

ANALISIS KOMPARATIF ALGORITMA MACHINE LEARNING UNTUK PREDIKSI PRODUK DOMESTIK REGIONAL BRUTO (PDRB) PROVINSI DI INDONESIA MENGGUNAKAN FITUR TIME LAG

Victor Asido Elyakim P, M. Alfathan Haris, Bagas Adi Nata,

Mirza Afif Pradipta, Muhammad Rizky Ramadhan

Sistem Informasi, STIKOM Tunas Bangsa

Jl. Jend. Sudirman Blok A, No.1, 2 & 3, Pematangsiantar, Sumatera Utara, Indonesia

victorasidoelyakim@gmail.com

ABSTRAK

Produk Domestik Regional Bruto (PDRB) merupakan indikator penting untuk menggambarkan tingkat aktivitas ekonomi wilayah yang mengalami dinamika fluktuatif akibat kondisi regional dan pandemi COVID-19, sehingga menuntut adanya pendekatan analitis untuk memprediksi perubahan ekonomi secara terukur. Penelitian ini membandingkan algoritma Linear Regression, Random Forest, dan XGBoost untuk memperkirakan distribusi PDRB provinsi Indonesia berdasarkan data BPS tahun 2020-2024. Data awal berbentuk matriks provinsi-tahun dirapikan, disusun ulang menjadi observasi provinsi per tahun, lalu diberi variabel *Lag_1* yang merekam nilai PDRB satu tahun sebelumnya. Dari 178 observasi awal, sebanyak 140 observasi digunakan setelah proses pembentukan fitur lag. Pengujian dilakukan dengan rasio pembagian data 80:20, validasi 5-fold, serta penyesuaian parameter melalui Grid Search untuk Random Forest dan XGBoost. Analisis memperlihatkan bahwa pola PDRB pada dataset ini lebih kuat dijelaskan oleh hubungan linier antartahun dibandingkan model ensemble yang kompleks. Berdasarkan hasil evaluasi kuantitatif, Linear Regression menghasilkan kesalahan (error) paling rendah dengan nilai akurasi prediksi yang direpresentasikan oleh MAE sebesar 0,042751, RMSE sebesar 0,070748, dan MAPE sebesar 3,28%. Selain itu, hasil validasi menggunakan 5-fold cross-validation juga membuktikan kestabilan model Linear Regression dengan perolehan rata-rata MAE sebesar 0,074940, yang jauh lebih baik dibandingkan model lainnya.

Kata kunci : PDRB, GRDP, Machine Learning, Linear Regression, Prediksi Ekonomi

1. PENDAHULUAN

Produk Domestik Regional Bruto (PDRB) merupakan indikator yang digunakan untuk menggambarkan tingkat aktivitas ekonomi suatu wilayah berdasarkan nilai tambah barang dan jasa yang dihasilkan dalam periode tertentu. Perubahan nilai PDRB antarprovinsi di Indonesia menunjukkan dinamika ekonomi yang berbeda-beda akibat pengaruh kondisi regional, kebijakan pembangunan, maupun perubahan ekonomi global. Periode setelah pandemi COVID-19 turut memengaruhi pola pertumbuhan ekonomi daerah sehingga diperlukan pendekatan analitis yang mampu membaca perubahan ekonomi secara lebih terukur.

Dalam penelitian ini, kondisi ekonomi regional direpresentasikan melalui distribusi nilai PDRB yang menggambarkan kontribusi masing-masing provinsi terhadap perekonomian nasional, yang mencerminkan perbedaan kontribusi ekonomi wilayah dari waktu ke waktu [1], [21].

Dalam konteks penelitian ini, PDRB tidak hanya dipakai sebagai ukuran ekonomi daerah, tetapi dijadikan data utama untuk membaca perubahan nilai antarprovinsi pada periode 2020–2024. Analisis terhadap nilai PDRB diperlukan karena perubahan kontribusi ekonomi setiap provinsi dapat menunjukkan dinamika perkembangan wilayah dan arah pergeseran aktivitas ekonomi daerah [14].

Oleh sebab itu, prediksi PDRB perlu dilakukan dengan pendekatan yang mampu memanfaatkan pola historis agar hasil analisis dapat mendukung pemahaman terhadap perkembangan ekonomi daerah secara lebih terukur [8].

Namun, perubahan nilai PDRB dari tahun ke tahun tidak selalu membentuk pola yang sederhana, sehingga proses prediksi perlu mempertimbangkan hubungan antarperiode melalui penggunaan fitur historis seperti lag [18].

Penelitian ini menggunakan data distribusi PDRB provinsi Indonesia periode 2020–2024 untuk menguji kemampuan beberapa algoritma machine learning dalam memprediksi nilai ekonomi antartahun menggunakan pendekatan fitur historis berbasis lag. Studi ini, fokus prediksi diarahkan pada hubungan nilai PDRB antarperiode menggunakan pendekatan berbasis fitur historis. Pendekatan tersebut cukup berguna pada data yang memiliki pola sederhana, tetapi sering kurang fleksibel ketika menghadapi hubungan non-linear dan interaksi antarvariabel ekonomi yang lebih rumit. Akibatnya, hasil prediksi dapat menjadi kurang optimal, terutama pada data yang berfluktuasi tinggi. Perlunya metode yang lebih mampu beradaptasi dengan pola data, salah satunya melalui penerapan Machine Learning [12].

Dalam penelitian ini, Machine Learning digunakan untuk menguji kemampuan model dalam membaca hubungan antara nilai PDRB tahun

sebelumnya dan nilai PDRB pada tahun berikutnya [4].

Studi ini, Linear Regression ditempatkan sebagai model dasar untuk membaca pola linier, sedangkan Random Forest dan XGBoost digunakan untuk menguji apakah model ensemble mampu menghasilkan prediksi yang lebih baik dibanding model linier pada data PDRB berbasis fitur historis.

Perbandingan ketiga model tersebut digunakan untuk menilai apakah data PDRB provinsi dengan satu fitur historis lebih tepat dijelaskan oleh model linier atau oleh model ensemble yang memiliki struktur lebih kompleks [2].

Selain pemilihan model, penelitian ini membentuk fitur lag dari nilai PDRB tahun sebelumnya agar setiap prediksi tetap mempertimbangkan hubungan antarperiode pada masing-masing provinsi [20].

Dengan pendekatan tersebut, model diharapkan dapat menghasilkan prediksi PDRB yang lebih baik.

Studi sebelumnya menunjukkan bahwa machine learning memiliki potensi dalam prediksi indikator ekonomi, namun performa model sangat dipengaruhi oleh karakteristik data dan jumlah fitur yang digunakan. Oleh karena itu, evaluasi empiris tetap diperlukan untuk menentukan model yang paling sesuai pada data distribusi PDRB provinsi Indonesia.

Penelitian terkait prediksi ekonomi umumnya berfokus pada pertumbuhan PDB nasional atau menggunakan banyak indikator makroekonomi. Kajian yang membandingkan model machine learning pada distribusi PDRB provinsi Indonesia menggunakan fitur lag tahunan masih relatif terbatas. Pemanfaatan variabel historis melalui fitur lag untuk meningkatkan akurasi prediksi PDRB provinsi di Indonesia masih perlu diteliti lebih lanjut.

Merujuk pada permasalahan tersebut, penelitian ini diarahkan untuk membandingkan performa beberapa algoritma Machine Learning dalam memprediksi distribusi PDRB provinsi di Indonesia. Tahapan yang dilakukan mencakup data preprocessing, transformasi data ke dalam format yang sesuai untuk Machine Learning, serta penerapan feature engineering berbasis lag [5].

Algoritma yang digunakan adalah Linear Regression, Random Forest, dan XGBoost. Evaluasi model dilakukan dengan menggunakan metrik Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan Mean Absolute Percentage Error (MAPE). Selain itu, penelitian ini juga menerapkan cross-validation dan Grid Search untuk memperoleh evaluasi yang lebih stabil serta konfigurasi parameter yang lebih sesuai.

Nilai kebaruan penelitian ini terletak pada penggabungan fitur lag, validasi menggunakan cross-validation, dan optimasi hyperparameter dalam satu rancangan eksperimen untuk membandingkan

beberapa algoritma Machine Learning pada kasus prediksi PDRB provinsi di Indonesia.

Penelitian ini dirancang sebagai eksperimen prediksi berbasis data sekunder untuk menguji hubungan antara nilai PDRB tahun sebelumnya dan nilai PDRB pada tahun target di tingkat provinsi Indonesia. Pendekatan kuantitatif digunakan karena data yang dianalisis berbentuk numerik dan hasil model dinilai melalui besaran kesalahan prediksi. Proses penelitian tidak hanya menerapkan algoritma Machine Learning, tetapi juga menyiapkan struktur data agar setiap observasi merepresentasikan kombinasi provinsi dan tahun. Dengan rancangan tersebut, model dapat diuji berdasarkan kemampuan membaca pola historis PDRB melalui fitur Lag₁.

2. TINJAUAN PUSTAKA

2.1. Studi Literatur

Penelitian mengenai peramalan indikator ekonomi telah banyak dieksplorasi menggunakan berbagai pendekatan algoritmik. Adewale et al. mengemukakan bahwa pemanfaatan metode ensemble machine learning mampu memberikan prediksi Produk Domestik Bruto (PDB) dengan baik [1].

Selain itu, Oancea dan Simionescu menunjukkan efektivitas pemanfaatan machine learning untuk peramalan PDB tingkat makro dalam pengaturan yang bersifat univariat [2].

Di konteks domestik, Passarella et al. melakukan komparasi beberapa model machine learning guna memprediksi pertumbuhan PDB di Indonesia [3].

Di ranah regional, Puttanapong et al. memprediksi PDRB tingkat provinsi dengan memanfaatkan algoritma machine learning yang digabungkan dengan data satelit [4].

Meskipun metode ensemble dan kompleks populer, pendekatan berbasis regresi sederhana terbukti tetap relevan, sebagaimana yang dikaji oleh Xiao dan Jin mengenai penelitian prediksi model regresi linier dalam peramalan ekonomi nasional [5].

2.2. Dinamika Produk Domestik Regional Bruto (PDRB)

Pertumbuhan ekonomi daerah yang direpresentasikan melalui PDRB dipengaruhi oleh berbagai determinan, termasuk otonomi daerah [6] serta modal sumber daya manusia [7].

Fluktuasi ini semakin terlihat di masa pandemi COVID-19 yang memberikan guncangan signifikan terhadap performa perekonomian tingkat provinsi [8], serta berdampak langsung pada ketimpangan pendapatan dan struktur ekonomi wilayah [8] [9].

Prediksi PDB/PDRB yang tepat melalui pendekatan machine learning karenanya krusial sebagai instrumen antisipasi dalam mencegah resesi ekonomi di masa mendatang [10].

2.3. Pendekatan Analisis Runtun Waktu dan Variabel Lag

Dalam melakukan peramalan indikator ekonomi berbasis runtun waktu, pemanfaatan informasi historis (lag) dan dekomposisi data sangat penting [11].

Goulet Coulombe et al. menyoroti bagaimana algoritma machine learning sangat berguna untuk melakukan peramalan makroekonomi yang melibatkan ketergantungan antarperiode [12].

Untuk mencapai performa terbaik, optimasi dalam model ensemble juga diperlukan, seperti yang dilakukan oleh Muhamad dan Matin menggunakan metode GridSearchCV pada model Random Forest [13].

2.4. Algoritma Pemodelan

Linear Regression digunakan sebagai model dasar untuk mengukur hubungan antara nilai PDRB periode sebelumnya dan nilai target pada periode berikutnya. Model ini dipilih karena sesuai untuk membaca hubungan linier sederhana pada data historis ekonomi berbasis fitur *Lag_1* Persamaan umum Linear Regression ditunjukkan pada Persamaan (1).

$$y_t = \beta^0 + \beta^1 y_{t-1} + \varepsilon \quad (1)$$

Keterangan: y_t merupakan nilai distribusi PDRB pada tahun ke- t ; y_{t-1} merupakan nilai distribusi PDRB pada tahun sebelumnya atau fitur lag; β_0 merupakan intercept; β_1 merupakan koefisien regresi; dan ε merupakan nilai error.

Random Forest diterapkan sebagai model pembandingan berbasis ensemble untuk mengevaluasi kemampuan pendekatan pohon keputusan dalam memprediksi nilai PDRB. Model ini bekerja melalui kombinasi sejumlah pohon keputusan untuk menghasilkan prediksi yang lebih stabil. Pada penelitian ini, Random Forest digunakan untuk menguji apakah pola hubungan PDRB dapat ditangkap lebih baik dibandingkan pendekatan regresi linier ketika memanfaatkan fitur historis yang terbatas. Persamaan umum Random Forest ditunjukkan pada Persamaan (2).

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(X) \quad (2)$$

Keterangan: $\hat{y}$ merupakan hasil prediksi model; $T_b(X)$ merupakan prediksi dari pohon keputusan ke- b ; dan B menunjukkan jumlah pohon keputusan yang digunakan dalam model.

XGBoost diterapkan untuk menguji kemampuan model boosting dalam memprediksi PDRB. Pada penelitian ini, XGBoost digunakan untuk melihat apakah pendekatan boosting tetap efektif ketika informasi prediktor yang tersedia hanya berasal dari nilai PDRB tahun sebelumnya. Pada penelitian ini, XGBoost ditempatkan sebagai pembandingan terhadap Linear Regression dan Random Forest untuk menilai apakah model boosting mampu memberi keuntungan pada data PDRB yang memiliki jumlah fitur terbatas. Fungsi objektif XGBoost ditunjukkan pada Persamaan (3).

$$Obj = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (3)$$

Keterangan: $L(y_i, \hat{y}_i)$ merupakan fungsi loss antara nilai aktual dan nilai prediksi; $\Omega(f_k)$ merupakan fungsi regularisasi; f_k merupakan fungsi pohon ke- k ; K menunjukkan jumlah pohon; sedangkan γ dan λ merupakan parameter regularisasi yang digunakan untuk mengontrol kompleksitas model.

Evaluasi model dilakukan menggunakan Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan Mean Absolute Percentage Error (MAPE). Ketiga metrik tersebut digunakan untuk mengukur tingkat kesalahan prediksi antara nilai aktual dan hasil prediksi model dari sudut pandang absolut, kuadrat kesalahan, dan persentase error. Semakin kecil nilai evaluasi yang diperoleh, semakin baik kemampuan model dalam melakukan prediksi. Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (6)$$

Keterangan: y_i merupakan nilai aktual, $\hat{y}_i$ merupakan nilai hasil prediksi, dan n menunjukkan jumlah data observasi yang digunakan dalam proses evaluasi.

3. METODE PENELITIAN

Penelitian ini dirancang sebagai eksperimen prediksi berbasis data sekunder untuk menguji hubungan antara nilai PDRB tahun sebelumnya dan nilai PDRB pada tahun target di tingkat provinsi Indonesia (Puttanapong, Prasertsoong, & Peechapat, 2023). Pendekatan kuantitatif digunakan karena data yang dianalisis berbentuk numerik dan hasil model dinilai melalui besaran kesalahan prediksi. Proses penelitian tidak hanya menerapkan algoritma Machine Learning, tetapi juga menyiapkan struktur data agar setiap observasi merepresentasikan kombinasi provinsi dan tahun. Dengan rancangan tersebut, model dapat diuji berdasarkan kemampuan membaca pola historis PDRB melalui fitur *Lag_1*.

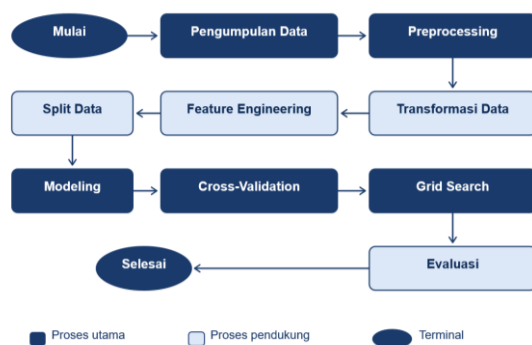
Tahapan penelitian dimulai dari penyusunan data distribusi PDRB provinsi Indonesia periode 2020–2024. Data awal masih berbentuk matriks provinsi-tahun, sehingga setiap kolom tahun perlu disusun ulang menjadi baris observasi terpisah. Proses transformasi dilakukan dari wide format ke long format agar satu baris data memuat satu provinsi, satu tahun, dan satu nilai PDRB. Setelah data tersusun dalam format tersebut, observasi diurutkan berdasarkan provinsi dan tahun untuk membentuk variabel *Lag_1*. Variabel ini berisi nilai PDRB pada tahun sebelumnya, sehingga data yang tidak memiliki nilai tahun sebelumnya tidak digunakan dalam proses pemodelan.

Tahap berikutnya adalah membentuk fitur lag. Fitur ini dibuat dari nilai PDRB tahun sebelumnya dan digunakan sebagai masukan model untuk memprediksi nilai PDRB pada tahun berikutnya. Contoh pembentukan variabel lag ditunjukkan pada Tabel 1.

Tabel 1. Contoh Transformasi Variabel Lag

Tahun	Lag_1	PDRB
2021	4.25	4.37
2022	4.37	4.51
2023	4.51	4.63

Berdasarkan Tabel 1, fitur Lag_1 dibentuk dengan mengambil nilai PDRB pada tahun sebelumnya sebagai masukan untuk memprediksi nilai PDRB pada tahun berikutnya. Dengan cara ini, setiap observasi tidak hanya memuat nilai PDRB tahun target, tetapi juga membawa informasi historis dari periode sebelumnya. Setelah proses pembentukan fitur selesai, dataset yang memiliki nilai Lag_1 digunakan dalam tahap pemodelan. Dataset yang telah melalui proses feature engineering selanjutnya dipisahkan menjadi data pelatihan dan data pengujian menggunakan proporsi 80% dan 20% untuk mengevaluasi kemampuan generalisasi model. Regresi linier dipilih sebagai baseline model untuk mengukur kemampuan pendekatan sederhana dalam memprediksi nilai PDRB berdasarkan informasi historis, sedangkan Random Forest dan XGBoost digunakan sebagai model pembanding. Evaluasi dilakukan menggunakan MAE, RMSE, dan MAPE untuk mengukur tingkat kesalahan prediksi model. Optimasi parameter dilakukan dengan Grid Search untuk memperoleh konfigurasi model yang lebih sesuai.

Gambar 1. Alur Tahapan Penelitian
Sumber. Hasil Rancangan Penelitian.

Metode dalam penelitian ini diarahkan untuk menguji hubungan antara nilai PDRB tahun sebelumnya sebagai variabel input dan nilai PDRB tahun berjalan sebagai target prediksi. Tiga algoritma yang dibandingkan adalah Linear Regression, Random

Forest, dan XGBoost. Pemilihan ketiga model tersebut dilakukan untuk membandingkan kemampuan model linier dan model ensemble dalam membaca pola PDRB yang hanya menggunakan satu fitur historis, yaitu Lag_1.

Dalam proses pemodelan, fitur utama yang digunakan adalah variabel lag, yaitu nilai PDRB pada tahun sebelumnya. Fitur ini digunakan karena data PDRB memiliki keterkaitan antarperiode, sehingga informasi pada periode sebelumnya dapat membantu model memperkirakan nilai pada periode berikutnya. Pendekatan ini diharapkan mampu menangkap pola temporal sederhana yang muncul pada data.

4. HASIL DAN PEMBAHASAN

Hasil eksperimen menunjukkan bahwa Linear Regression menghasilkan performa prediksi terbaik dibandingkan Random Forest dan XGBoost pada data distribusi PDRB provinsi Indonesia periode 2020–2024. Nilai MAE, RMSE, dan MAPE yang lebih rendah menunjukkan bahwa hubungan antara nilai PDRB tahun sebelumnya dan tahun target pada dataset cenderung mengikuti pola linier. Kondisi ini membuat model regresi sederhana lebih efektif dalam memanfaatkan informasi historis dibandingkan model ensemble yang memiliki tingkat kompleksitas lebih tinggi.

Performa Random Forest dan XGBoost yang belum optimal dapat dikaitkan dengan karakteristik dataset yang hanya menggunakan satu fitur utama, yaitu *Lag_1*. Pada kondisi tersebut, model berbasis pohon belum memperoleh variasi informasi yang cukup untuk membentuk pola keputusan yang lebih kompleks. Sebaliknya, hubungan sederhana antara nilai historis dan nilai target justru lebih mudah dipelajari oleh pendekatan linier.

Temuan penelitian ini menunjukkan bahwa peningkatan kompleksitas algoritma tidak selalu menghasilkan akurasi prediksi yang lebih baik. Pemilihan model perlu mempertimbangkan karakteristik data dan jumlah fitur yang tersedia. Pada kasus prediksi PDRB provinsi Indonesia dengan fitur historis terbatas, Linear Regression terbukti lebih sesuai dibanding model ensemble.

Meskipun demikian, penggunaan satu variabel historis masih menjadi keterbatasan penelitian ini. Penambahan indikator ekonomi lain, seperti inflasi, tingkat pengangguran, investasi, atau indeks pembangunan manusia, berpotensi meningkatkan kemampuan model dalam menangkap variasi ekonomi antarprovinsi secara lebih komprehensif.

Hasil evaluasi model pada data uji disajikan pada Tabel 3. Model dengan nilai MAE, RMSE, dan MAPE yang lebih rendah menunjukkan kemampuan prediksi yang lebih baik karena memiliki tingkat kesalahan yang lebih kecil terhadap nilai aktual.

Tabel 3. Perbandingan Kinerja Model Machine Learning

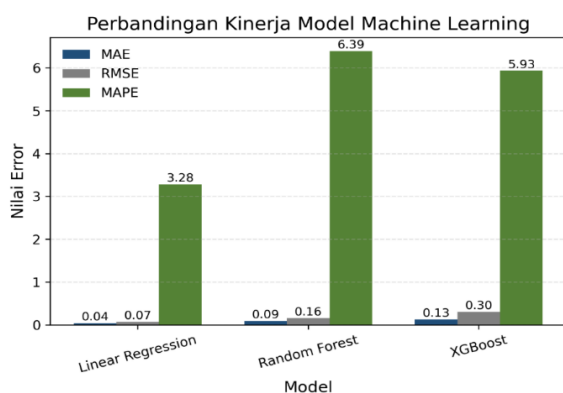
Model	MAE	RMSE	MAPE (%)
Linear Regression	0.042751	0.070748	3.280902
Random Forest	0.091229	0.156842	6.390301
XGBoost	0.128372	0.304261	5.932379

Berdasarkan Tabel 3, Linear Regression menjadi model dengan tingkat kesalahan paling kecil pada data uji. Model ini memperoleh MAE sebesar 0,042751, RMSE sebesar 0,070748, dan MAPE sebesar 3,280902%. Nilai tersebut lebih rendah dibandingkan Random Forest dan XGBoost, sehingga menunjukkan bahwa hubungan antara nilai PDRB tahun sebelumnya dan nilai PDRB tahun target pada dataset ini lebih mudah ditangkap melalui pola linier. Hasil ini juga mengindikasikan bahwa penggunaan satu fitur historis, yaitu Lag_1, belum memberikan informasi yang cukup bagi model ensemble untuk menghasilkan prediksi yang lebih baik.

Random Forest dan XGBoost menghasilkan nilai kesalahan yang lebih tinggi dibandingkan Linear Regression. Hasil ini menunjukkan bahwa kedua model ensemble tersebut belum memberikan performa yang lebih optimal pada dataset yang digunakan. Salah satu kemungkinan penyebabnya adalah pola data PDRB dalam penelitian ini cenderung linier, sementara jumlah fitur yang digunakan masih terbatas.

Penggunaan fitur lag tetap memberikan kontribusi penting dalam proses prediksi. Fitur ini membantu model mengenali pola historis PDRB antarperiode, sehingga informasi dari tahun sebelumnya dapat dimanfaatkan untuk memperkirakan nilai PDRB pada tahun berikutnya.

Perbandingan hasil evaluasi setiap model divisualisasikan pada Gambar 2. Visualisasi tersebut menunjukkan bahwa Linear Regression memiliki nilai MAE, RMSE, dan MAPE yang paling rendah. Model ini menjadi algoritma dengan performa terbaik pada eksperimen yang dilakukan.

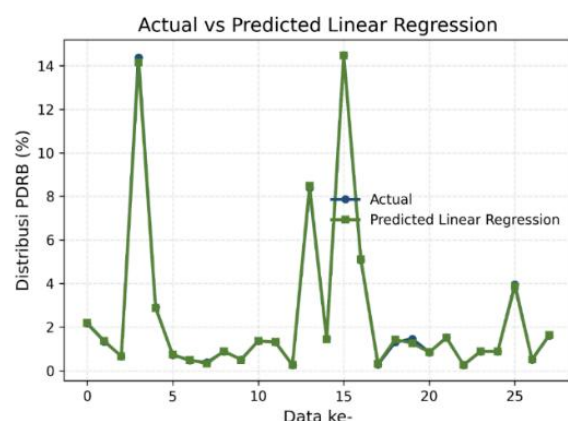
Gambar 2. Perbandingan Kinerja Model Machine Learning
Sumber: Hasil Penelitian

Pada Gambar 2, perbedaan performa antar model terlihat dari jarak nilai error yang cukup jelas. Linear Regression berada pada posisi paling rendah untuk seluruh metrik, sedangkan Random Forest dan XGBoost menunjukkan error yang lebih besar. Pola visual ini memperkuat hasil pada Tabel 3 bahwa model sederhana lebih sesuai untuk dataset yang hanya memanfaatkan hubungan PDRB antarperiode melalui Lag_1.

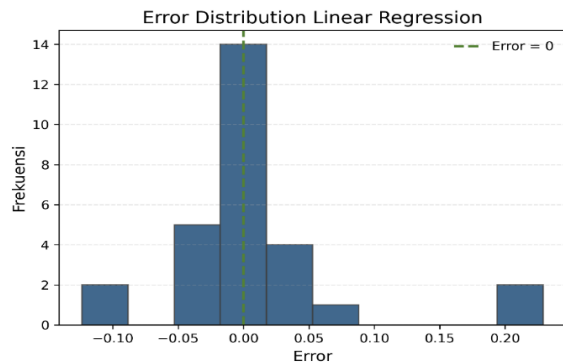
Random Forest dan XGBoost belum memberikan hasil yang lebih baik pada eksperimen ini. Meskipun keduanya termasuk model ensemble, performanya masih berada di bawah Linear Regression. Maka dapat terjadi karena data yang digunakan hanya memiliki fitur utama berupa Lag_1, sehingga pola hubungan antarperiode lebih sesuai dimodelkan dengan pendekatan linier.

Hasil validasi juga memperlihatkan kecenderungan yang sama. Linear Regression mampu mempertahankan nilai error yang lebih rendah, sehingga model ini dapat dikatakan lebih stabil pada data PDRB yang digunakan dalam penelitian.

Hasil penelitian ini sejalan dengan penelitian Adewale et al. (2024) dan Oancea & Simionescu (2024) yang menyatakan bahwa model sederhana seperti Linear Regression dapat memberikan performa optimal pada data ekonomi dengan pola hubungan yang cenderung linier.

Gambar 3. Grafik Actual vs Predicted Linear Regression
Sumber: Hasil Penelitian

Gambar 3 menunjukkan bahwa hasil prediksi Linear Regression masih berada dekat dengan pola nilai aktual, terutama pada observasi yang memiliki perubahan PDRB relatif kecil. Selisih yang muncul pada beberapa titik menunjukkan bahwa model masih memiliki keterbatasan ketika perubahan nilai antarperiode tidak sepenuhnya mengikuti pola linier.



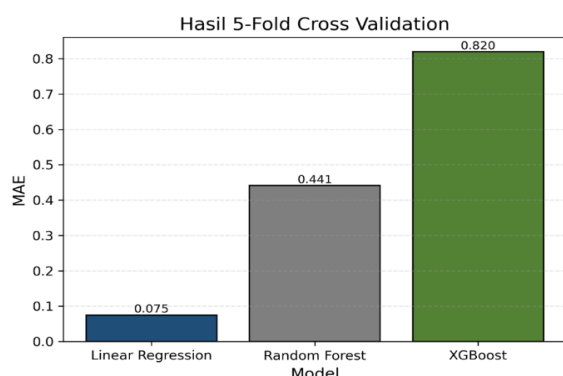
Gambar 4. Error Distribution Linear Regression
Sumber: Hasil Penelitian

Sebaran residual pada Gambar 4 menunjukkan bahwa sebagian besar kesalahan prediksi berada pada rentang kecil. Pola ini mendukung hasil MAE dan RMSE, tetapi juga memperlihatkan bahwa masih terdapat beberapa observasi dengan selisih prediksi yang lebih besar. Kondisi tersebut wajar karena model hanya menggunakan satu prediktor historis. Nilai MAPE sebesar 3,28% menunjukkan bahwa rata-rata kesalahan prediksi masih berada pada kategori rendah. Hasil prediksi yang diperoleh dapat dikatakan cukup dekat dengan nilai aktual pada data uji.

Pengujian tambahan dilakukan melalui 5-fold cross-validation untuk melihat kestabilan model pada beberapa pembagian data. Berdasarkan Tabel 4, Linear Regression memperoleh MAE rata-rata sebesar 0,074940, lebih rendah dibandingkan Random Forest sebesar 0,441107 dan XGBoost sebesar 0,819558. Perbedaan ini menunjukkan bahwa keunggulan Linear Regression tidak hanya terjadi pada pembagian data uji tertentu, tetapi juga tetap terlihat ketika data diuji melalui beberapa lipatan validasi. Model linier memiliki kestabilan error yang lebih baik pada struktur data PDRB yang digunakan.

Tabel 4. Hasil Cross-Validation (MAE)

Model	MAE
Linear Regression	0.074940
Random Forest	0.441107
XGBoost	0.819558



Gambar 5. Grafik Hasil 5-Fold Cross Validation
Sumber: Hasil Penelitian

Optimasi parameter dilakukan menggunakan Grid Search. Pada eksperimen akhir, Random Forest menggunakan `n_estimators` sebesar 100 dan `max_depth` sebesar 10. Sementara itu, XGBoost menggunakan `learning_rate` sebesar 0,1, `n_estimators` sebesar 200, dan `max_depth` sebesar 6. Parameter tersebut digunakan dalam pemodelan akhir untuk memperoleh hasil evaluasi yang lebih terarah.

Berdasarkan Gambar 5, Linear Regression tetap menunjukkan nilai MAE paling rendah dibandingkan Random Forest dan XGBoost. Hasil ini memperkuat bahwa Linear Regression memiliki performa yang lebih stabil pada data PDRB yang digunakan.

4.1. Pembahasan

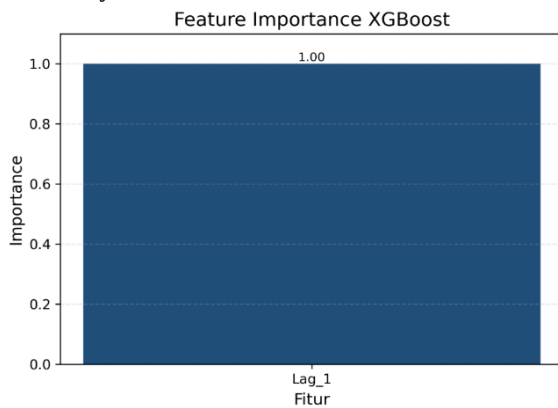
Hasil eksperimen memperlihatkan bahwa Dominasi performa Linear Regression menunjukkan bahwa hubungan antarperiode pada data PDRB cenderung bersifat linier. Kondisi ini mengindikasikan bahwa perubahan nilai PDRB dari satu tahun ke tahun berikutnya masih dapat dijelaskan secara memadai tanpa memerlukan struktur model yang terlalu kompleks. Keunggulan model ini tidak hanya tampak pada data uji, tetapi juga terlihat pada hasil 5-fold cross-validation. Hal tersebut menunjukkan bahwa hubungan antara `Lag_1` dan PDRB target pada periode 2020–2024 cenderung membentuk pola yang dapat dijelaskan secara linier. Dengan jumlah fitur yang terbatas, model linier mampu memanfaatkan informasi historis secara lebih langsung, sedangkan model ensemble membutuhkan variasi fitur yang lebih kaya untuk menunjukkan keunggulannya.

Performa Linear Regression yang lebih baik dapat dikaitkan dengan karakteristik dataset. Penelitian ini hanya menggunakan fitur utama berupa `Lag_1`, yaitu nilai PDRB pada periode sebelumnya. Dengan jumlah fitur yang terbatas, hubungan antara input dan target cenderung lebih sederhana sehingga lebih mudah ditangkap oleh model linier. Oleh karena itu, Linear Regression mampu memberikan hasil yang lebih efektif dibandingkan model ensemble.

Random Forest dan XGBoost belum menghasilkan peningkatan performa pada eksperimen ini karena informasi prediktor yang digunakan masih sangat terbatas. Dataset hanya memanfaatkan `Lag_1` sebagai fitur utama, sehingga model berbasis pohon tidak memiliki banyak variabel untuk membentuk pemisahan pola yang lebih kompleks. Pada kondisi tersebut, kompleksitas model justru tidak selalu memberi keuntungan karena variasi informasi yang tersedia belum cukup untuk membedakan pola antarprovinsi secara lebih mendalam.

Fitur `Lag_1` tetap menjadi komponen penting dalam proses prediksi. Fitur ini membantu model memanfaatkan nilai PDRB pada periode sebelumnya untuk memperkirakan nilai pada periode berikutnya. Hal ini juga diperkuat oleh feature importance XGBoost yang menunjukkan bahwa `Lag_1` menjadi fitur utama dalam proses prediksi.

Hasil penelitian ini menunjukkan bahwa pemilihan algoritma perlu disesuaikan dengan bentuk data dan jumlah fitur yang digunakan. Model yang lebih kompleks tidak selalu menghasilkan performa lebih tinggi. Pada kasus prediksi PDRB provinsi di Indonesia, Linear Regression lebih sesuai karena data yang digunakan memiliki pola yang relatif linier dan berbasis pada informasi historis satu periode sebelumnya.



Gambar 6. Feature Importance XGBoost
Sumber: Hasil Penelitian

Gambar 6 menunjukkan bahwa fitur Lag_1 memiliki nilai importance sebesar 1,00 pada model XGBoost. Nilai tersebut muncul karena hanya terdapat satu fitur utama yang digunakan dalam proses prediksi. Artinya, seluruh keputusan model bergantung pada informasi PDRB periode sebelumnya. Penemuan ini memperkuat bahwa rancangan penelitian sangat berpusat pada hubungan historis antarperiode. Namun, hasil tersebut juga menunjukkan keterbatasan model karena belum ada variabel ekonomi lain yang dapat membantu menjelaskan variasi PDRB antarprovinsi.

Hasil penelitian menunjukkan bahwa peningkatan kompleksitas algoritma tidak selalu berbanding lurus dengan peningkatan akurasi prediksi. Pada dataset yang hanya memiliki satu prediktor utama, model ensemble belum memperoleh informasi yang cukup untuk membentuk pola keputusan yang lebih unggul dibandingkan model linier.

Hasil penelitian dapat dimanfaatkan oleh pemerintah daerah dan perencana pembangunan sebagai referensi awal dalam memproyeksikan kondisi ekonomi wilayah. Model yang sederhana namun akurat dapat digunakan sebagai alat bantu pengambilan keputusan tanpa memerlukan sumber daya komputasi yang besar.

5. KESIMPULAN DAN SARAN

Penelitian ini menyimpulkan bahwa algoritma Linear Regression memberikan performa terbaik dibandingkan Random Forest dan XGBoost dalam memprediksi distribusi PDRB provinsi di Indonesia menggunakan fitur historis Lag_1. Keunggulan tersebut dibuktikan dengan hasil pengujian kuantitatif di mana Linear Regression memperoleh nilai kesalahan terendah, yaitu MAE sebesar 0,042751,

RMSE sebesar 0,070748, dan MAPE sebesar 3,28%. Hasil ini mengindikasikan bahwa pola perubahan PDRB pada periode 2020–2024 cenderung linier dan sangat bergantung pada nilai periode sebelumnya sehingga dapat dimodelkan secara efektif dengan regresi sederhana, sedangkan kinerja model ensemble kurang optimal akibat terbatasnya jumlah prediktor.

Oleh karena itu, disarankan bagi penelitian selanjutnya untuk memperluas rentang waktu pengamatan serta menambahkan variabel makroekonomi pendukung, seperti inflasi, tingkat pengangguran, investasi, Indeks Pembangunan Manusia (IPM), dan jumlah penduduk. Penambahan variabel ini penting agar karakteristik ekonomi antarprovinsi dapat dipetakan lebih komprehensif. Selain itu, jika jumlah periode data di masa depan sudah mencukupi, penggunaan metode pemodelan deret waktu (time series) tingkat lanjut seperti LSTM atau GRU sangat direkomendasikan agar model mampu menangkap pola historis non-linear yang lebih kompleks dari beberapa periode sekaligus.

DAFTAR PUSTAKA

- [1] Abdillah, K., Maulana, A., & Aminatuzzuhro. (2023). Determinants of Indonesia's Regional Economic Growth During The Era of Regional Autonomy. In *Jurnal Ekonomi Pembangunan* (Vol. 21, Number 02). <https://doi.org/10.22219/jep.v21i02.27934>
- [2] Adewale, M. D., Ebem, D. U., Awodele, O., Sambo-Magaji, A., Aggrey, E. M., Okechalu, E. A., Donatus, R. E., Olayanju, K. A., Owolabi, A. F., Oju, J. U., Ubadike, O. C., Otu, G. A., Muhammed, U. I., Danjuma, O. R., & Oluyide, O. P. (2024). Predicting gross domestic product using the ensemble machine learning method. *Systems and Soft Computing*, 6. <https://doi.org/10.1016/j.sasc.2024.200132>
- [3] Akita, T., & Alisjahbana, A. S. (2023). The Initial Impacts of the COVID-19 Pandemic on Regional Economies in Indonesia: Structural Changes and Regional Income Inequality. *Sustainability (Switzerland)*, 15(18). <https://doi.org/10.3390/su151813709>
- [4] Anesti, N., Kalamara, E., & Kapetanios, G. (2024). Forecasting with Machine Learning methods and multiple large datasets. *Econometrics and Statistics*. <https://doi.org/10.1016/j.ecosta.2024.08.003>
- [5] Bojer, C. S. (2022). Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. *International Journal of Forecasting*, 38(4), 1555–1561. <https://doi.org/10.1016/j.ijforecast.2021.11.003>
- [6] Foroni, C., Marcellino, M., & Stevanovic, D. (2022). Forecasting the Covid-19 recession and recovery: Lessons from the financial crisis. *International Journal of Forecasting*, 38(2),

- 596–612.
<https://doi.org/10.1016/j.ijforecast.2020.12.005>
- [7] Goulet Coulombe, P., Leroux, M., & Stevanovic, D. (2022). *How is Machine Learning Useful for Macroeconomic Forecasting?* *.
<https://doi.org/10.1002/jae.2910>
- [8] Herlina, H., & Viddy, A. (2023). *Projection of Gross Regional Domestic Product (GRDP) Indonesia-Malaysia of North Kalimantan Using the Least Squares Method* (pp. 155–163).
https://doi.org/10.2991/978-2-38476-202-6_23
- [9] Muhamad, I., & Matin, M. (2023). *Hyperparameter Tuning menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware*.
<https://doi.org/10.32722/multinetics.v9i1.5578>
- [10] Novianti, T., & Panjaitan, D. V. (2022). Income Inequality in Indonesia: Before and during the Covid-19 Pandemic. *International Journal of Economics and Financial Issues*, 12(3), 29–37.
<https://doi.org/10.32479/ijefi.12996>
- [11] Nugrahadi, E. W., Maipita, I., Fitrawaty, Hidayat, N., & Rinaldi, M. (2023). Impact of Labor and Health on Economic Growth in Indonesia During the COVID-19 Pandemic: A Panel Data Regression Analysis. *International Journal of Sustainable Development and Planning*, 18(12), 4015–4021.
<https://doi.org/10.18280/ijstdp.181232>
- [12] Oancea, B., & Simionescu, M. (2024). Gross Domestic Product Forecasting: Harnessing Machine Learning for Accurate Economic Predictions in a Univariate Setting †. *Electronics (Switzerland)*, 13(24).
<https://doi.org/10.3390/electronics13244918>
- [13] Passarella, R., Setiawan, M. I., & Yamani, Z. (2025). Comparative Analysis of Machine Learning Models for Predicting Indonesia's GDP Growth. *Acadlore Transactions on AI and Machine Learning*, 4(3), 157–173.
<https://doi.org/10.56578/ataiml040302>
- [14] Perkasa Jafba, G., & Juanda, R. (2023). *INFLUENCE OF GROSS REGIONAL DOMESTIC PRODUCT AND OPEN UNEMPLOYMENT RATE ON THE HUMAN DEVELOPMENT INDEKS IN LAMPUNG PROVINCE FOR THE YEARS 2017-2021*.
<http://ojs.unimal.ac.id/index.php/Jompe>
- [15] Primandari, A. H., & Arum, W. K. (2024). Analyzing the impact of the pandemic on Indonesia's economic growth using dynamic time warping. *Enthusiastic International Journal of Applied Statistics and Data Science*, 4(1), 11–20.
<https://journal.uui.ac.id/ENTHUSIASTIC>
- [16] Puttanapong, N., Prasertsong, N., & Peechapat, W. (2023). Predicting Provincial Gross Domestic Product Using Satellite Data and Machine Learning Methods: A Case Study of Thailand. *Asian Development Review*, 40(2), 39–85.
<https://doi.org/10.1142/S0116110523400024>
- [17] Rahmana Subian, A., Ali Mulkan, D., Hilmy Ahmady, H., Kartiasih, F., Studi DIV Komputasi Statistik, P., & Statistika STIS, P. (2023). SISTEMASI: Jurnal Sistem Informasi Comparison Methods of Machine Learning and Deep Learning to Forecast The GDP of Indonesia. In *Januari* (Vol. 13, Number 1).
<http://sistemasi.ftik.unisi.ac.id>
- [18] Saputri, P. D., Angrenani, A. B., & Fitriana, I. N. L. (2023). Preventing recession through GDP growth prediction: A classical and machine learning classification approach. *Data Science: Journal of Computing and Applied Informatics*, 7(2), 51–68.
<https://doi.org/10.32734/jocai.v7.i2-10507>
- [19] Tulus Rahayu, N., & Muharam, H. (2021). The Impact of The Covid-19 Pandemic on Provincial Economic Performance In Indonesia Article Information. In *Management Analysis Journal*.
<http://maj.unnes.ac.id>
- [20] Vianita, E., & Tantyoko, H. (2025). A Comparative Study of Machine Learning Models for Short-Term Load Forecasting. *Jurnal Masyarakat Informatika*, 16(1), 93–103.
<https://doi.org/10.14710/jmasif.16.1.73130>
- [21] Windhani, K., Purwaningsih, Y., Mulyaningsih, T., Samudro, B. R., & Hardoyono, F. (2023). Human Capital and Regional Economic Growth in Indonesia: A Spatial Analysis Approach. *Indonesian Journal of Geography*, 55(3), 473–487.
<https://doi.org/10.22146/ijg.88241>
- [22] Xiao, Y., & Jin, Z. (2021). The Forecast Research of Linear Regression Forecast Model in National Economy. *OALib*, 08(08), 1–17.
<https://doi.org/10.4236/oalib.1107797>