

Prediksi PDRB Provinsi di Indonesia Menggunakan Pendekatan Machine Learning Berbasis Data BPS 2021–2025

Provincial GRDP Prediction in Indonesia Using a Machine Learning Approach Based on BPS Data 2021–2025

Joel Fransiskus Parulian Sitorus¹, Ahmad Fariz Gustav², Ferdinan Silaen³, Diah Septiani⁴

^{1,2,3,4} Program Studi S1 Sains Data, Telkom University, Purwokerto, Jawa Tengah

e-mail: ¹fransiskusjoel@student.telkomuniversity.ac.id,

²ahmadgustav@student.telkomuniversity.ac.id, ³ferdinan@student.telkomuniversity.ac.id,

⁴diahseptiani@telkomuniversity.ac.id

Abstrak - Produk Domestik Regional Bruto (PDRB) merupakan indikator penting kinerja ekonomi daerah. Penelitian ini membandingkan lima model machine learning (Linear Regression, Ridge Regression, Decision Tree, Random Forest, dan Gradient Boosting) untuk memprediksi PDRB ADHB berdasarkan Indeks Pembangunan Manusia, persentase penduduk miskin, jumlah penduduk, dan Tingkat Pengangguran Terbuka menggunakan 178 data BPS periode 2021–2025. Data dibagi 80:20, dievaluasi dengan 5-fold cross-validation, dan dioptimasi menggunakan GridSearchCV. Hasil menunjukkan bahwa Gradient Boosting memberikan performa terbaik dengan nilai R^2 sebesar 0,9642 pada data uji dan meningkat menjadi 0,9724 setelah hyperparameter tuning. Analisis feature importance menunjukkan bahwa jumlah penduduk merupakan faktor paling berpengaruh dalam prediksi PDRB, diikuti oleh Indeks Pembangunan Manusia.

Kata kunci – gradient boosting; machine learning; PDRB; prediksi ekonomi regional; sosial-ekonomi.

Abstract - Gross Regional Domestic Product (GRDP) is an important indicator of regional economic performance. This study compared five machine learning models (Linear Regression, Ridge Regression, Decision Tree, Random Forest, and Gradient Boosting) to predict GRDP at Current Prices using four predictors: Human Development Index, poverty rate, total population, and Open Unemployment Rate, based on 178 observations from Statistics Indonesia for 2021–2025. The data were split into an 80:20 ratio, evaluated using 5-fold cross-validation, and optimized with GridSearchCV. The results showed that Gradient Boosting achieved the best performance with an R^2 of 0.9642 on the test set, improving to 0.9724 after hyperparameter tuning. Feature importance analysis indicated that total population was the most influential predictor of GRDP, followed by the Human Development Index.

Keywords – gradient boosting; machine learning; GRDP; regional economic prediction; socio-economic.

I. PENDAHULUAN

Produk Domestik Regional Bruto (PDRB) merupakan indikator utama untuk mengukur kinerja ekonomi suatu wilayah karena mencerminkan nilai tambah seluruh aktivitas ekonomi yang dihasilkan daerah dalam periode tertentu. Di Indonesia, pengukuran PDRB antarprovinsi menjadi penting mengingat masih terdapat disparitas pembangunan yang

cukup besar antara berbagai wilayah, sebagaimana dijelaskan dalam penelitian mengenai hubungan desentralisasi fiskal, pembangunan manusia, dan pertumbuhan ekonomi regional [1]. Oleh karena itu, PDRB tidak hanya berfungsi sebagai indikator ekonomi, tetapi juga sebagai gambaran tingkat ketimpangan pembangunan daerah.

Badan Pusat Statistik (BPS) menyediakan berbagai data sosial-ekonomi tingkat provinsi, seperti PDRB, Indeks Pembangunan Manusia (IPM), tingkat kemiskinan, jumlah penduduk, dan Tingkat Pengangguran Terbuka (TPT) [2]. Data tersebut dapat dimanfaatkan untuk membangun model prediksi PDRB. Namun, hubungan antara indikator sosial-ekonomi dan PDRB sering kali bersifat kompleks dan nonlinear sehingga sulit dijelaskan menggunakan pendekatan regresi konvensional.

Perkembangan machine learning memungkinkan pemodelan hubungan nonlinear secara lebih efektif. Algoritma berbasis pohon keputusan seperti Random Forest [3] dan Gradient Boosting [4] telah terbukti mampu memberikan akurasi tinggi dalam berbagai studi prediksi ekonomi. Penelitian terdahulu menunjukkan bahwa model ensemble, khususnya Random Forest dan Gradient Boosting, menghasilkan performa prediksi yang lebih baik dibandingkan model regresi linear maupun regresi teraturisasi [5], [6]. Selain itu, TPT juga dipilih sebagai prediktor karena memiliki keterkaitan dengan pertumbuhan output ekonomi sebagaimana dijelaskan dalam kajian Hukum Okun di Indonesia [7].

Penggunaan pustaka scikit-learn [8] dan pandas [9] telah menjadi standar dalam penelitian data science modern. Namun, penelitian yang membandingkan berbagai algoritma machine learning secara sistematis untuk prediksi PDRB ADHB provinsi Indonesia periode pascapandemi (2021–2025), disertai validasi silang dan hyperparameter tuning, masih relatif terbatas.

Kebaruan penelitian ini terletak pada perbandingan lima algoritma machine learning, yaitu Linear Regression, Ridge Regression, Decision Tree, Random Forest, dan Gradient Boosting, pada dataset sosial-ekonomi provinsi Indonesia periode 2021–2025. Seluruh model dievaluasi menggunakan 5-fold cross-validation dan dioptimasi melalui GridSearchCV untuk memperoleh model terbaik.

Berdasarkan hal tersebut, penelitian ini bertujuan menganalisis karakteristik data sosial-ekonomi provinsi Indonesia, mengidentifikasi faktor yang paling berpengaruh terhadap PDRB, membandingkan performa berbagai model machine learning dalam memprediksi PDRB, serta mengoptimalkan model terbaik melalui hyperparameter tuning guna menghasilkan prediksi yang lebih akurat dan bermanfaat bagi analisis ekonomi daerah.

II. TINJAUAN TEORI

A. PDRB Atas Dasar Harga Berlaku (ADHB)

PDRB ADHB menunjukkan nilai tambah barang dan jasa berdasarkan harga tahun berjalan. Indikator ini mencerminkan kapasitas ekonomi nominal suatu wilayah dan digunakan untuk melihat struktur ekonomi serta pendapatan per kapita [2].

B. Indeks Pembangunan Manusia (IPM)

IPM mengukur kualitas pembangunan manusia melalui dimensi kesehatan, pendidikan, dan standar hidup. IPM yang lebih tinggi umumnya berkaitan dengan produktivitas ekonomi yang lebih baik dan berkontribusi positif terhadap pertumbuhan ekonomi regional [1].

C. Tingkat Kemiskinan

Tingkat kemiskinan menunjukkan persentase penduduk yang hidup di bawah garis kemiskinan. Tingginya kemiskinan biasanya berkaitan dengan rendahnya daya beli, kualitas sumber daya manusia, dan produktivitas, sehingga berpotensi menurunkan kinerja ekonomi daerah.

D. Jumlah Penduduk

Jumlah penduduk memengaruhi skala ekonomi suatu wilayah karena berperan sebagai tenaga kerja dan konsumen. Semakin besar jumlah penduduk, semakin besar pula potensi pasar dan output ekonomi yang dapat dihasilkan, sehingga diperkirakan berkorelasi positif dengan PDRB.

E. Tingkat Pengangguran Terbuka (TPT)

TPT merupakan persentase angkatan kerja yang belum memperoleh pekerjaan. Berdasarkan Hukum Okun, pengangguran memiliki hubungan negatif dengan pertumbuhan ekonomi. Pada konteks Indonesia, hubungan tersebut terbukti ada meskipun relatif lemah dibandingkan negara maju [7].

F. Machine Learning untuk Prediksi Ekonomi

Metode machine learning banyak digunakan untuk memprediksi indikator ekonomi. Random Forest mampu menangkap hubungan nonlinear melalui kombinasi banyak pohon keputusan [3], sedangkan Gradient Boosting membangun model secara bertahap untuk memperbaiki kesalahan sebelumnya sehingga menghasilkan akurasi tinggi [4]. Linear Regression dan Ridge Regression digunakan sebagai model dasar yang mudah diinterpretasikan, sementara Decision Tree memberikan visualisasi pola keputusan yang sederhana. Penelitian terdahulu menunjukkan bahwa Gradient Boosting dan Random Forest secara konsisten menghasilkan performa terbaik dalam prediksi indikator ekonomi regional [5] maupun Produk Domestik Bruto [6].

III. METODE**A. Dataset**

Penelitian menggunakan data Badan Pusat Statistik (BPS) Indonesia periode 2021–2025 [2]. Dataset terdiri atas variabel target PDRB ADHB serta empat variabel prediktor, yaitu Indeks Pembangunan Manusia (IPM), persentase penduduk miskin, jumlah penduduk, dan Tingkat Pengangguran Terbuka (TPT). Variabel Provinsi dan Tahun digunakan sebagai identifikasi data.

Tabel 1. Deskripsi Variabel Dataset (BPS, 2021–2025)

No.	Nama Kolom	Keterangan	Peran
1	Nama Provinsi	Nama provinsi di Indonesia	Identifikasi
2	Tahun	Tahun observasi (2021–2025)	Identifikasi

No.	Nama Kolom	Keterangan	Peran
3	PDRB ADHB (Miliar Rp)	Produk Domestik Regional Bruto Atas Dasar Harga Berlaku	Target (Y)
4	IPM	Indeks Pembangunan Manusia	Prediktor (X1)
5	Kemiskinan (%)	Persentase Penduduk Miskin	Prediktor (X2)
6	Jumlah Penduduk (Ribuan Jiwa)	Jumlah penduduk provinsi	Prediktor (X3)
7	TPT (%)	Tingkat Pengangguran Terbuka	Prediktor (X4)

Tabel 1. di atas menyajikan deskripsi variabel-variabel yang digunakan dalam dataset penelitian ini, yang bersumber dari Badan Pusat Statistik (BPS) untuk periode tahun 2021–2025. Dataset terdiri atas tujuh kolom dengan peran yang berbeda-beda, yaitu dua variabel identifikasi berupa Nama Provinsi dan Tahun yang berfungsi untuk menandai unit observasi pada setiap data, satu variabel target (Y) yaitu PDRB Atas Dasar Harga Berlaku (ADHB) dalam satuan miliar rupiah yang merupakan variabel yang ingin diprediksi, serta empat variabel prediktor (X1–X4) yang terdiri dari Indeks Pembangunan Manusia (IPM), persentase penduduk miskin, jumlah penduduk dalam ribuan jiwa, dan Tingkat Pengangguran Terbuka (TPT) dalam persen. Keempat variabel prediktor tersebut dipilih karena secara teoritis maupun empiris memiliki keterkaitan dengan kondisi perekonomian suatu provinsi, sehingga relevan digunakan untuk menjelaskan atau memprediksi besaran PDRB sebagai indikator utama pertumbuhan ekonomi regional.

B. Alur Penelitian

Tahapan penelitian meliputi pengumpulan data, pembersihan data, analisis deskriptif, transformasi data, pembangunan model, evaluasi model, dan hyperparameter tuning. Implementasi dilakukan menggunakan Python dengan pustaka pandas [9] dan scikit-learn [9] pada Google Colab.

C. Pembersihan Data (Data Cleaning)

Proses pembersihan mencakup penanganan nilai hilang dengan menghapus baris yang memiliki data numerik tidak lengkap, penghapusan data duplikat, serta deteksi outlier menggunakan metode Interquartile Range (IQR) [10]. Outlier yang ditemukan tetap dipertahankan karena merepresentasikan kondisi riil provinsi dengan aktivitas ekonomi dan populasi yang sangat besar.

D. Analisis Deskriptif dan Visualisasi

Analisis dilakukan melalui statistik deskriptif (rata-rata, median, standar deviasi, minimum, dan maksimum) serta visualisasi berupa histogram, grafik tren PDRB, heatmap korelasi, scatter plot, boxplot, dan diagram batang provinsi dengan rata-rata PDRB tertinggi.

E. Transformasi Data

Fitur yang digunakan sebagai input model adalah IPM, Kemiskinan, Jumlah Penduduk, dan TPT, sedangkan PDRB ADHB digunakan sebagai target. Standarisasi menggunakan StandardScaler [8] diterapkan pada model regresi, sementara model berbasis pohon menggunakan data asli karena tidak sensitif terhadap skala fitur.

F. Pembangunan Model

Data dibagi menjadi data latih dan data uji dengan rasio 80:20 menggunakan `train_test_split`

(random_state=42). Lima model yang dibandingkan adalah Linear Regression, Ridge Regression, Decision Tree, Random Forest [3], dan Gradient Boosting [4]. Evaluasi stabilitas model dilakukan menggunakan 5-fold cross-validation. Model terbaik kemudian dioptimasi menggunakan GridSearchCV dengan kombinasi parameter *n_estimators*, *learning_rate*, dan *max_depth*.

G. Metrik Evaluasi

Kinerja model dievaluasi menggunakan Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 Score, dan Mean Absolute Percentage Error (MAPE). Kombinasi metrik tersebut digunakan untuk mengukur tingkat kesalahan dan kemampuan model dalam menjelaskan variasi nilai PDRB.

IV. HASIL DAN PEMBAHASAN

A. Eksplorasi Data Awal (EDA)

1) Pembersihan Data

Pada tahap pembersihan data, nilai hilang dan data duplikat diperiksa, kemudian baris yang mengandung missing values dihapus. Jumlah data berkurang dari 190 menjadi 178 observasi, dan hasil akhir menunjukkan dataset telah bersih serta siap untuk analisis lebih lanjut.

2) Deteksi dan Penanganan Outlier

Deteksi outlier menggunakan metode IQR menunjukkan adanya outlier pada seluruh variabel, terutama PDRB_ADHB (19 data) dan Jumlah_Penduduk (15 data). Outlier tersebut tetap dipertahankan karena mencerminkan kondisi nyata provinsi dengan skala ekonomi dan populasi yang jauh lebih besar dibandingkan provinsi lain. Hasil boxplot juga mengonfirmasi adanya variasi data yang tinggi, terutama pada variabel PDRB_ADHB dan Jumlah Penduduk.

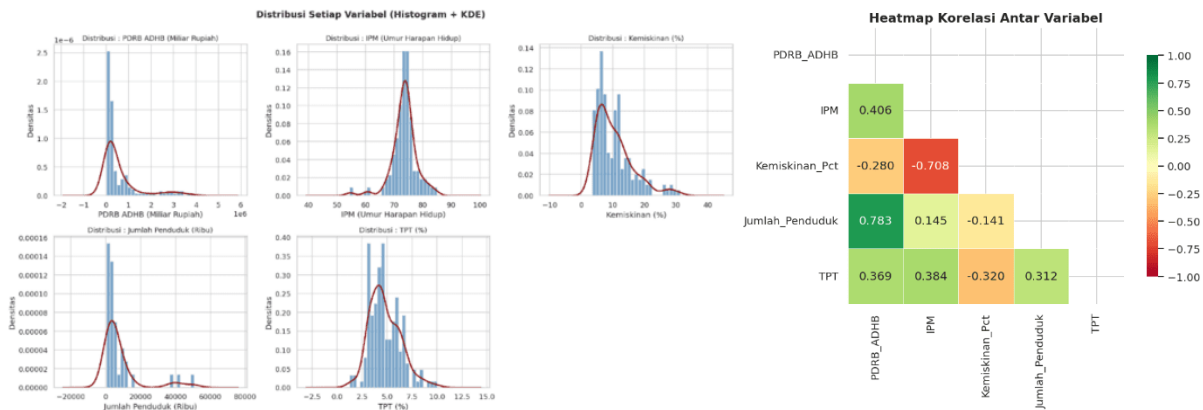
3) Analisis Deskriptif

Tabel 2. Statistik Deskriptif (Count, Mean, Std, Min, Q1, Median, Q3, Max)

Variabel	Count	Mean	Std Dev	Min	Q1	Median	Q3	Max
PDRB_ADHB	178	573293,62	829954,59	26561,43	117289,42	242128,92	646778,24	3926153,30
IPM	178	73,56	4,28	54,43	71,83	73,84	75,33	85,05
Kemiskinan_Pct	178	10,47	5,77	3,57	6,18	8,72	12,76	31,32
Jumlah_Penduduk	178	7844,31	11335,52	542,10	1992,98	4114,00	8195,55	50759,00
TPT	178	4,80	1,54	1,25	3,65	4,48	5,88	10,02

Tabel 2. menyajikan statistik deskriptif untuk kelima variabel numerik dengan jumlah data sebanyak 178 observasi pada masing-masing variabel. PDRB_ADHB menunjukkan rentang nilai yang sangat lebar, dari minimum 26.561,43 hingga maksimum 3.926.153,30 miliar rupiah dengan standar deviasi yang besar (829.954,59), mengindikasikan tingginya kesenjangan ekonomi antarprovinsi. Variabel IPM relatif lebih stabil dengan rata-rata 73,56 dan standar deviasi 4,28, sedangkan Kemiskinan_Pct, Jumlah_Penduduk, dan TPT juga menunjukkan variasi yang cukup besar antarprovinsi, terlihat dari selisih signifikan antara nilai minimum dan maksimum pada masing-masing variabel.

4) Visualisasi Distribusi dan Hubungan Antarvariabel



Gambar 1. Distribusi Setiap Variabel (Histogram + KDE) & Heatmap Korelasi Antar Variabel

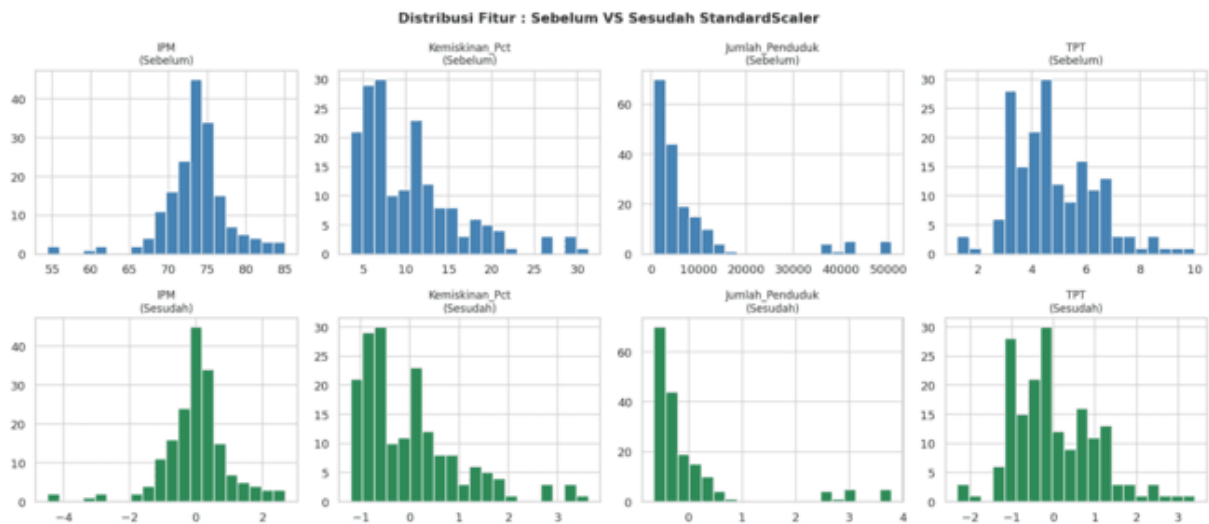
Gambar 1. menunjukkan distribusi data pada variabel PDRB_ADHB, Jumlah_Penduduk, dan Kemiskinan_Pct cenderung right-skewed, dengan sebagian besar observasi berada pada nilai rendah dan beberapa provinsi memiliki nilai yang jauh lebih tinggi. Sementara itu, IPM dan TPT relatif mendekati distribusi normal. Analisis korelasi menunjukkan bahwa Jumlah_Penduduk memiliki hubungan positif terkuat dengan PDRB_ADHB ($r = 0,783$), diikuti oleh IPM ($r = 0,406$), sedangkan Kemiskinan_Pct berkorelasi negatif terhadap PDRB_ADHB ($r = -0,280$) dan IPM ($r = -0,708$). Temuan ini mengindikasikan bahwa provinsi dengan jumlah penduduk dan kualitas pembangunan manusia yang lebih tinggi cenderung memiliki PDRB yang lebih besar, sementara tingkat kemiskinan yang tinggi berasosiasi dengan IPM dan PDRB yang lebih rendah.

B. Transformasi Data

1) Seleksi Fitur dan Target

Berdasarkan hasil EDA, empat variabel dipilih sebagai fitur prediktor, yaitu IPM, Kemiskinan_Pct, Jumlah_Penduduk, dan TPT. Kolom Provinsi tidak dimasukkan karena merupakan variabel kategorikal nominal, dan kolom Tahun tidak dimasukkan karena tujuan model adalah memprediksi PDRB berdasarkan kondisi sosial-ekonomi, bukan berdasarkan waktu. Variabel target adalah PDRB_ADHB.

2) Standarisasi Data



Gambar 2. Distribusi Fitur : Sebelum VS Sesudah StandardScaler

Gambar 2. menampilkan perbandingan distribusi keempat fitur prediktor sebelum dan sesudah dilakukan standarisasi menggunakan StandardScaler. Terlihat bahwa bentuk distribusi masing-masing variabel, baik IPM, Kemiskinan_Pct, Jumlah_Penduduk, maupun TPT, tetap sama atau tidak berubah secara bentuk setelah proses scaling, namun skala nilainya berubah menjadi terstandarisasi dengan rata-rata mendekati 0 dan rentang nilai yang lebih seragam antarvariabel. Hal ini menunjukkan bahwa proses StandardScaler berhasil menyamakan skala antarfitur tanpa mengubah pola atau bentuk distribusi aslinya, sehingga data siap digunakan untuk tahap pemodelan yang sensitif terhadap perbedaan skala antarvariabel.

C. Pembangunan dan Evaluasi Model

1) Pembagian Data Latih dan Uji

Data dibagi dengan rasio 80:20 menggunakan `random_state=42`. Versi data yang telah distandarisasi digunakan untuk Linear Regression dan Ridge Regression, sedangkan versi data asli (unscaled) digunakan untuk model berbasis pohon (Decision Tree, Random Forest, Gradient Boosting) karena model tersebut tidak bergantung pada skala fitur.

2) Perbandingan Model

Lima model dibandingkan, yaitu Linear Regression, Ridge Regression, Decision Tree, Random Forest [3], dan Gradient Boosting [4]. Evaluasi dilakukan menggunakan MAE, RMSE, MAPE, R^2 , dan 5-fold cross-validation.

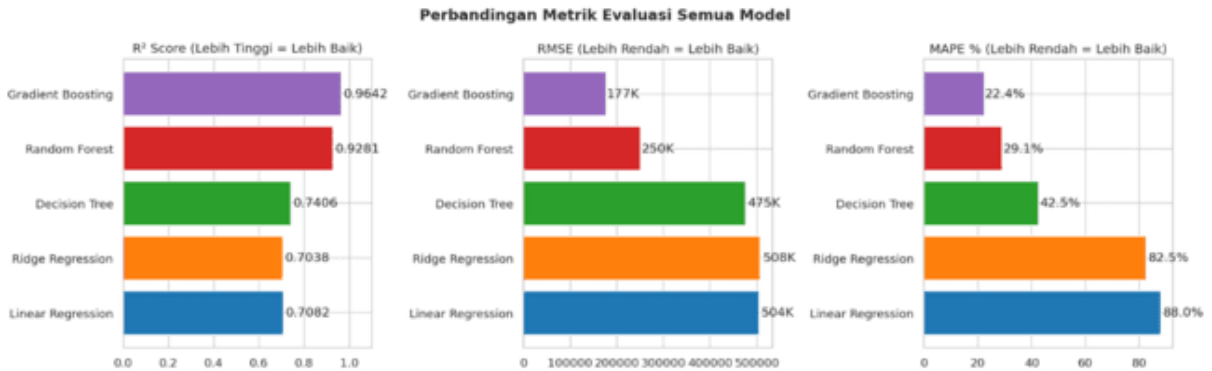
Tabel 3. Rekap Evaluasi Semua Model (R^2 , RMSE, CV R^2)

Model	MAE	MSE	RMSE	R^2	MAPE (%)	CV R^2 Mean	CV R^2 Std
Linear Regression	276962,99	2,542047e+11	504187,16	0,7082	87,9598	0,5968	0,0974
Ridge Regression	269221,56	2,580169e+11	507953,61	0,7038	82,4918	0,6134	0,1077
Decision Tree	231663,17	2,259504e+11	475342,43	0,7406	42,5065	0,8410	0,1645
Random Forest	142969,26	6,261899e+10	250237,87	0,9281	29,1374	0,8943	0,0687

Model	MAE	MSE	RMSE	R ²	MAPE (%)	CV R ² Mean	CV R ² Std
Gradient Boosting	95183,70	3,119589e+10	176623,57	0,9642	22,4049	0,8899	0,1433

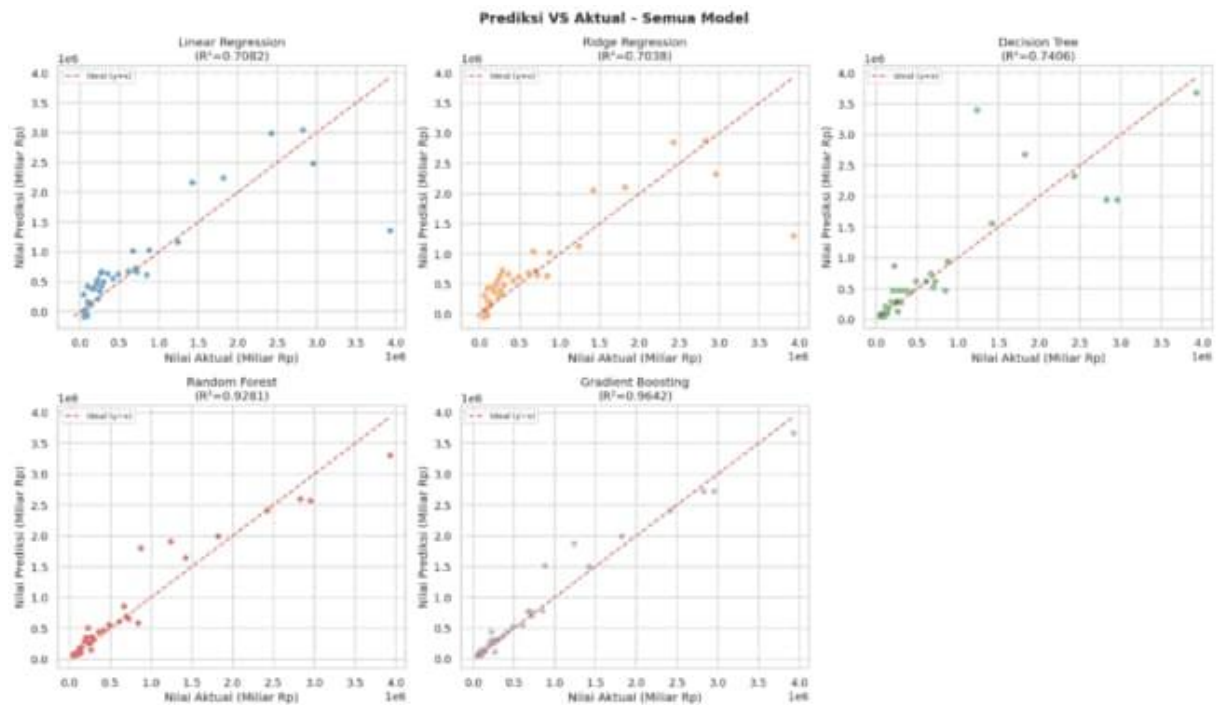
Tabel 3. menyajikan rekapitulasi hasil evaluasi lima model regresi yang digunakan untuk memprediksi PDRB_ADHB, berdasarkan metrik MAE, MSE, RMSE, R², MAPE, serta hasil cross-validation (CV R² Mean dan Std). Berdasarkan tabel tersebut, Gradient Boosting menunjukkan performa terbaik dengan nilai R² tertinggi (0,9642) dan error terendah (MAE 95.183,70; MAPE 22,41%), diikuti oleh Random Forest yang juga mencatat performa baik (R² = 0,9281; CV R² Mean = 0,8943) dengan tingkat stabilitas validasi yang paling konsisten (CV R² Std = 0,0687). Sebaliknya, model linear seperti Linear Regression dan Ridge Regression menunjukkan performa paling rendah dengan R² di bawah 0,71 dan MAPE di atas 80%, mengindikasikan bahwa hubungan antara variabel prediktor dan target bersifat non-linear, sehingga model berbasis ensemble tree seperti Gradient Boosting dan Random Forest lebih mampu menangkap pola kompleks dalam data dibandingkan model regresi linear.

3) Hasil Evaluasi Model



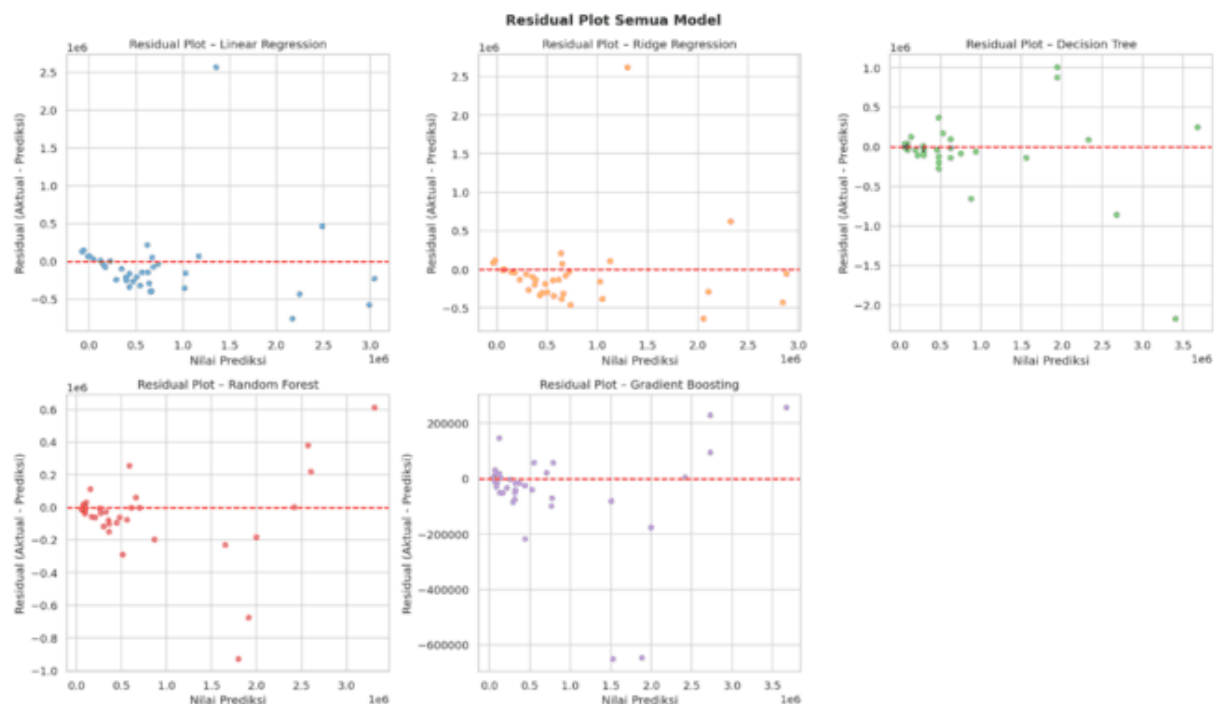
Gambar 3. Perbandingan Metrik Evaluasi Semua Model

Gambar 3. menyajikan perbandingan visual ketiga metrik evaluasi utama, yaitu R² Score, RMSE, dan MAPE, untuk kelima model yang diuji. Pada ketiga grafik tersebut, Gradient Boosting secara konsisten menempati posisi terbaik dengan R² tertinggi (0,9642), RMSE terendah (177K), dan MAPE terendah (22,4%), sedangkan Linear Regression dan Ridge Regression menunjukkan performa paling lemah pada seluruh metrik, mengonfirmasi superioritas model berbasis ensemble tree dalam memprediksi PDRB_ADHB.



Gambar 4. Prediksi VS Aktual – Semua Model

Gambar 4. memperlihatkan perbandingan antara nilai prediksi dan nilai aktual untuk seluruh model, dengan garis putus-putus merah merepresentasikan kondisi ideal ($y=x$). Titik-titik pada model Gradient Boosting dan Random Forest terlihat paling rapat mengikuti garis ideal tersebut, menandakan akurasi prediksi yang tinggi, sedangkan pada Linear Regression, Ridge Regression, dan Decision Tree terlihat lebih banyak titik yang menyebar jauh dari garis ideal, terutama pada nilai aktual yang besar.



Gambar 5. Residual Plot Semua Model

Gambar 5. menyajikan residual plot kelima model untuk mengevaluasi pola error prediksi

terhadap nilai prediksi. Residual pada model Random Forest dan Gradient Boosting cenderung tersebar lebih rapat di sekitar garis nol dibandingkan model lainnya, meski masih terdapat beberapa nilai residual yang cukup besar pada prediksi dengan nilai tinggi, sedangkan pada Linear Regression dan Ridge Regression terlihat pola residual yang lebih menyebar, mengindikasikan tingkat kesalahan prediksi yang lebih besar dan kurang konsisten.

D. Pemilihan Model Terbaik dan Hyperparameter Tuning

1) Pemilihan Model Terbaik

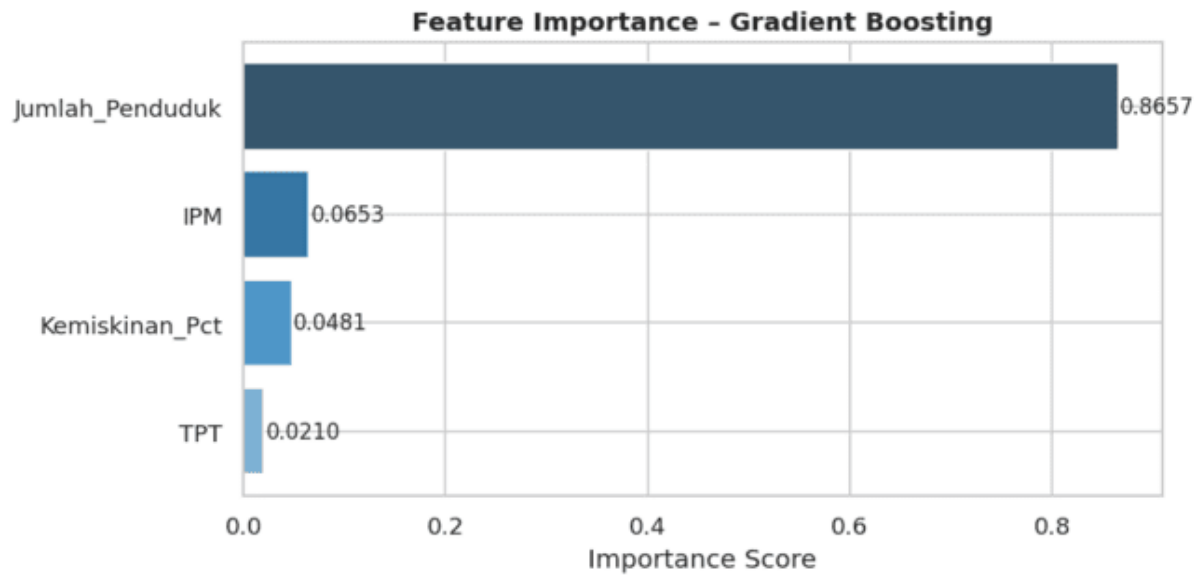
Model terbaik dipilih berdasarkan kombinasi nilai R^2 tertinggi pada data uji dan nilai CV R^2 yang tinggi dengan standar deviasi rendah. Berdasarkan hasil evaluasi, Gradient Boosting Regressor terpilih sebagai model terbaik secara konsisten pada semua metrik evaluasi.

MODEL TERBAIK : Gradient Boosting		
=====		
MAE	:	95,183.6986
MSE	:	31,195,885,188.2909
RMSE	:	176,623.5692
R^2	:	0.9642
MAPE (%)	:	22.4049
CV R^2 Mean	:	0.8899
CV R^2 Std	:	0.1433

Gambar 6. Identifikasi Model Terbaik

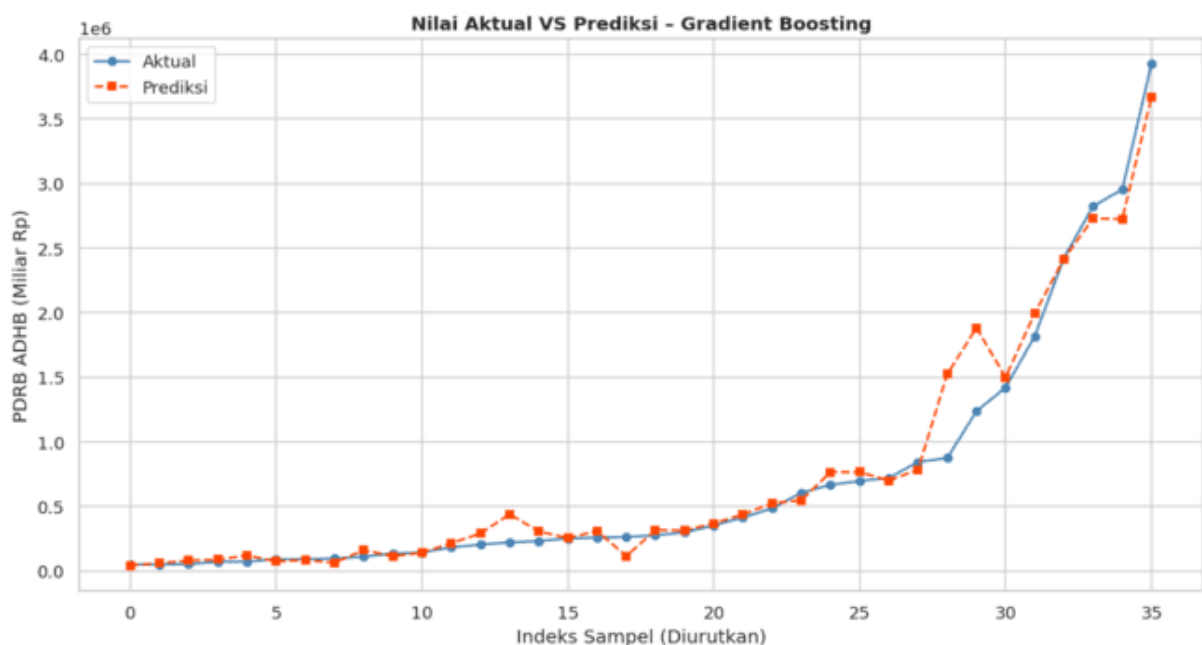
Gambar 6. menampilkan ringkasan identifikasi model terbaik berdasarkan hasil evaluasi sebelumnya, yaitu Gradient Boosting, dengan nilai R^2 sebesar 0,9642 dan RMSE sebesar 176.623,57, yang berarti model ini mampu menjelaskan sekitar 96,42% variasi PDRB_ADHB dengan tingkat kesalahan prediksi rata-rata (MAPE) sebesar 22,40%. Hasil cross-validation juga menunjukkan rata-rata R^2 sebesar 0,8899 dengan standar deviasi 0,1433, mengonfirmasi bahwa Gradient Boosting merupakan model dengan performa dan generalisasi terbaik dibandingkan keempat model lainnya dalam memprediksi PDRB ADHB pada dataset ini.

2) Feature Importance



Gambar 7. Feature Importance – Gradient Boosting

Gambar 7. menampilkan hasil analisis tingkat kepentingan fitur (feature importance) pada model Gradient Boosting sebagai model terbaik. Variabel Jumlah_Penduduk mendominasi kontribusi terhadap prediksi PDRB_ADHB dengan skor importance sebesar 0,8657, jauh melampaui ketiga fitur lainnya, yaitu IPM (0,0653), Kemiskinan_Pct (0,0481), dan TPT (0,0210). Hasil ini menunjukkan bahwa jumlah penduduk merupakan faktor paling dominan dan determinan dalam menjelaskan variasi PDRB antarprovinsi, sejalan dengan temuan korelasi yang kuat pada analisis sebelumnya, sementara ketiga variabel lainnya hanya memberikan kontribusi yang relatif kecil terhadap performa model.



Gambar 8. Nilai Aktual VS Prediksi – Gradient Boosting

Gambar 8. menampilkan perbandingan nilai aktual dan nilai prediksi pada data uji yang telah diurutkan berdasarkan besaran nilai, menggunakan model Gradient Boosting. Secara umum, garis prediksi mengikuti pola garis aktual dengan cukup baik di sepanjang rentang data,

mengindikasikan kemampuan model dalam menangkap tren peningkatan PDRB ADHB. Namun, terdapat beberapa selisih yang cukup terlihat pada beberapa indeks sampel, terutama pada sampel dengan nilai tinggi (indeks 28–35), di mana model sedikit melebihi atau mengurangi nilai prediksi dibandingkan nilai aktualnya, yang mengindikasikan model masih memiliki tingkat kesalahan yang lebih besar pada provinsi dengan PDRB sangat tinggi.

3) Hyperparameter Tuning

```
Hyperparameter Tuning : Gradient Boosting
Best Params : {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 300}
Best CV R2 : 0.8466

Hasil Setelah Tuning :
R2 = 0.9724
RMSE = 155,076.73
MAE = 97,036.09
```

Gambar 9. Hyperparameter Tuning : Gradient Boosting

Gambar 9. menampilkan hasil hyperparameter tuning pada model Gradient Boosting, dengan parameter terbaik yang diperoleh berupa `learning_rate` sebesar 0,1, `max_depth` sebesar 3, dan `n_estimators` sebanyak 300, yang menghasilkan nilai Best CV R² sebesar 0,8466. Setelah dilakukan tuning, performa model meningkat dibandingkan model sebelum tuning, terlihat dari nilai R² yang naik menjadi 0,9724 serta RMSE dan MAE yang menurun menjadi masing-masing 155.076,73 dan 97.036,09, menunjukkan bahwa proses optimasi hyperparameter berhasil meningkatkan akurasi dan generalisasi model Gradient Boosting dalam memprediksi PDRB ADHB.

E. Interpretasi Hasil

1) Kinerja Model Terbaik

Gradient Boosting Regressor menghasilkan nilai R² tertinggi pada data uji, yang berarti sebagian besar variansi PDRB antarprovinsi dapat dijelaskan oleh keempat prediktor. RMSE dan MAE yang relatif rendah dibandingkan model lain menunjukkan kesalahan prediksi rata-rata yang lebih kecil. Gradient Boosting bekerja dengan membangun pohon-pohon keputusan secara sekuensial, di mana setiap pohon baru mengoreksi kesalahan pohon sebelumnya, sehingga efektif menangkap pola nonlinear dalam data [4]. Temuan ini sejalan dengan hasil pada penelitian sebelumnya yang membandingkan algoritma serupa untuk memprediksi indikator pembangunan dan ekonomi pada level wilayah maupun nasional, di mana model berbasis boosting secara konsisten mengungguli model regresi linear ketika hubungan antarvariabel bersifat nonlinear [5], [6].

2) Faktor Dominan yang Memengaruhi PDRB

Jumlah Penduduk merupakan faktor paling dominan dalam memprediksi PDRB, karena populasi besar menghasilkan angkatan kerja lebih banyak, pasar konsumsi lebih luas, dan skala ekonomi yang lebih besar. IPM berperan penting sebagai cerminan kualitas sumber daya manusia, di mana provinsi dengan IPM tinggi memiliki tenaga kerja lebih terdidik dan produktif [1]. Kemiskinan dan TPT berkontribusi negatif, mencerminkan rendahnya produktivitas dan daya beli masyarakat, sejalan dengan hubungan invers antara pengangguran dan output ekonomi yang juga ditemukan pada konteks Indonesia secara

nasional [7]. Model ensemble mengungguli model linear karena mampu menangkap hubungan nonlinear dan interaksi antarvariabel yang tidak terdeteksi oleh regresi linear.

V. KESIMPULAN DAN SARAN

Berdasarkan hasil analisis data sosial-ekonomi BPS Indonesia periode 2021–2025, ditemukan bahwa PDRB ADHB dan Jumlah Penduduk memiliki distribusi yang tidak merata, mencerminkan adanya disparitas ekonomi antarprovinsi. Jumlah Penduduk menjadi faktor yang paling berpengaruh terhadap PDRB, diikuti oleh Indeks Pembangunan Manusia (IPM), sedangkan Kemiskinan dan Tingkat Pengangguran Terbuka (TPT) menunjukkan hubungan negatif dengan PDRB. Dari lima model yang dibandingkan, yaitu Linear Regression, Ridge Regression, Decision Tree, Random Forest, dan Gradient Boosting, model Gradient Boosting Regressor memberikan performa terbaik berdasarkan nilai R^2 , RMSE, MAE, dan MAPE. Hasil 5-fold cross-validation menunjukkan bahwa model memiliki kemampuan generalisasi yang baik, sedangkan hyperparameter tuning menggunakan GridSearchCV berhasil meningkatkan performa model meskipun tidak secara signifikan.

Untuk penelitian selanjutnya, disarankan menambahkan variabel seperti investasi daerah, infrastruktur, rasio elektrifikasi, kontribusi sektoral, serta ekspor-impor daerah. Selain itu, periode data dapat diperpanjang menjadi 10–15 tahun, serta mempertimbangkan penggunaan model deret waktu seperti ARIMA, SARIMA, atau LSTM. Analisis spasial, transformasi logaritmik pada variabel target, dan pemodelan berdasarkan kluster ekonomi juga dapat diterapkan untuk meningkatkan akurasi prediksi dan mengurangi heterogenitas data antarprovinsi.

VI. DAFTAR PUSTAKA

- [1] E. Sofilda, M. Z. Hamzah, dan S. Kusairi, "Analysis of fiscal decentralisation, human development, and regional economic growth in Indonesia," *Cogent Economics & Finance*, vol. 11, no. 1, art. 2220520, 2023.
- [2] Badan Pusat Statistik, *Produk Domestik Regional Bruto Provinsi Atas Dasar Harga Berlaku Menurut Pengeluaran 2021–2025*. Jakarta: Badan Pusat Statistik Republik Indonesia, 2025.
- [3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, hal. 5–32, Okt. 2001.
- [4] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, hal. 1189–1232, Okt. 2001.
- [5] Y. J. Garcia-Lopez dan L. A. del Carpio Castro, "Assessing regional competitiveness in Peru: an approach using nonlinear machine learning models," *PLOS ONE*, vol. 20, no. 2, hal. e0318813, Feb. 2025.
- [6] C. V. Raghavendran, K. C. Mouli, S. B. Latha, dan A. U. Motagi, "A lightweight neural network approach for predicting national Gross Domestic Product (LightNet-GDP) with regression benchmarks," *Scientific Reports*, vol. 16, art. 6634, Jan. 2026.
- [7] C. Sutopo-Putri, "Empirical validity of the Okun's law: evidence from Indonesia," *International Review of Business and Economics*, vol. 9, no. 1, art. 1, hal. 1–9, Mar. 2024.
- [8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, dan E. Duchesnay, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, hal. 2825–2830, 2011.
- [9] W. McKinney, "Data structures for statistical computing in Python," dalam *Proc. 9th Python in*

- Science Conf. (SciPy 2010), Austin, TX, 2010, hal. 56–61.
- [10] B. Dastjerdy, A. Saeidi, dan S. Heidarzadeh, “Review of applicable outlier detection methods to treat geomechanical data,” *Geotechnics*, vol. 3, no. 2, hal. 375–396, Mei 2023.