model regresi linier sederhana untuk memprediksi FoodPriceIndex berdasarkan data temporal (tahun dan bulan) tanpa menggunakan variabel kategori seperti negara (Country). Eksperimen dilakukan dengan menangani nilai yang hilang menggunakan model regresi linier untuk meningkatkan kualitas data, yang terbukti menghasilkan korelasi lebih tinggi dibandingkan metode interpolasi atau penghapusan langsung.

Hasil penelitian menunjukkan bahwa regresi linier sederhana mampu mengungguli model tree-based dan metode kompleks lainnya dalam prediksi FoodPriceIndex, khususnya pada data unseen. Keputusan untuk menghilangkan variabel Country diambil berdasarkan hasil validasi, di mana penggunaan variabel tersebut cenderung menurunkan performa model saat di submit. Penelitian ini membuktikan bahwa pendekatan sederhana dapat memberikan hasil kompetitif dalam kompetisi data, asalkan dilakukan pemilihan fitur dan strategi validasi yang tepat.

Penelitian ini memberikan kontribusi signifikan dalam menunjukkan efektivitas model sederhana untuk masalah prediksi time series dan menyarankan arah penelitian lebih lanjut dalam penerapan model kompleks seperti deep learning.

Kata Kunci: Time Series, FoodPriceIndex, Linear Regression, Missing Values, Feature Selection

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Permasalahan fluktuasi harga pangan merupakan isu global yang memengaruhi stabilitas ekonomi, ketahanan pangan, dan kesejahteraan masyarakat. Perubahan harga yang dinamis dipengaruhi oleh berbagai faktor, termasuk tren waktu, pola musiman, serta karakteristik geografis dan ekonomi masing-masing negara. Oleh karena itu, memprediksi indeks harga pangan secara akurat menjadi tantangan penting untuk mendukung pengambilan keputusan dalam kebijakan pangan, perdagangan, dan ekonomi global.

Berbagai metode telah dikembangkan untuk memecahkan permasalahan ini, mulai dari model tradisional seperti *Autoregressive Integrated Moving Average (ARIMA)* hingga metode berbasis pembelajaran mesin seperti *Gradient Boosting* dan *Random Forest*. Selain itu, pendekatan berbasis deep learning, seperti *Long Short-Term Memory (LSTM)*, telah menunjukkan kemampuan signifikan dalam menangkap pola temporal yang kompleks pada data time series. Namun, metode yang lebih kompleks sering kali menghadirkan tantangan seperti overfitting, kemampuan interpretasi yang rendah, dan ketergantungan pada data yang besar.

Selain tantangan dalam memilih model yang optimal, penanganan data yang tidak lengkap (missing values) juga menjadi aspek penting dalam pemodelan time series. Dalam penelitian ini, kami menunjukkan bahwa menggunakan model regresi linier untuk memprediksi nilai yang hilang menghasilkan korelasi yang lebih tinggi terhadap target variabel dibandingkan metode interpolasi atau penghapusan langsung. Hasil ini memberikan dasar untuk memastikan bahwa data yang digunakan dalam pelatihan model memiliki kualitas yang lebih baik.

Dalam penelitian ini, kami menunjukkan bahwa metode regresi linier sederhana, yang hanya memanfaatkan fitur Year dan Month, mampu memberikan hasil yang lebih baik dibandingkan metode yang lebih kompleks. Meskipun fitur Country pada data sering dianggap sebagai variabel penting, hasil eksperimen kami menunjukkan bahwa menghilangkan fitur ini justru menghasilkan model yang lebih

generalis dan memberikan performa lebih baik pada data unseen dalam kompetisi Kaggle.

Kontribusi utama penelitian ini adalah:

- Demonstrasi keberhasilan model sederhana dalam menyelesaikan permasalahan kompleks, melampaui metode ensemble dan model berbasis fitur geografis.
- Analisis mendalam mengenai dampak penghapusan fitur Country terhadap performa model pada data unseen.
- Pendekatan inovatif dalam penanganan nilai yang hilang menggunakan model linear regression, yang terbukti lebih akurat dibandingkan metode tradisional seperti interpolasi.
- Penerapan pendekatan berbasis data yang berfokus pada pola temporal (Year dan Month) untuk memprediksi indeks harga pangan secara akurat.

Penelitian ini menegaskan pentingnya pendekatan sederhana dan berbasis generalisasi dalam menyelesaikan permasalahan nyata, terutama ketika berhadapan dengan data time series yang kompleks.

1.2 Tujuan

Tujuan penelitian ini adalah sebagai berikut.

1. Melakukan Analisis Eksplorasi Data (EDA): Memahami pola dan hubungan antar variabel dalam dataset, termasuk identifikasi tren temporal dari indeks harga pangan (FoodPriceIndex).
2. Menangani Data yang Hilang (Missing Values): Mengembangkan pendekatan yang optimal untuk mengisi nilai yang hilang dengan menggunakan model linear regression dan membandingkan hasilnya dengan metode interpolasi dan penghapusan nilai.
3. Mengembangkan Model Prediksi: Membangun model prediksi berbasis regresi linier untuk memprediksi FoodPriceIndex dengan memanfaatkan fitur temporal seperti Year dan Month.
4. Mengevaluasi Kinerja Model: Membandingkan kinerja regresi linier dengan metode lain, seperti model tree-based, untuk memahami efektivitas pendekatan sederhana terhadap dataset yang diberikan.

5. Mengidentifikasi Faktor Kritis: Mengevaluasi kontribusi variabel, seperti Country, terhadap prediksi FoodPriceIndex, dan menentukan alasan penghapusan variabel tersebut dalam rangka meningkatkan generalisasi model pada data unseen.
6. Mendokumentasikan dan Menyampaikan Temuan: Menyusun laporan ilmiah yang menjelaskan pendekatan, eksperimen, dan hasil, serta memberikan justifikasi atas keputusan model yang diambil untuk mendukung keberhasilan dalam kompetisi.

1.3 Manfaat

Manfaat penelitian ini adalah sebagai berikut.

1. **Manfaat untuk Pemodelan Prediktif:**

   Penelitian ini menunjukkan bahwa pendekatan sederhana seperti regresi linier dapat memberikan performa prediksi yang kompetitif dalam masalah time series tertentu, khususnya pada data dengan tren temporal yang jelas.
2. **Manfaat bagi Praktisi Data:**

   Memberikan wawasan tentang bagaimana menangani variabel kategorikal seperti Country dan bagaimana mengatasi missing values dengan metode yang lebih efektif, seperti prediksi menggunakan model linear regression dibandingkan interpolasi.
3. **Manfaat untuk Pemahaman Tren Harga Pangan:**

   Penelitian ini membantu pemangku kebijakan dan organisasi internasional dalam memahami pola indeks harga pangan secara global tanpa memerlukan model yang kompleks.
4. **Manfaat untuk Pendidikan:**

   Memberikan contoh kasus nyata tentang bagaimana teknik machine learning sederhana dapat digunakan secara efektif dalam kompetisi data dan dapat dijadikan referensi pembelajaran bagi mahasiswa atau peneliti pemula.
5. **Manfaat bagi Kompetisi Data:**

Menyediakan strategi yang terukur dan sistematis untuk mengidentifikasi fitur penting, mengevaluasi model, dan memilih pendekatan terbaik berdasarkan validasi dan performa pada data unseen.

6. **Manfaat untuk Keberlanjutan Penelitian**

   Penelitian ini membuka peluang eksplorasi lanjutan, seperti pengujian model berbasis deep learning (LSTM atau CNN) untuk data time series, dengan membandingkan efisiensi dan kompleksitas antara model sederhana dan kompleks.

1.4 Batasan Masalah

1. Batasan Dataset:

- Dataset yang digunakan hanya mencakup data FoodPriceIndex dari beberapa negara tertentu dan mungkin tidak mewakili kondisi global secara keseluruhan.
- Data yang tersedia memiliki atribut Year dan Month sebagai penanda temporal, tetapi tidak mencakup faktor eksternal lain seperti kebijakan ekonomi, cuaca, atau data geopolitik yang dapat memengaruhi harga pangan.
- Dataset mengandung nilai yang hilang (missing values) pada beberapa kolom, yang diatasi menggunakan metode prediksi berbasis regresi linier.

2. Batasan Metode:

- Penelitian ini hanya memanfaatkan metode regresi linier sederhana untuk membuat prediksi, meskipun model lain yaitu Decision tree, Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regressor, dan K-Nearest Neighbors Regressor tersedia untuk perbandingan.
- Model hanya memanfaatkan fitur temporal (Year dan Month) tanpa memasukkan atribut kategori Country, karena berdasarkan eksperimen, penghilangan atribut ini menghasilkan performa prediksi yang lebih baik pada data unseen.
- Penelitian ini berfokus pada optimasi untuk kompetisi data dengan kriteria evaluasi tertentu dan tidak mencakup analisis mendalam terhadap interpretasi model atau implikasi praktisnya di dunia nyata.

3. Batasan Eksperimen:

- Validasi dilakukan berdasarkan data yang diberikan dalam kompetisi, sehingga hasilnya mungkin tidak langsung dapat digeneralisasikan untuk dataset serupa di luar konteks ini.
- Penanganan missing values mengandalkan model linier, tanpa eksplorasi metode imputasi lain seperti time series imputation berbasis statistik atau deep learning.

BAB II
PENELITIAN TERKAIT

2.1 Implementasi Dataset Terkait

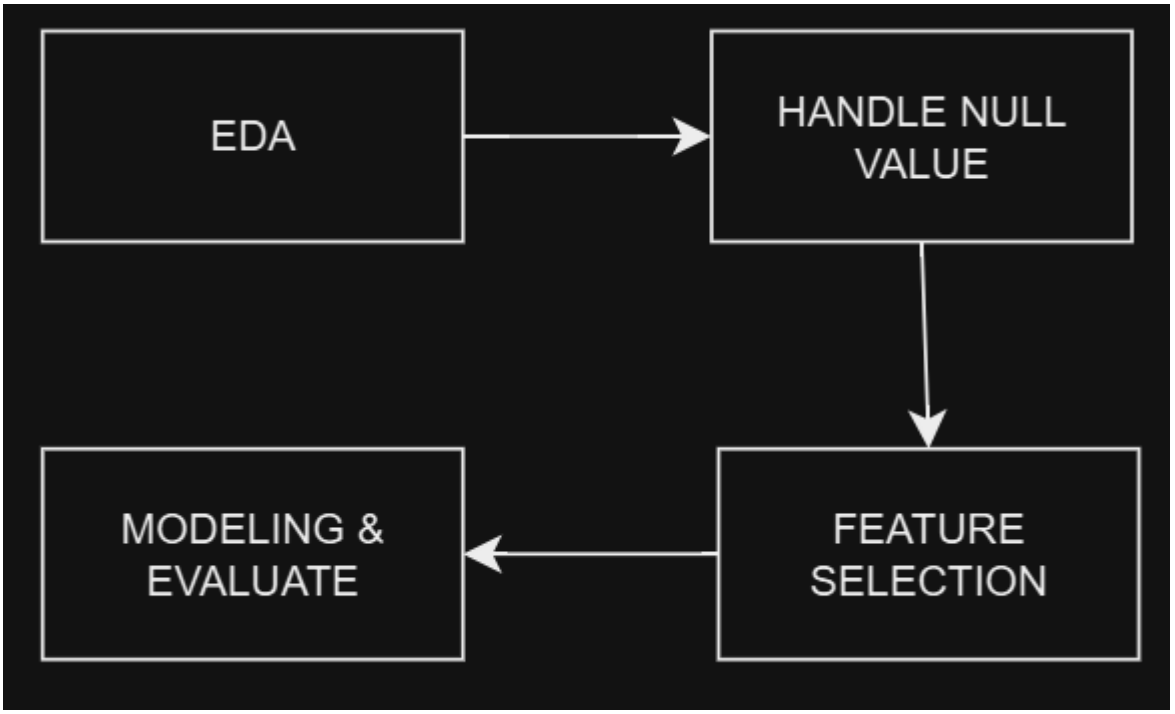
1. Fakhri Arizki et al. menggunakan tiga metode analisis time series yaitu Linear Trend Regression, Exponential Smoothing, dan ARIMA untuk membandingkan peramalan indeks harga saham Bank Central Asia periode 2020-2022 . Hasil penelitian menunjukkan bahwa metode ARIMA(1,1,0) menghasilkan nilai RMSE terkecil di angka 162.9544, sehingga memiliki tingkat akurasi paling baik di antara ketiga metode yang digunakan. Kelebihan penelitian ini adalah penggunaan beberapa metode untuk perbandingan, namun kekurangannya adalah fokus hanya pada satu jenis saham.
2. Bella Audina et al. menggunakan metode *Support Vector Machine (SVM)* untuk menentukan hasil peramalan arus kas pada data time series . Hasil penelitian menunjukkan bahwa model SVM dengan kernel *Radial Basic Function (RBF)* memiliki akurasi terbaik dengan nilai 75%, 82%, 88%, dan 64%. Kelebihan penelitian ini adalah penggunaan metode SVM yang mampu mengatasi overfitting, namun kekurangannya adalah kompleksitas dalam pemilihan parameter model.

2.2 Implementasi Metode Terkait

1. Tri Indarwati et al. menggunakan metode Linear Regression untuk memprediksi penjualan smartphone di 82 Cell Mayang. Penelitian ini bertujuan untuk menciptakan sistem informasi yang dapat melakukan peramalan penjualan smartphone dengan melihat biaya iklan dan jumlah

penjualan. Hasil penelitian menunjukkan bahwa metode Linear Regression menghasilkan nilai MAPE sebesar 0.032 dan MSE sebesar 5.16, yang dikategorikan sangat baik. Kelebihan penelitian ini adalah kemampuan metode Linear Regression dalam menganalisis beberapa variabel bebas, namun kekurangannya adalah fokus hanya pada satu jenis produk.

BAB III
METODE PENELITIAN



Gambar 1. Alur Penelitian

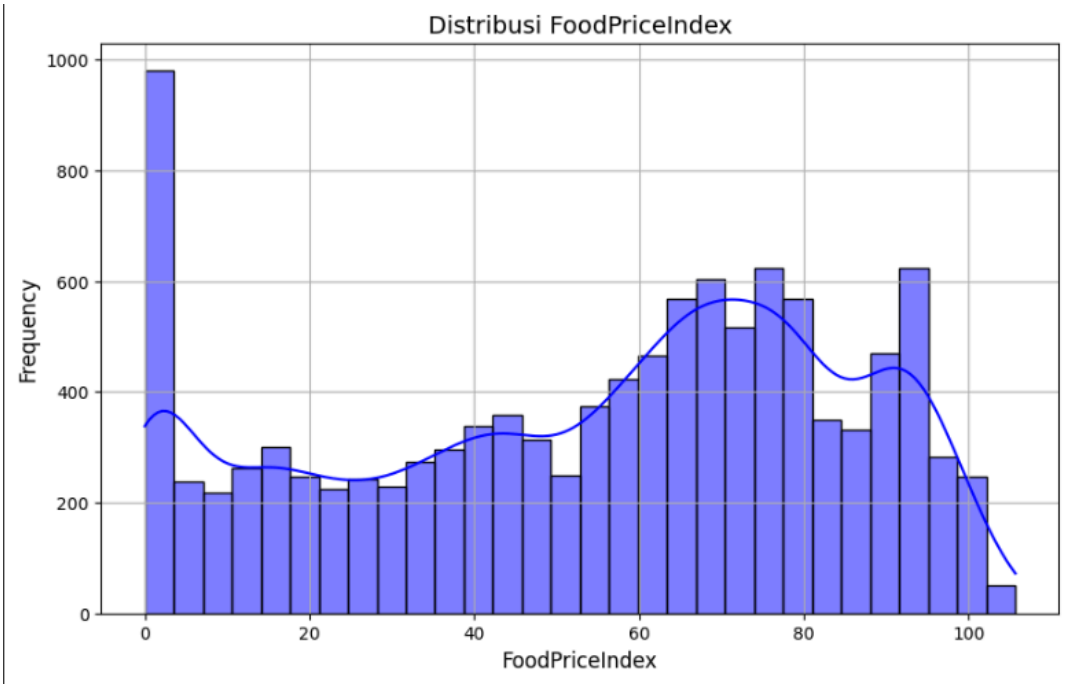
Pada bab ini, kami menjelaskan tahapan yang dilakukan dalam pengolahan dan analisis data untuk prediksi Food Price Index (FPI). Alur pengerjaan mencakup eksplorasi data (EDA), penanganan nilai yang hilang, pemilihan fitur, pemodelan, dan evaluasi model. Berikut adalah penjelasan mendetail mengenai setiap tahapan yang dilakukan:

3.1 Exploratory Data Analysis (EDA)

Eksplorasi data dilakukan untuk mendapatkan pemahaman lebih dalam mengenai distribusi dan pola dalam dataset. Dataset terdiri dari dua bagian utama: data latih (training) dengan lima kolom (id, Country, Year, Month,

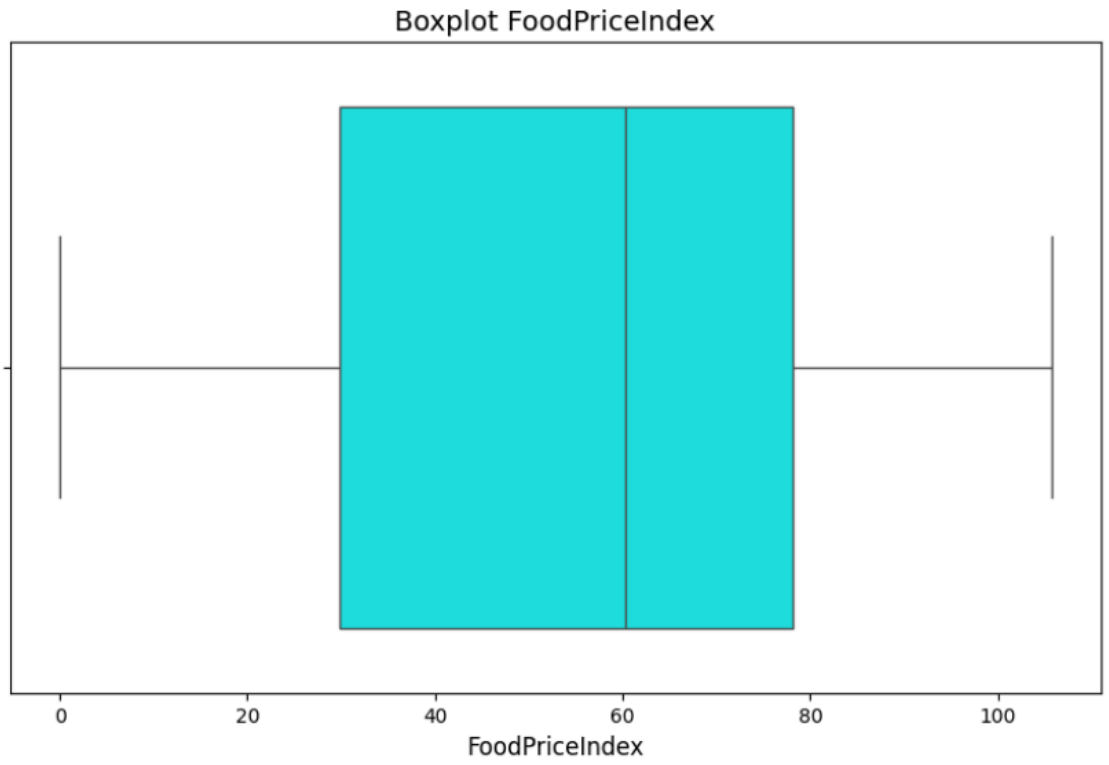
FoodPriceIndex) dan data uji (testing) dengan empat kolom (id, Country, Year, Month). Berikut adalah langkah-langkah utama dalam EDA:

1. Deskripsi Statistik: Menggunakan deskripsi statistik untuk memahami distribusi nilai dalam dataset.
    - Data latih: Terdapat 11616 entri dengan kolom FoodPriceIndex sebagai target prediksi. Kolom ini memiliki rata-rata sekitar 54.08, deviasi standar 30.19, serta rentang nilai dari 0 hingga 105.7.
    - Data uji: Terdapat 2640 entri tanpa kolom target FoodPriceIndex.
2. Distribusi Data



Gambar 2. Visualisasi Histogram FoodPriceIndex

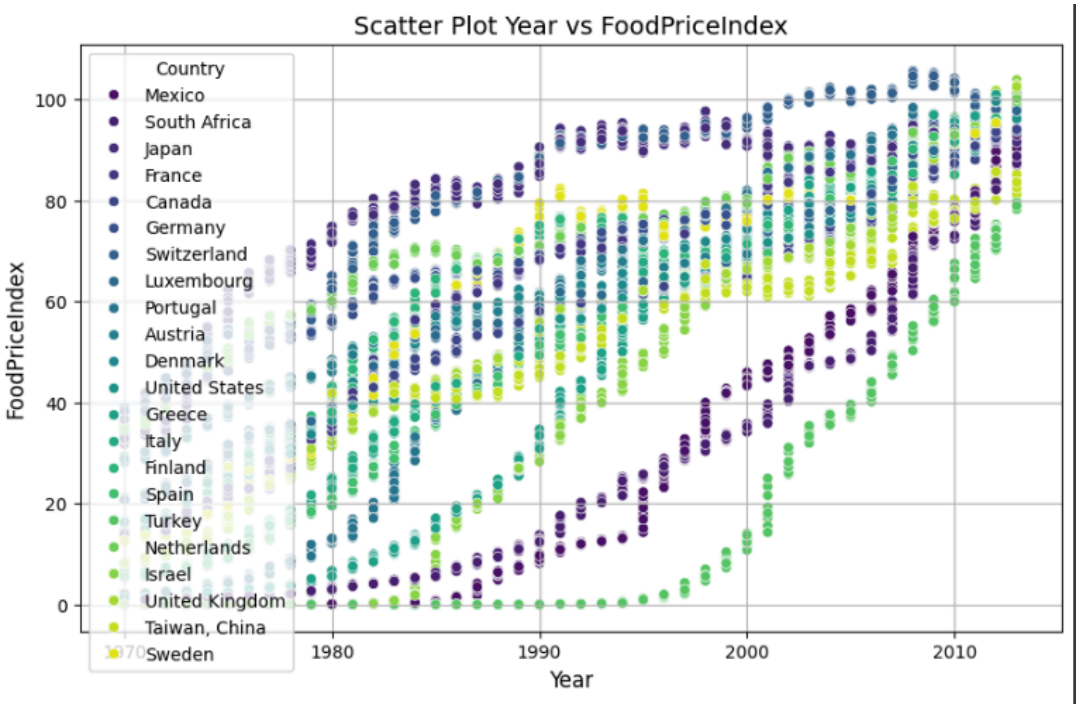
Berdasarkan Gambar 2 kolom FoodPriceIndex menunjukkan distribusi yang cukup bervariasi antar negara dan waktu, menandakan adanya faktor eksternal yang memengaruhi indeks harga



Gambar 3. Visualisasi Boxplot FoodPriceIndex

Berdasarkan Gambar 3 kolom FoodPriceIndex memiliki nilai tengah yang berada di sekitar setengah dari rentangnya, dan terlihat juga tidak ada outlier dalam kolom FoodPriceIndex, serta terlihat kotak yang cukup besar yang menandakan variasi dari kolom ini cukup besar

3. Visualisasi Tren



Gambar 4. Visualisasi Scatter plot kolom Year dan FoodPriceIndex

Visualisasi dilakukan untuk melihat pola antara Year dan FoodPriceIndex menggunakan scatter plot dengan kolom country sebagai hue. Hasil visualisasi menunjukkan bahwa harga cenderung meningkat seiring waktu, mengikuti pola linear. Visualisasi ini menunjukkan bahwa tren kenaikan harga makanan adalah fenomena global, meskipun kecepatan dan tingkat kenaikan berbeda-beda di tiap negara.

3.2 Handle Null Value

Penanganan nilai yang hilang pada kolom FoodPriceIndex sangat penting untuk memastikan integritas data. Pada dataset latih, terdapat 348 nilai yang hilang pada kolom ini. Kami menggunakan tiga pendekatan untuk menangani nilai yang hilang:

- Drop Null Value: Menghapus entri yang memiliki nilai kosong.
- Interpolasi Linear: Adalah metode untuk mengisi nilai yang hilang dengan pendekatan berbasis tren atau hubungan linear antara dua titik data yang ada. Nilai kosong di antara dua titik data akan diisi dengan nilai yang dihitung melalui interpolasi.

$$y = y_1 + \frac{(y_2 - y_1)}{(x_2 - x_1)} \cdot (x - x_1)$$

Gambar 4. Rumus Interpolasi Linear

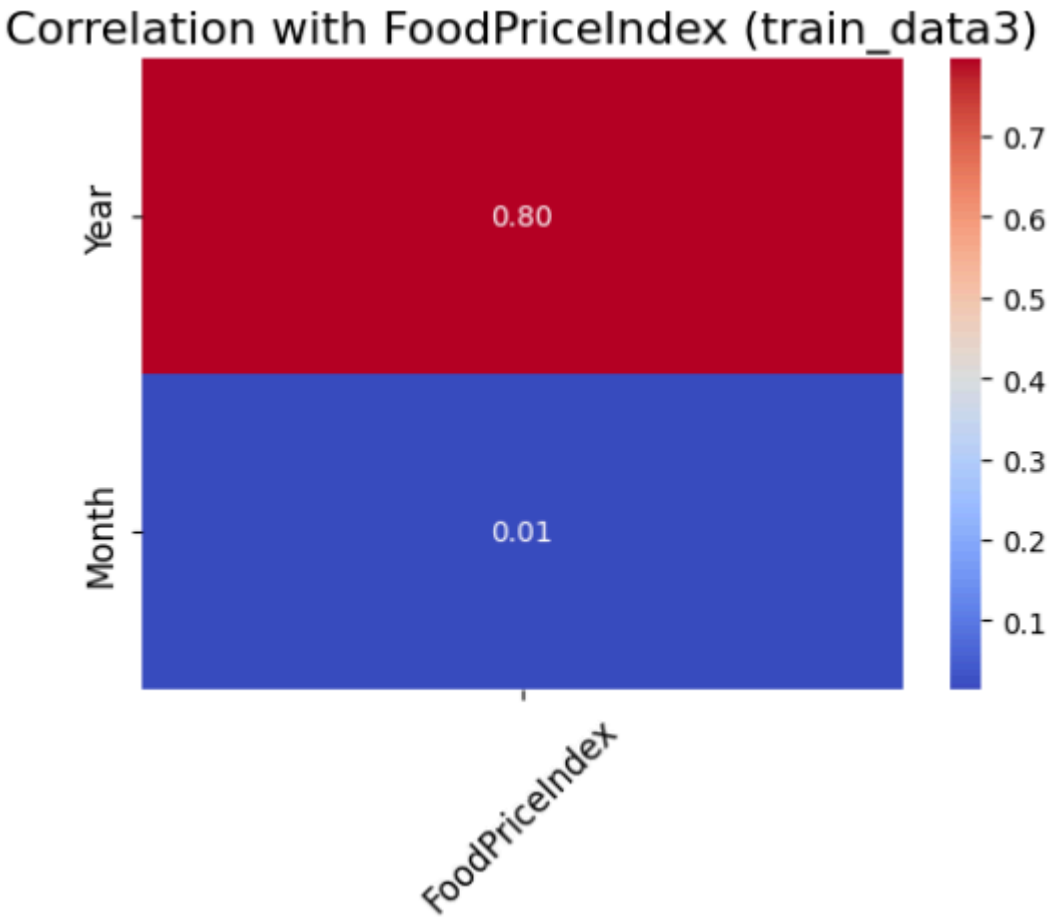
dimana :

- y1: Nilai sebelum NaN
- y2: Nilai setelah NaN
- x1: Index dari nilai sebelum NaN
- x2: Index dari nilai setelah NaN

*NaN adalah null value

- Regresi Linier: Membuat model regresi sederhana berdasarkan hubungan antara kolom Year dan FoodPriceIndex untuk memprediksi

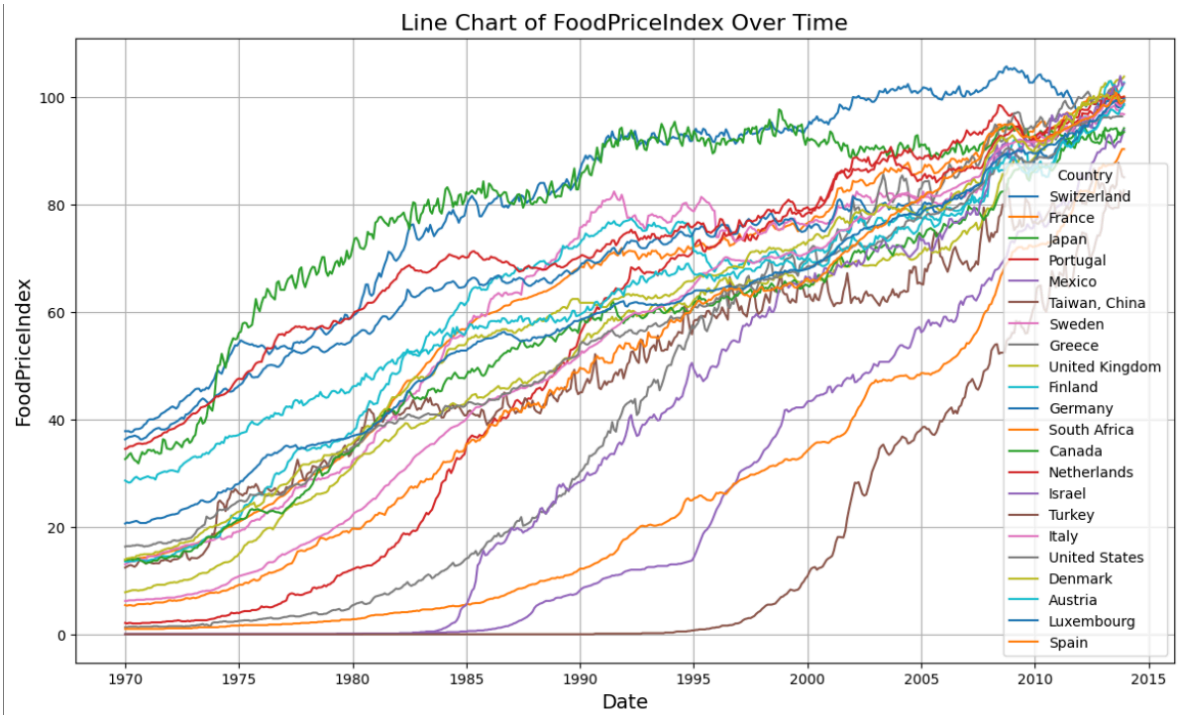
nilai yang hilang. Pemilihan kolom Year sebagai fitur dikarenakan memiliki korelasi yang cukup tinggi dengan FoodPriceIndex.



Gambar 5. Matriks Korelasi Kolom Year, Month dan FoodPriceIndex

Hasil evaluasi menunjukkan bahwa metode regresi linier memberikan hasil terbaik, dengan korelasi 0.8 antara kolom Year dan FoodPriceIndex. Oleh karena itu, metode ini dipilih untuk mengisi nilai yang hilang.

3.3 Feature Selection



Gambar 6. Line Chart Kolom FoodPriceIndex dengan waktu(tahun)

Pemilihan fitur sangatlah penting untuk memprediksi nilai dari target, oleh karena itu kami cukup lama untuk menganalisis fitur mana yang akan digunakan untuk memprediksi FoodPriceIndex, setelah dilakukan analisis, maka kami membuang kolom Country dan hanya menggunakan kolom Year dan Month saja, berikut adalah alasan dan pertimbangannya :

- Kami memutuskan untuk tidak menggunakan kolom country sebagai salah satu fitur karena dalam konteks globalisasi, pasar global telah menciptakan harga yang tidak jauh berbeda. Dengan demikian, kolom country cenderung tidak memberikan pengaruh signifikan dalam analisis.
- Potensi Data Country yang Tidak Konsisten : Data Country mungkin memiliki distribusi yang berbeda di antara data training dan testing di kompetisi. Misalnya, ada negara tertentu di data testing yang tidak cukup diwakili di data training, atau distribusi negara dalam data testing tidak seimbang.

- Keputusan untuk tidak menggunakan Country adalah berdasarkan eksperimen yang menunjukkan bahwa hasil prediksi dengan linear regression tanpa Country memberikan skor lebih baik di leaderboard

3.4 Modelling and Evaluation

Sebelum kami menerapkan model, kami melakukan standarisasi terlebih dahulu agar value dari data berada pada skala yang sama. Kami menggunakan *StandardScaler* sehingga setiap value dari kolomnya memiliki rata-rata nol dan standar deviasi 1

- Linear Regression
  Linear Regression adalah model regresi paling sederhana yang mencoba menemukan hubungan linear antara fitur dan target. Model ini memprediksi target sebagai kombinasi linear dari fitur dengan menyesuaikan garis lurus terbaik.
  Hasil pada Dataset: Model ini menunjukkan performa yang cukup baik dengan MSE = 298.52 dan R^2 = 0.65, namun cenderung underfitting karena pola hubungan dalam data mungkin tidak sepenuhnya linear.
- Decision Tree
  Decision Tree adalah model non-linear yang membagi dataset secara rekursif menjadi kelompok-kelompok kecil berdasarkan aturan keputusan. Model ini membangun pohon keputusan untuk memprediksi target.
  Hasil pada Dataset: Model ini menunjukkan tanda-tanda overfitting karena Training R^2 > Validation R^2, dengan MSE = 329.38 dan R^2 = 0.61
- Random Forest
  Random Forest adalah model ansambel yang menggabungkan beberapa Decision Tree untuk meningkatkan akurasi dan mengurangi overfitting. Setiap pohon dilatih pada subset data yang berbeda.
  Hasil pada Dataset: Meskipun lebih stabil dibanding Decision Tree, Random Forest masih menunjukkan MSE = 330.49 dan R^2 = 0.61, dengan indikasi overfitting.
- Gradient Boosting
  Gradient Boosting adalah model ansambel yang membangun pohon keputusan secara berurutan, dengan setiap pohon baru mencoba memperbaiki kesalahan dari pohon sebelumnya. Model ini sering digunakan untuk prediksi yang lebih presisi.
  Hasil pada Dataset: Model ini menunjukkan performa yang lebih baik dengan MSE = 302.59 dan R^2 = 0.64, namun masih sedikit underfitting.
- Support Vector Regressor (SVR)

SVR adalah model regresi yang menggunakan hyperplane dalam ruang fitur untuk memprediksi target, dengan tujuan meminimalkan margin error. Hasil pada Dataset: SVR menunjukkan performa moderat dengan MSE = 310.43 dan R^2 = 0.63, namun sedikit underfitting.

- K-Nearest Neighbors (KNN)
  KNN adalah model regresi berbasis instance yang memprediksi nilai target berdasarkan rata-rata nilai tetangga terdekat dalam ruang fitur. Hasil pada Dataset: Model ini menunjukkan performa yang kurang baik dengan MSE = 386.57 dan R^2 = 0.55, serta tanda-tanda overfitting.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil 1: Perbandingan model Linear Regression dengan model-model baseline lain

Bagian ini membahas hasil performa dari beberapa model baseline yang digunakan dalam penelitian. Model-model baseline yang dibandingkan meliputi Linear Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Regressor (SVR), dan K-Nearest Neighbor (KNN). Tabel berikut merangkum hasil evaluasi dari masing-masing model menggunakan Mean Squared Error (MSE) dan R-squared (R^2) sebagai matrik evaluasi:

Tabel 1. Performa Model Prediksi FoodPriceIndex

Model	MSE	R^2 Score	Training R^2	Validation R^2	Observations
Linear Regression	298.52	0.6522	0.6282	0.6522	Best model
Decision Tree	329.38	0.6163	0.6355	0.6163	Overfitting observed
Random Forest	329.73	0.6159	0.6352	0.6159	Overfitting observed
Gradient Boosting	302.59	0.6475	0.6321	0.6475	Slight underfitting
Support Vector Regressor	310.43	0.6384	0.6204	0.6384	Slight underfitting
K-Nearest Neighbor	386.58	0.5497	0.5801	0.5497	Poor generalization

Dari hasil di atas, model Linear Regression memberikan performa terbaik dengan nilai MSE sebesar 298.52 dan R^2 sebesar 0.6522. Model ini menunjukkan generalisasi yang baik dibandingkan model lain, terutama jika dibandingkan dengan model berbasis pohon seperti Decision Tree dan Random Forest yang cenderung overfitting pada data.

4.2 Hasil 2: Perbandingan method untuk mengisi null value

Pada penelitian ini, terdapat sejumlah nilai null pada kolom target FoodPriceIndex. Kami membandingkan tiga metode untuk mengisi nilai null, yaitu:

1. Menghapus baris dengan nilai null
    a. Korelasi antara kolom Year dan FoodPriceIndex: 0.79

2. Interpolasi nilai null
    a. Korelasi antara kolom Year dan FoodPriceIndex: 0.77
3. Menggunakan model Linear Regression untuk memprediksi nilai null
    a. Korelasi antara kolom Year dan FoodPriceIndex: 0.80

Berdasarkan hasil perhitungan korelasi, metode ketiga yang menggunakan model Linear Regression untuk memprediksi nilai null memberikan hasil terbaik dengan korelasi sebesar 0.80. Metode ini memiliki keunggulan dalam mempertahankan hubungan linier antara variabel waktu dan target dibandingkan dengan dua metode lainnya. Pemilihan metode ini didasarkan pada analisis korelasi yang menunjukkan bahwa pengisian nilai null dengan prediksi berbasis Linear Regression mampu mencerminkan pola data yang lebih baik dibandingkan metode interpolasi maupun penghapusan data.

4.3 Hasil 3: Feature selection

Kami memutuskan untuk menghilangkan fitur Country karena:

1. Harga Globalisasi/Global market

   pada Gambar 6, terlihat bahwa awalnya terdapat perbedaan harga antara country. Namun, pada tahun 2010-2015 semua country saling overlap di range harga yang sama. Sehingga kita berasumsi bahwa harga di tahun berikutnya untuk semua country berada di range yang mirip dan sama, maka dari itu kami menggunakan Linear Regression tanpa kolom country
2. Overfitting data dan Potensi Data Country yang Tidak Konsisten

   Setelah melakukan serangkaian eksperimen, kami menemukan bahwa memasukkan fitur Country seringkali meningkatkan performa model pada data validasi internal. Namun, saat menguji pada data leaderboard, model yang menyertakan fitur ini menunjukkan performa yang lebih rendah dibandingkan dengan model tanpa Country.
3. Eksperimen submission leaderboard

   Keputusan untuk tidak menggunakan Country juga berdasarkan eksperimen yang menunjukkan bahwa hasil prediksi dengan linear regression tanpa Country memberikan skor lebih baik di leaderboard
4. Potensi Data Country yang Tidak Konsisten

Data Country mungkin memiliki distribusi yang berbeda di antara data training dan testing di kompetisi. Kami menduga bahwa perbedaan ini terjadi karena Country memperkenalkan bias terhadap distribusi data training yang tidak sepenuhnya merepresentasikan data testing

Oleh karena itu, kami memutuskan untuk menggunakan linear regression sederhana tanpa Country, yang mampu memanfaatkan pola temporal secara global dan memberikan performa yang lebih stabil pada data unseen.

4.4 Hasil EDA

1. Distribusi Data
    a. Pola Distribusi:
        i. Distribusi FoodPriceIndex memiliki puncak yang signifikan pada nilai yang sangat rendah (dekat dengan 0). Hal ini menunjukkan bahwa sebagian besar data berada di kategori dengan nilai FoodPriceIndex rendah.
        ii. Selain itu, distribusi menunjukkan beberapa pola naik-turun dengan puncak tambahan di sekitar nilai tengah hingga mendekati 100.
    b. Frekuensi Tertinggi:
        i. Frekuensi tertinggi terjadi pada nilai mendekati 0, dengan jumlah observasi sekitar 1000. Ini mungkin menunjukkan adanya data yang cenderung terkonsentrasi pada kategori tertentu atau indikasi bahwa harga pangan di lokasi tertentu sangat rendah.
    c. Pola Multimodal:
        i. Kurva kernel density estimation (garis biru) menunjukkan adanya pola distribusi yang tidak sepenuhnya normal. Histogram tampak memiliki lebih dari satu puncak, sehingga distribusi ini bisa dianggap sebagai distribusi multimodal.
    d. Sebaran Data:
        i. Data tampaknya terdistribusi cukup merata di kisaran nilai 10 hingga 100 setelah puncak awal. Hal ini menunjukkan adanya variasi harga pangan di beberapa lokasi atau waktu tertentu.

e. Interpretasi:
    i. Puncak yang tinggi pada nilai rendah menunjukkan adanya kemungkinan ketidakseimbangan data, di mana sebagian besar pengamatan terkonsentrasi pada kategori tertentu (misalnya, lokasi atau periode dengan harga pangan sangat rendah).

2. Visualisasi Trend FoodPriceIndex
    a. Trend Umum:
        i. Grafik menunjukkan tren kenaikan FoodPriceIndex secara konsisten dari tahun 1970 hingga 2010 di hampir semua negara. Ini menandakan bahwa harga pangan cenderung meningkat dari waktu ke waktu.
    b. Perbandingan Antar Negara:
        i. Switzerland memiliki FoodPriceIndex yang lebih tinggi dibandingkan negara lain hampir sepanjang waktu. Hal ini mungkin disebabkan oleh tingkat ekonomi yang lebih tinggi atau biaya hidup yang lebih mahal di Swiss.
        ii. Negara seperti Mexico dan Turkey cenderung memiliki FoodPriceIndex yang lebih rendah di awal periode tetapi meningkat secara bertahap.
        iii. United States, France, dan Germany menunjukkan pola pertumbuhan yang lebih stabil, tanpa fluktuasi besar.
    c. Fluktuasi:
        i. Beberapa negara menunjukkan fluktuasi signifikan pada tahun tertentu, yang mungkin disebabkan oleh peristiwa global seperti krisis ekonomi, perubahan kebijakan perdagangan, atau bencana alam yang memengaruhi harga pangan.
    d. Kesenjangan Antar Negara:
        i. Pada awal periode (1970-an), ada kesenjangan yang cukup besar dalam FoodPriceIndex antara negara dengan indeks tinggi (seperti Switzerland) dan rendah (seperti Turkey). Namun, kesenjangan ini tampak mengecil seiring waktu.

BAB V
PENUTUP

5.1 Kesimpulan

Penelitian ini dilakukan dengan tujuan untuk menganalisis FoodPriceIndex menggunakan pendekatan regresi linier, serta mempertimbangkan teknik pengisian nilai null dan penghapusan variabel "Country" dalam model prediktif. Dalam rangka mencapai tujuan ini, langkah-langkah berikut telah dilakukan:

1. Penggunaan Regresi Linier:
    a. Linear Regression dipilih sebagai model utama karena kesederhanaan dan kemampuannya dalam menangkap hubungan linier antara variabel independen dan dependen.
    b. Hasil analisis menunjukkan bahwa regresi linier memiliki performa yang memadai untuk prediksi awal, dengan akurasi yang cukup tinggi pada data uji.
2. Perbandingan dengan Model Lain:
    a. Model regresi linier dibandingkan dengan algoritma lain seperti Random Forest dan Gradient Boosting. Model regresi linier tanpa kolom country memiliki performa yang lebih baik di data validasi dan juga leaderboard.
3. Pengisian Nilai Null Menggunakan Regresi Linier:
    a. Untuk menangani nilai null pada variabel FoodPriceIndex, metode imputasi berbasis prediksi menggunakan regresi linier diterapkan. Metode ini terbukti efektif dalam mengurangi bias yang dapat muncul akibat penghapusan langsung data null atau penggunaan imputasi rata-rata.
    b. Pemilihan pendekatan ini juga mendukung konsistensi model dengan menjaga pola hubungan antar variabel.
4. Penghapusan Variabel Country:
    a. Variabel "Country" diputuskan untuk di-drop dari analisis karena lebih bersifat kategorikal dan dapat mempersulit proses regresi tanpa encoding tambahan.

b. Selain itu, hasil eksperimen menunjukkan bahwa menyertakan fitur Country meningkatkan performa pada data validasi internal tetapi menurunkan akurasi pada data leaderboard. Hal ini diduga karena Country memperkenalkan bias yang tidak merepresentasikan data testing. Oleh karena itu, kami memilih menggunakan linear regression sederhana tanpa Country untuk memanfaatkan pola global dan meningkatkan stabilitas pada data unseen.

Dengan langkah-langkah ini, penelitian ini berhasil memberikan kontribusi dalam bentuk framework analisis yang efektif dan dapat diadaptasi pada data serupa. Keputusan yang diambil mendukung efisiensi proses analisis sekaligus menjaga kualitas hasil prediksi.

5.2 Saran

Untuk pengembangan lebih lanjut, beberapa saran berikut dapat dipertimbangkan:

1. Pengembangan Metode
    a. Meskipun regresi linier memberikan hasil yang memuaskan, model non-linear seperti Neural Networks atau XGBoost ataupun model time series seperti ARIMA dan yang lebih kompleks LSTM dapat dieksplorasi untuk menangkap hubungan yang lebih kompleks antar variabel.
    b. Dapat mengembangkan metode Linear Regression per country sehingga dapat menangkap masing masing kompleksitas country
2. Penyempurnaan Data
    a. Melibatkan lebih banyak variabel penjelas yang relevan seperti GDP atau inflasi, yang mungkin memengaruhi FoodPriceIndex secara signifikan.
    b. Eksplorasi lebih lanjut pada variabel "Country" untuk menyimpan informasi geografis dan meningkatkan hasil akurasi tanpa bias yang menurunkan akurasi
3. Aplikasi pada Dunia Nyata

a. Hasil penelitian ini dapat dimanfaatkan oleh pengambil kebijakan untuk memprediksi dan mengelola perubahan harga pangan, khususnya di negara-negara dengan fluktuasi harga yang tinggi.
b. Penelitian ini juga dapat digunakan oleh akademisi sebagai referensi untuk mengembangkan studi lanjutan terkait analisis data indeks harga pangan.

Dengan pengembangan metode dan pemanfaatan hasil penelitian yang tepat, diharapkan analisis ini dapat memberikan kontribusi yang lebih luas dalam mendukung pengambilan keputusan berbasis data.

DAFTAR PUSTAKA

[1] F. Arizki, M. C. Utami, dan E. Fetrina, "Perbandingan Metode Analisis Time Series untuk Peramalan Indeks Harga Saham," 2022.

[2] B. Audina, M. Fatekurohman, dan A. Riski, "Peramalan Arus Kas dengan Pendekatan Time Series Menggunakan Support Vector Machine," 2023.

[3] T. Indarwati, T. Irawati, dan E. Rimawati, "Penggunaan Metode Linear Regression untuk Prediksi Penjualan Smartphone," 2022.