

Prediksi Jumlah Kendaraan Bermotor Berbasis Data Provinsi Indonesia Menggunakan *Linear Regression* dan *Random Forest*

Prediction of Motor Vehicle Numbers Using Indonesian Provincial Data with Linear Regression and Random Forest

Dill Thafa Jausha¹, Rosa Nur Aliana Sawafi², Indy Aurellia AzZahra³, Diah Septiani⁴

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

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

²alinxzy@student.telkomuniversity.ac.id, ³indyaurelly@student.telkomuniversity.ac.id,

⁴diahseptiani@telkomuniversity.ac.id

Abstrak—Penelitian ini membahas pemodelan prediksi jumlah kendaraan bermotor di Indonesia berdasarkan variabel jumlah penduduk, kepadatan penduduk, dan panjang jalan. Tujuan penelitian ini adalah untuk membandingkan kinerja dua model regresi dalam memprediksi jumlah kendaraan serta menganalisis tren pertumbuhan kendaraan hingga tahun 2026. Metode yang digunakan meliputi penerapan model *Linear Regression* dan *Random Forest Regressor* dengan pembagian data latih dan data uji sebesar 80:20. Evaluasi kinerja model dilakukan menggunakan metrik MAE, RMSE, dan R^2 Score untuk mengukur tingkat kesalahan dan akurasi prediksi. Selain itu, dilakukan analisis visual untuk melihat kesesuaian antara nilai aktual dan nilai prediksi serta peramalan jumlah kendaraan pada tahun 2026. Hasil penelitian menunjukkan bahwa *Random Forest Regressor* memiliki performa yang lebih baik dibandingkan *Linear Regression* dengan nilai MAE sebesar 7.18×10^5 , RMSE sebesar 1.75×10^6 , dan R^2 sebesar 0.945. Model *Random Forest* juga mampu menjelaskan sekitar 94,5% variasi data dan menunjukkan pola prediksi yang lebih mendekati nilai aktual. Hasil prediksi tahun 2026 menunjukkan bahwa provinsi di Pulau Jawa masih mendominasi jumlah kendaraan tertinggi, dengan Jawa Barat sebagai peringkat pertama. Secara keseluruhan, model *Random Forest* dinilai lebih akurat dan stabil dalam memprediksi jumlah kendaraan bermotor di Indonesia.

Kata kunci-aplikasi; machine learning; model regresi; prediksi kendaraan; *Random Forest*

Abstract—This study examines the modeling of motor vehicle population prediction in Indonesia based on variables such as population size, population density, and road length. The objective of this research is to compare the performance of two regression models in predicting the number of vehicles and to analyze vehicle growth trends up to 2026. The methods applied include *Linear Regression* and *Random Forest Regressor* with an 80:20 train-test data split. Model performance was evaluated using MAE, RMSE, and R^2 Score to measure prediction error and accuracy. In addition, visual analysis was conducted to assess the relationship between actual and predicted values and to forecast vehicle numbers for 2026. The results show that the *Random Forest Regressor* outperforms *Linear Regression* with an MAE of 7.18×10^5 , RMSE of 1.75×10^6 , and an R^2 value of 0.945. The *Random Forest* model explains approximately 94.5% of the variance in the data and produces predictions that are closer to actual values. The 2026 prediction results indicate that provinces on Java Island still dominate the highest number of vehicles, with West Java ranking first. Overall, the *Random Forest* model demonstrates better accuracy and stability in predicting motor vehicle population in Indonesia.

Keywords—application; machine learning; regression model; vehicle prediction; Random Forest

I. PENDAHULUAN

Indonesia merupakan salah satu negara dengan populasi terbesar di dunia, dengan jumlah penduduk yang terus meningkat setiap tahunnya. Seiring dengan pertumbuhan penduduk dan perkembangan ekonomi, kebutuhan akan mobilitas masyarakat juga meningkat pesat, yang tercermin dari lonjakan jumlah kendaraan bermotor di seluruh wilayah nusantara. Berdasarkan data yang disajikan dalam berbagai penelitian, jumlah kendaraan bermotor di Indonesia terus bertambah secara signifikan dari tahun ke tahun [1]. Fenomena ini tidak hanya terjadi di pusat-pusat ekonomi seperti Jakarta, Bandung, dan Surabaya, tetapi juga merata di berbagai provinsi di Indonesia, dengan pertumbuhan yang sangat pesat dan cenderung tidak terkendali [2]. Di Kota Sorong, misalnya, pertumbuhan kendaraan bermotor roda dua dan roda empat menunjukkan peningkatan yang konsisten setiap tahunnya, sejalan dengan pertumbuhan penduduk dan peningkatan PDRB per kapita [2]. Sementara itu, di Kabupaten Sikka, Provinsi Nusa Tenggara Timur, peningkatan jumlah kendaraan juga terjadi secara signifikan seiring dengan bertambahnya jaringan jalan dan panjang jalan yang tersedia [3]. Data historis menunjukkan bahwa secara nasional, terjadi ketidakseimbangan antara pertumbuhan jumlah kendaraan yang sangat cepat dengan pertumbuhan infrastruktur jalan yang relatif lambat, sehingga berpotensi menimbulkan berbagai permasalahan transportasi seperti kemacetan, polusi udara, dan kecelakaan lalu lintas [1][3].

Peningkatan jumlah kendaraan bermotor tidak terjadi secara independen, melainkan dipengaruhi oleh berbagai faktor fundamental yang saling terkait. Penelitian menunjukkan bahwa jumlah penduduk memiliki pengaruh signifikan terhadap pertumbuhan kendaraan, di mana setiap peningkatan populasi akan mendorong kebutuhan mobilitas yang lebih tinggi [4] [2]. Studi di Kota Sorong membuktikan bahwa jumlah penduduk dan PDRB per kapita secara signifikan berpengaruh terhadap pertumbuhan kendaraan bermotor, dengan kontribusi yang sangat besar terhadap jumlah kendaraan roda dua dan roda empat [2]. Penelitian di tingkat nasional juga mengkonfirmasi bahwa jumlah penduduk berpengaruh positif dan signifikan terhadap konsumsi energi minyak, yang mencerminkan tingginya penggunaan kendaraan bermotor di Indonesia [4][5]. Di sisi lain, ketersediaan infrastruktur jalan atau panjang jalan memiliki hubungan yang erat dengan peningkatan jumlah kendaraan. Penelitian di Kabupaten Sikka menunjukkan bahwa konektivitas jaringan jalan yang meliputi jumlah jaringan jalan dan panjang jalan memiliki hubungan yang sangat kuat dengan peningkatan jumlah kendaraan, dengan nilai korelasi mencapai 0,978 [3]. Hal ini mengindikasikan bahwa semakin baik akses dan konektivitas jalan, semakin tinggi pula jumlah kendaraan yang beroperasi di wilayah tersebut. Namun, ketidakseimbangan antara pertumbuhan panjang jalan yang lambat dengan pertumbuhan kendaraan yang sangat cepat justru menciptakan masalah kapasitas jalan dan tingkat pelayanan yang buruk [1][2]. Semua faktor ini menunjukkan bahwa pertumbuhan kendaraan adalah fenomena multidimensi yang memerlukan pemahaman mendalam untuk dapat diprediksi dan dikelola dengan tepat.

Dalam upaya memahami dan mengantisipasi dinamika pertumbuhan kendaraan, pendekatan prediktif menggunakan *machine learning* menjadi sangat penting. Teknologi ini memungkinkan analisis pola dari data historis untuk menghasilkan proyeksi masa depan yang lebih akurat, yang sangat dibutuhkan dalam perencanaan infrastruktur, pengaturan lalu

lintas, dan perumusan kebijakan lingkungan [6][7]. Penelitian di berbagai bidang telah menunjukkan efektivitas metode *machine learning* dalam memprediksi berbagai fenomena kompleks, termasuk di bidang transportasi [6][7]. Dalam konteks prediksi jumlah kendaraan bermotor di Indonesia, penggunaan metode *machine learning* masih relatif terbatas [7][8]. Penelitian yang ada umumnya masih menggunakan pendekatan statistik tradisional seperti analisis regresi linear berganda dan metode Monte Carlo [2][8]. Misalnya, penelitian di Kota Sorong menggunakan analisis regresi linear berganda untuk memodelkan pertumbuhan kendaraan bermotor berdasarkan faktor jumlah penduduk dan PDRB per kapita [2]. Sementara itu, penelitian di Pulau Sumatra menggunakan metode Monte Carlo untuk memprediksi jumlah kendaraan bermotor dengan tingkat akurasi antara 84% hingga 89% [8]. Namun, kedua pendekatan tersebut memiliki keterbatasan dalam menangani hubungan non-linear dan interaksi antar variabel yang kompleks, sehingga diperlukan pengembangan model prediktif yang lebih adaptif dan akurat.

Di antara berbagai algoritma *machine learning*, *Linear Regression* dan *Random Forest* menawarkan pendekatan yang berbeda namun saling melengkapi untuk tugas prediksi. *Linear Regression* adalah metode statistik klasik yang sederhana, mudah diinterpretasikan, dan efektif untuk data dengan hubungan linear antara variabel dependen dan independen, seperti yang digunakan dalam penelitian di Kota Sorong untuk memodelkan pengaruh jumlah penduduk dan PDRB terhadap jumlah kendaraan [2]. Namun, metode ini memiliki kelemahan ketika dihadapkan pada hubungan non-linear atau data dengan interaksi variabel yang kompleks [6]. Di sisi lain, *Random Forest* adalah algoritma berbasis pohon keputusan yang termasuk dalam metode ensemble learning, yang mampu menangani hubungan non-linear, robust terhadap overfitting, dan memiliki akurasi tinggi dalam berbagai studi klasifikasi dan regresi [9][10]. Penelitian dalam bidang kesehatan dan perbankan telah membuktikan keunggulan *Random Forest* dibandingkan algoritma lain seperti Support Vector Machine dan Naive Bayes, dengan tingkat akurasi yang sangat tinggi mencapai 97,88% untuk prediksi diabetes dan 83% untuk klasifikasi kelayakan kredit [9][10]. Dalam konteks prediksi jumlah kendaraan bermotor, *Random Forest* dapat mengakomodasi kompleksitas interaksi antara variabel demografi, infrastruktur, dan ekonomi yang tidak selalu linear, sehingga berpotensi menghasilkan prediksi yang lebih akurat dibandingkan *Linear Regression* [6][7]. Namun, penelitian tentang penerapan kedua algoritma ini secara komparatif untuk prediksi jumlah kendaraan bermotor di Indonesia masih sangat terbatas.

Meskipun telah banyak penelitian tentang prediksi kendaraan bermotor, sebagian besar masih terbatas pada pendekatan statistik tradisional atau penggunaan satu jenis algoritma *machine learning* di wilayah spesifik. Penelitian tentang faktor-faktor yang mempengaruhi kepemilikan dan pertumbuhan kendaraan di Indonesia telah dilakukan di berbagai kota seperti Jakarta, Bandung, dan Surabaya [1], serta di Kabupaten Sikka [3] dan Kota Sorong [2]. Namun, studi yang secara khusus membandingkan kinerja algoritma *Linear Regression* dan *Random Forest* dalam konteks prediksi jumlah kendaraan bermotor di tingkat nasional dengan mempertimbangkan variabel jumlah penduduk, kepadatan penduduk, dan panjang jalan masih belum banyak dilakukan. Gap analisis yang diidentifikasi adalah kurangnya studi komprehensif yang membandingkan kedua metode ini secara langsung pada data nasional yang mencakup periode waktu yang panjang dan beragam wilayah, serta belum optimalnya pemanfaatan teknik *machine learning* untuk prediksi jumlah kendaraan di Indonesia. Dengan demikian, penelitian ini bertujuan untuk mengembangkan model prediksi jumlah kendaraan

bermotor di Indonesia dengan membandingkan kinerja algoritma *Linear Regression* dan *Random Forest*, serta menganalisis faktor-faktor dominan yang memengaruhinya. Kebaruan penelitian ini terletak pada penggunaan data nasional yang lebih komprehensif dan komparasi sistematis antara dua pendekatan *machine learning* yang berbeda, yang diharapkan dapat memberikan kontribusi bagi pemerintah dan pemangku kepentingan dalam perencanaan infrastruktur transportasi, kebijakan pengendalian kendaraan, dan pengembangan kota yang berkelanjutan.

II. TINJAUAN TEORI

2.1 Prediksi Kendaraan Bermotor

Prediksi jumlah kendaraan bermotor merupakan bagian dari pemodelan transportasi yang bertujuan memperkirakan kebutuhan infrastruktur dan kebijakan transportasi. Penelitian sebelumnya menemukan bahwa jumlah penduduk dan PDRB per kapita berpengaruh signifikan secara statistik terhadap pertumbuhan kendaraan bermotor di Kota Sorong[2]]. Ini menunjukkan bahwa konektivitas jaringan jalan berkorelasi positif dengan jumlah kendaraan terdaftar, karena peningkatan aksesibilitas mendorong kepemilikan kendaraan pribadi

2.2 Linear Regression

Regresi linear mengasumsikan hubungan linear antara variabel independen dan dependen, meminimalkan jumlah kuadrat residual (OLS) untuk mencari koefisien terbaik. Penelitian sebelumnya menggunakan regresi linear untuk prediksi penjualan kendaraan di industri otomotif dan memperoleh prediksi yang cukup akurat sebagai *baseline* [6]. Model linear mudah diinterpretasikan namun cenderung *underestimate* pada wilayah dengan nilai ekstrem tinggi akibat keterbatasan dalam menangkap hubungan non-linear.

2.3 Random Forest

Random Forest adalah metode ensemble yang membangun sejumlah pohon keputusan secara independen pada subhimpunan data latih (*bootstrap sampling*), kemudian mengagregasi prediksi seluruh pohon sebagai rata-rata. Mekanisme ini efektif mengurangi *overfitting* yang umum pada pohon tunggal sekaligus menangkap hubungan non-linear dan interaksi antar fitur [9].

III. METODE

3.1 Dataset

Data bersumber dari publikasi resmi BPS Republik Indonesia, mencakup tiga dataset periode 2019–2025. variabel penelitian disajikan pada Tabel 1.

Tabel 1. Karakteristik Dataset

Variabel	Deskripsi	Satuan	Peran
Tahun	Tahun observasi	2019–2025	Fitur (X)
Jml. Penduduk	Jumlah penduduk	Ribu jiwa	Fitur (X)
Kepadatan	Kepadatan/km ²	Jiwa/km ²	Fitur (X)
Panjang Jalan	Total jalan	Km	Fitur (X)

Variabel	Deskripsi	Satuan	Peran
Jml. Kendaraan	Total kendaraan	Unit	Target (Y)

3.2 Data Cleaning

Proses praproses data diawali dengan seleksi kolom, di mana hanya variabel yang relevan dengan tujuan penelitian dipertahankan pada masing-masing dataset. Selanjutnya dilakukan standardisasi dan pembersihan data, meliputi penyamaan format nama provinsi, konversi tahun menjadi tipe numerik, serta penghapusan karakter non-numerik agar seluruh variabel dapat dikonversi ke bentuk numerik *float*.

Ketiga dataset kemudian digabungkan menggunakan metode inner join berdasarkan provinsi dan tahun sehingga menghasilkan satu dataset terpadu. Pada tahap berikutnya, dilakukan penanganan data hilang menggunakan imputasi median dengan SimpleImputer karena lebih robust terhadap *outlier*. Selain itu, data agregat nasional seperti "INDONESIA" dihapus, serta baris dengan nilai target yang tidak lengkap juga dikeluarkan untuk menjaga konsistensi data dalam proses pemodelan.

3.3 Exploratory Data Analysis (EDA)

Eksplorasi data dilakukan untuk memahami karakteristik distribusi, pola, serta hubungan antar variabel dalam dataset. Analisis dilakukan menggunakan *histogram*, *boxplot*, *heatmap* korelasi, serta analisis tren tahunan dan peringkat provinsi.

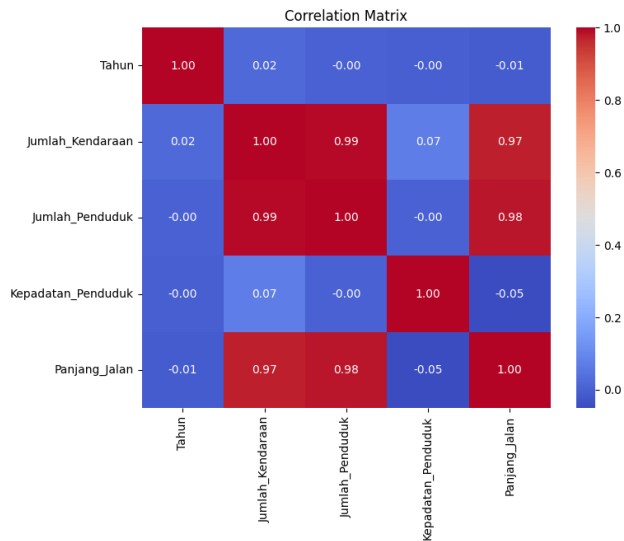
3.3.1 Boxplot Variabel

Analisis *boxplot* menunjukkan bahwa variabel jumlah kendaraan memiliki sebaran yang sangat lebar dengan beberapa nilai pencilan yang mendekati 10 juta. Hal ini menunjukkan adanya perbedaan yang signifikan antarwilayah. Sebagian besar data terkonsentrasi pada nilai yang lebih rendah dengan variasi yang tinggi.

Pada variabel jumlah penduduk, distribusi data relatif lebih stabil dengan median berada pada kisaran menengah. Namun demikian, terdapat beberapa pencilan pada nilai yang lebih tinggi yang menunjukkan adanya wilayah dengan populasi sangat besar.

Variabel kepadatan penduduk menunjukkan sebaran yang cukup bervariasi antarwilayah, dengan beberapa wilayah memiliki kepadatan yang jauh lebih tinggi dibandingkan wilayah lainnya. Sementara itu, variabel panjang jalan memiliki distribusi yang relatif moderat dengan beberapa nilai ekstrem pada bagian bawah dan atas, yang menunjukkan variasi kondisi infrastruktur antarprovinsi.

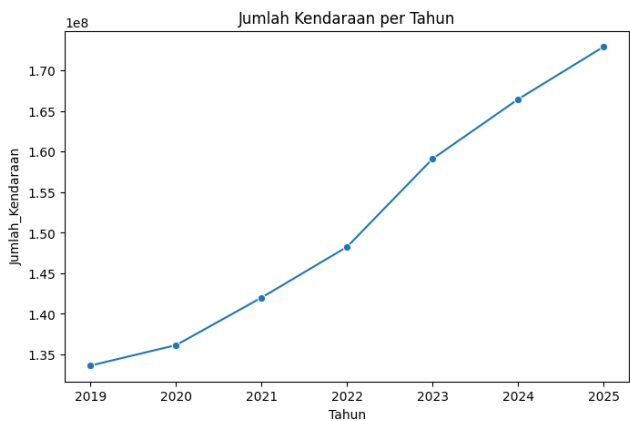
3.3.2 Matrix Korelasi



Gambar 1. Correlation Matrix

Hasil analisis korelasi menunjukkan bahwa terdapat hubungan yang sangat kuat antara jumlah kendaraan dengan jumlah penduduk (0.99). Hal ini menunjukkan bahwa peningkatan jumlah penduduk cenderung diikuti oleh peningkatan jumlah kendaraan. Selain itu, terdapat korelasi kuat antara jumlah kendaraan dengan panjang jalan (0.97), yang mengindikasikan bahwa infrastruktur jalan berhubungan erat dengan jumlah kendaraan.

3.3.3 Tren Jumlah Kendaraan (2019-2025)



Gambar 2. Trend Jumlah Kendaraan

Hasil analisis menunjukkan bahwa jumlah kendaraan mengalami peningkatan yang konsisten setiap tahun. Pada tahun 2019 jumlah kendaraan tercatat sebesar 1.33 dan meningkat menjadi 1.74 pada tahun 2025. Tren ini menunjukkan pertumbuhan yang stabil dengan rata-rata kenaikan tahunan sekitar 0.05 hingga 0.08.

Peningkatan ini mencerminkan adanya pertumbuhan mobilitas masyarakat, perkembangan ekonomi, serta urbanisasi yang terus berlangsung. Jika tren ini berlanjut, maka jumlah kendaraan diperkirakan akan terus meningkat di masa mendatang.

3.4 Preprocessing

Tahap praproses data dalam penelitian ini mencakup standarisasi fitur serta pembagian data menjadi data latih dan data uji. Pada model *Linear Regression*, fitur terlebih dahulu distandarisasi menggunakan *StandardScaler*, agar setiap variabel memiliki skala yang seragam. Hal ini diperlukan karena regresi linear sensitif terhadap perbedaan skala antarfitur. Sebaliknya, pada model *Random Forest* tidak dilakukan standarisasi karena algoritma berbasis pohon keputusan tidak bergantung pada skala data.

Selanjutnya, dataset dibagi menjadi data latih dan data uji dengan rasio 80:20 menggunakan parameter *test_size=0.2* dan *random_state=42* untuk memastikan hasil yang dapat direproduksi. Selain pembagian tersebut, dilakukan pula validasi silang (*cross-validation*) dengan *5-fold* untuk memperoleh estimasi performa model yang lebih stabil dan dapat digeneralisasi. Konsistensi pembagian data ini juga bertujuan untuk memberikan perbandingan yang adil antara model yang digunakan dalam penelitian.

3.5 Modeling

Metode yang digunakan dalam penelitian ini terdiri dari dua model regresi, yaitu *Linear Regression* dan *Random Forest Regressor*. *Linear Regression* merupakan model statistik dasar untuk permasalahan regresi yang mengasumsikan adanya hubungan linear antara variabel independen (fitur) dan variabel dependen (target). Model ini bekerja dengan meminimalkan jumlah kuadrat residual melalui metode *Ordinary Least Squares* (OLS) untuk memperoleh koefisien terbaik. Meskipun sederhana, model ini sering digunakan sebagai baseline karena dapat memberikan performa yang cukup baik ketika hubungan antar variabel bersifat linear. Dalam penelitian ini, *Linear Regression* diimplementasikan menggunakan kelas *LinearRegression* dari pustaka *scikit-learn* dengan parameter default.

Sementara itu, *Random Forest Regressor* merupakan metode ensemble berbasis kumpulan pohon keputusan yang dibangun secara independen menggunakan teknik *bootstrap sampling*. Setiap pohon menghasilkan prediksi, kemudian hasil akhir diperoleh dari rata-rata seluruh prediksi pohon tersebut. Pendekatan ini mampu mengurangi risiko *overfitting* yang umum terjadi pada pohon keputusan tunggal, serta efektif dalam menangkap hubungan non-linear dan interaksi antar fitur. Model ini menggunakan parameter *n_estimators=200*, *max_depth=10*, dan *random_state=42* untuk menjaga keseimbangan antara performa dan kompleksitas model serta memastikan hasil yang dapat direproduksi. Berbeda dengan *Linear Regression*, *Random Forest* tidak memerlukan proses standarisasi fitur karena bersifat tidak sensitif terhadap skala data (*scale-invariant*).

3.6 Evaluasi Model

Kinerja kedua model dievaluasi menggunakan tiga metrik regresi, yaitu *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), dan koefisien determinasi (R^2). MAE digunakan untuk mengukur rata-rata kesalahan absolut antara nilai prediksi dan aktual, RMSE memberikan penalti lebih besar pada kesalahan yang tinggi sehingga sensitif terhadap *outlier*, sedangkan R^2 menunjukkan seberapa besar variansi data yang dapat dijelaskan oleh model.

Selain evaluasi pada data uji, dilakukan juga validasi silang (*5-fold cross-validation*) untuk memperoleh estimasi performa yang lebih stabil dan tidak bergantung pada satu pembagian data, sehingga hasil evaluasi lebih andal dan representatif.

Bab ini berisi informasi teknis dari penelitian yang disajikan secara jelas, baik berupa

metodologi, spesifikasi material dan peralatan yang diperlukan, pendekatan, prosedur hingga metode analisis data yang digunakan.

3.7 Prediksi

Untuk menghasilkan prediksi jumlah kendaraan tahun 2026, nilai-nilai fitur prediktor (Jumlah_Penduduk, Kepadatan_Penduduk, Panjang_Jalan) untuk masing-masing provinsi pada tahun 2026 terlebih dahulu diestimasi menggunakan model regresi linear *time series* per provinsi. Untuk setiap provinsi, model regresi linear dilatih dengan Tahun sebagai variabel independen dan masing-masing fitur sebagai variabel dependen, kemudian digunakan untuk mengekstrapolasi nilai fitur pada tahun 2026.

IV. HASIL DAN PEMBAHASAN

4.1 Hasil Evaluasi

Tabel 2 merangkum hasil evaluasi kinerja kedua model pada data uji (20% dari keseluruhan data):

Tabel 2. Hasil Evaluasi Model

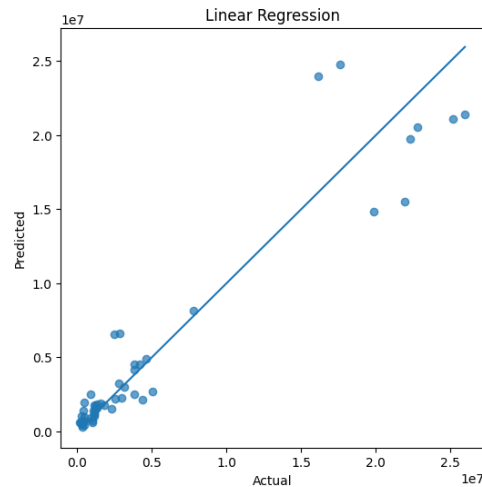
Model	MAE	RMSE	R2 Score
<i>Linear Regression</i>	1.409131e+06	2.403468e+06	0.897283
<i>Random Forest</i>	7.181727e+05	1.751000e+06	0.945482

Hasil evaluasi menunjukkan bahwa model *Random Forest* memiliki performa yang cukup baik dalam melakukan prediksi jumlah kendaraan. Nilai MAE sebesar 7.18×10^5 (≈ 718.173 kendaraan) menunjukkan bahwa rata-rata kesalahan prediksi model berada pada kisaran ratusan ribu kendaraan dari nilai aktual. Sementara itu, RMSE sebesar 1.75×10^6 ($\approx 1.751.000$ kendaraan) menunjukkan bahwa terdapat beberapa kesalahan prediksi yang cukup besar, karena RMSE lebih sensitif terhadap *outlier* dan menghukum error yang besar secara lebih signifikan.

Selain itu, nilai R^2 sebesar 0.945 menunjukkan bahwa model mampu menjelaskan sekitar 94,5% variasi data jumlah kendaraan, yang mengindikasikan kemampuan prediksi yang sangat baik. Secara keseluruhan, meskipun masih terdapat deviasi pada beberapa kasus ekstrem, model *Random Forest* menunjukkan kinerja yang kuat dan mampu merepresentasikan pola hubungan antar variabel dengan baik.

4.2 Perbandingan Nilai Aktual *Linear Regression*

Perbandingan nilai aktual dan prediksi dapat dilihat lewat Gambar 3.

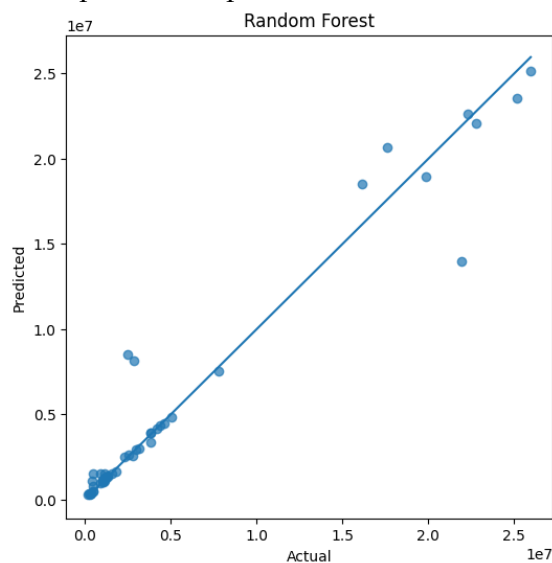


Gambar 3. Visualisasi Nilai Aktual dan Prediksi *Linear Regression*

Scatter plot menunjukkan hubungan antara nilai aktual dan nilai prediksi dari model *Linear Regression*. Pada rentang nilai aktual rendah hingga sekitar 5×10^6 , titik prediksi cenderung mengikuti pola garis diagonal meskipun masih terdapat penyebaran. Pada rentang nilai menengah sekitar $1,5 \times 10^7$ hingga $1,8 \times 10^7$, terlihat adanya beberapa titik yang menunjukkan nilai prediksi lebih tinggi dibandingkan nilai aktual, yang mengindikasikan terjadinya overestimasi pada sebagian data. Sementara itu, pada rentang nilai aktual yang lebih tinggi sekitar 2×10^7 ke atas, nilai prediksi masih mendekati nilai aktual namun dengan variasi yang lebih besar. Secara umum, sebaran titik menunjukkan adanya penyimpangan dari garis ideal yang mencerminkan adanya kesalahan prediksi pada beberapa observasi.

4.3 Perbandingan Nilai Aktual *Random Forest*

Perbandingan nilai aktual dan prediksi dapat dilihat lewat Gambar 4.



Gambar 4. Visualisasi Nilai Aktual dan Prediksi *Random Forest*

Scatter plot ini menunjukkan hubungan antara nilai aktual (sumbu X) dan nilai prediksi (sumbu Y) dari model *Random Forest*. Secara umum, titik-titik mengikuti garis diagonal yang menunjukkan bahwa model mampu menangkap pola data dengan baik.

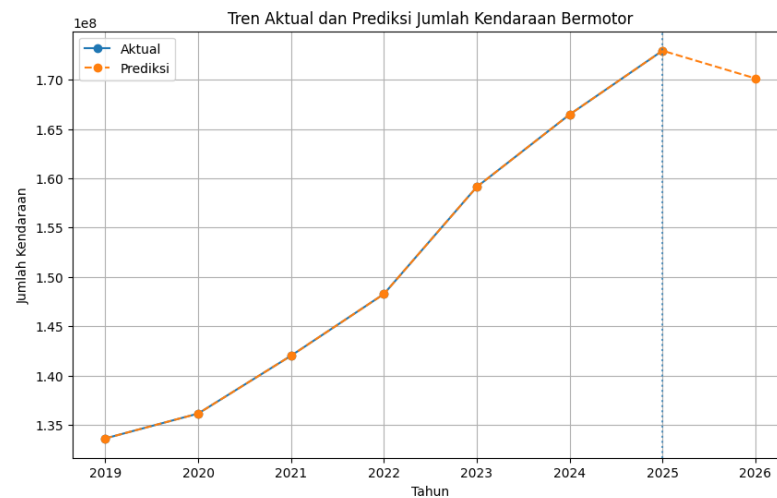
Pada nilai aktual rendah hingga sekitar 5×10^6 , titik-titik berada dekat dengan garis diagonal yang menandakan kesalahan prediksi relatif kecil. Pada rentang menengah sekitar $1,5 \times 10^7$ hingga 2×10^7 , terdapat beberapa titik di atas garis diagonal yang menunjukkan kecenderungan overestimasi. Sementara pada nilai aktual tinggi sekitar $2,2 \times 10^7$ ke atas, terdapat satu titik yang berada jauh di bawah garis diagonal yang menunjukkan underestimasi pada nilai yang sangat besar. Secara keseluruhan, pola sebaran masih mengikuti tren yang konsisten dengan garis ideal.

4.4 Hasil Prediksi 10 Provinsi Terbanyak 2026

Peringkat	Provinsi	Prediksi Kendaraan
1	JAWA BARAT	26.517.209
2	JAWA TIMUR	25.582.910
3	JAWA TENGAH	22.633.533
4	DKI JAKARTA	12.325.388
5	BANTEN	8.758.554
6	SUMATERA UTARA	8.219.752
7	SULAWESI SELATAN	5.667.216
8	BALI	5.021.785
9	LAMPUNG	4.850.863
10	RIAU	4.763.744

Hasil prediksi jumlah kendaraan tahun 2026 menunjukkan bahwa nilai estimasi telah diurutkan berdasarkan nilai tertinggi hingga terendah untuk memudahkan analisis peringkat antarprovinsi. Berdasarkan hasil tersebut, Jawa Barat menempati posisi pertama dengan prediksi sebesar 26.517.209 unit kendaraan, diikuti oleh Jawa Timur dengan 25.582.910 unit dan Jawa Tengah dengan 22.633.533 unit. Selanjutnya, DKI Jakarta berada pada posisi keempat dengan 12.325.388 unit kendaraan, kemudian Banten dan Sumatera Utara masing-masing sebesar 8.758.554 unit dan 8.219.752 unit. Provinsi lainnya yang termasuk dalam 10 besar adalah Sulawesi Selatan, Bali, Lampung, dan Riau dengan nilai prediksi yang berada pada kisaran 4 hingga 5 juta unit kendaraan. Hasil ini merupakan keluaran model *Random Forest* yang telah diurutkan untuk menunjukkan peringkat prediksi tertinggi pada tahun 2026 di masing-masing provinsi.

4.5 Tren Aktual dan Prediksi Jumlah Kendaraan Bermotor



Gambar 5. Tren Aktual dan Prediksi Jumlah Kendaraan Bermotor

Terlihat bahwa jumlah kendaraan di Indonesia mengalami peningkatan yang konsisten dari tahun 2019 hingga 2025 dengan pola pertumbuhan yang stabil dan cenderung semakin meningkat pada periode 2021 hingga 2024, yang menunjukkan adanya percepatan pertumbuhan jumlah kendaraan. Peningkatan paling signifikan terlihat pada rentang 2022 hingga 2024, yang mencerminkan tingginya pertumbuhan mobilitas dan kepemilikan kendaraan di Indonesia. Pada tahun 2025, jumlah kendaraan mencapai titik tertinggi pada data aktual sebelum memasuki tahap prediksi. Sementara itu, hasil prediksi tahun 2026 menunjukkan adanya sedikit penurunan dibandingkan tahun 2025, namun tetap berada pada level yang tinggi sehingga tidak mengubah pola umum tren yang cenderung meningkat dan stabil. Secara keseluruhan, grafik menunjukkan bahwa model mampu mengikuti pola historis dengan baik, meskipun terdapat sedikit deviasi pada tahun prediksi yang dapat dipengaruhi oleh keterbatasan model dalam menangkap seluruh dinamika pertumbuhan secara kompleks.

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa model *machine learning* khususnya *Random Forest Regressor* mampu digunakan untuk memprediksi jumlah kendaraan bermotor di Indonesia berdasarkan variabel jumlah penduduk, kepadatan penduduk, dan panjang jalan. Hasil evaluasi menunjukkan bahwa *Random Forest* memiliki kinerja yang lebih baik dibandingkan *Linear Regression* dengan nilai MAE sebesar 7.18×10^5 , RMSE sebesar 1.75×10^6 , dan R^2 sebesar 0.945, yang menunjukkan kemampuan model dalam menjelaskan variasi data secara cukup tinggi.

Hasil analisis juga menunjukkan bahwa prediksi jumlah kendaraan tahun 2026 masih mengikuti pola distribusi sebelumnya, dengan konsentrasi tertinggi berada pada provinsi-provinsi di Pulau Jawa seperti Jawa Barat, Jawa Timur, dan Jawa Tengah. Selain itu, hasil tren menunjukkan peningkatan jumlah kendaraan dari tahun 2019 hingga 2025, meskipun pada tahun 2026 terdapat sedikit penurunan berdasarkan hasil prediksi model. Secara keseluruhan, model yang digunakan mampu menangkap pola hubungan antarvariabel dengan baik dan menghasilkan prediksi yang cukup akurat.

5.2 Saran

Untuk penelitian selanjutnya, disarankan agar penggunaan variabel tambahan seperti tingkat urbanisasi, pertumbuhan ekonomi, atau jumlah pengguna transportasi umum dapat dipertimbangkan untuk meningkatkan akurasi model. Selain itu, penggunaan metode *time series forecasting* yang lebih kompleks seperti ARIMA atau LSTM dapat menjadi alternatif untuk menangkap pola temporal secara lebih mendalam.

Dari sisi pemodelan, eksplorasi algoritma lain seperti *Gradient Boosting* atau XGBoost juga dapat dilakukan untuk dibandingkan dengan *Random Forest* guna memperoleh performa yang lebih optimal. Selain itu, peningkatan kualitas data, seperti penanganan *outlier* yang lebih adaptif atau penambahan data historis yang lebih panjang, juga dapat membantu meningkatkan stabilitas dan generalisasi model.

VI. DAFTAR PUSTAKA

- [1] I. Sefriyadi, I. G. A. Andani, A. Raditya, P. F. Belgiawan, and N. A. Windasari, "Private car ownership in Indonesia: Affecting factors and policy strategies," *Transp. Res. Interdiscip. Perspect.*, vol. 19, May 2023, doi: 10.1016/j.trip.2023.100796.
- [2] Johaness Eudes Ola, "PEMODELAN PERTUMBUHAN KENDARAAN DI KOTA SORONG," *Jurnal Karkasa*, vol. 7, no. 1, Jul. 2021, doi: <https://doi.org/10.32531/jkar.v7i1.248>.
- [3] Gado Alfridus, Tan Valentinus, and Gare Masuetus, "Analisis Pengaruh Konektivitas Jaringan Jalan Terhadap Peningkatan Jumlah Kendaraan Di Kabupaten Sikka," *TEKNOSIAR*, vol. Volume 18, Sep. 2024.
- [4] A. Rahmawati, "PENGARUH JUMLAH PENDUDUK, JUMLAH KENDARAAN BERMOTOR, PDRB PER KAPITA DAN KEBIJAKAN FISKAL TERHADAP KONSUMSI ENERGI MINYAK DI INDONESIA," *Jurnal Pembangunan dan Pemerataan (JPP)*, vol. 10, Jun. 2021.
- [5] R. Bhaskara Praja and S. Wijaya, "THE EFFECT OF POPULATION SIZE AND NUMBER OF MOTOR VEHICLES ON LOCAL TAX REVENUE WITH GRDP AS A MODERATING VARIABLE," *EDUCORETAX*, vol. 4, no. 1, 2024.
- [6] A. Ali, A. Kalatian, and C. F. Choudhury, "Comparing and contrasting choice model and machine learning techniques in the context of vehicle ownership decisions," *Transp. Res. Part A Policy Pract.*, vol. 173, Jul. 2023, doi: 10.1016/j.tra.2023.103727.
- [7] A. Darmawan Sidik, A. Ansawarman, K. Kunci, J. Kendaraan Bermotor, M. Regresi, and F. Jalan, "Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning," *Formosa Journal of Multidisciplinary Research (FJMR)*, vol. 1, no. 3, pp. 559–568, 2022, doi: 10.55927.
- [8] Y. Mufida, I. Muslim, and K. Karo, "PREDIKSI JUMLAH KENDARAAN BERMOTOR DI PULAU SUMATRA MENGGUNAKAN METODE MONTE CARLO," 2024.
- [9] B. Prasajo and E. Haryatmi, "Analisa Prediksi Kelayakan Pemberian Kredit Pinjaman dengan Metode Random Forest," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 2, pp. 79–89, Sep. 2021, doi: 10.25077/teknosi.v7i2.2021.79-89.
- [10] W. Apriliah *et al.*, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *SISTEMASI: Jurnal Sistem Informasi*, vol. 10, no. 1, pp. 2540–9719, Jan. 2021, [Online]. Available: <http://sistemasi.ftik.unisi.ac.id>