Seminar Nasional Sains Data 2023 (SENADA 2023)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Model Prediksi Kepadatan Lalu Lintas: Perbandingan Antara Algoritma Random Forest dan XGBoost

Fitri Indah Sari[1], Ellexia Leonie Gunawan[2], Chelsea Ayu Adhigiadany[3] , Angela Lisanthoni[4]
[1,2,3,4] *Sains Data, UPN "Veteran" Jawa Timur*
[1] 21083010025@student.upnjatim.ac.id
[2] 21083010027@student.upnjatim.ac.id
[3] 21083010028@student.upnjatim.ac.id
[4] 21083010032@student.upnjatim.ac.id
Corresponding author email: 21083010032@student.upnjatim.ac.id

Abstract: *Population in general will continue to grow so traffic congestion is a phenomenon that needs to be overcome. Therefore, it is necessary to predict accurate traffic density to develop the right strategy to reduce the level of congestion. So, the aim of this study is making a comparison between the Random Forest and XGBoost algorithms which are one of the forecasting models forming algorithms. The results will be compared based on three types of evaluation parameters including the coefficient of determination* (R^2)*, Mean Absolute Error* (MAE), dan *Root Mean Square Error* (RMSE). *Based on the research, it proves that there is no significant difference in predicting traffic volume between the Random Forest and XGBoost algorithms. The two algorithms being compared have an accuracy of approximately 95%, however, XGBoost has the advantage that the time required for prediction is 532% faster than Random Forest. By considering accuracy and efficiency, the XGBoost algorithm is the right choice in building traffic forecasting models.*
Keywords: *Regression Supervised Learning, Traffic Forecasting, Random Forest, XGBoost*

Abstrak: Populasi manusia secara umum akan terus bertambah sehingga kepadatan lalu lintas adalah fenomena yang perlu diatasi. Oleh sebab itu, diperlukan prediksi kepadatan lalu lintas yang akurat untuk menyusun strategi yang tepat guna mengurangi tingkat kemacetan. Sehingga tujuan dari penelitian ini adalah membandingkan algoritma *Random Forest* dan *XGBoost* yang merupakan salah satu algoritma pembentuk model prediksi. Hasil akan dibandingkan berdasarkan tiga jenis evaluasi parameter diantaranya koefisien determinasi (R^2), *Mean Absolute Error* (MAE), dan *Root Mean Square Error* (RMSE). Berdasarkan penelitian yang dilakukan, menunjukkan bahwa tidak ada perbedaan signifikan dalam melakukan prediksi volume lalu lintas antara algoritma *Random Forest* dan *XGBoost*. Kedua algoritma yang dibandingkan memiliki hasil akurasi kurang lebih 95% namun, XGBoost memiliki kelebihan yakni waktu yang dibutuhkan dalam prediksi 532% lebih cepat dibanding *Random Forest*. Dengan mempertimbangkan akurasi dan efisiensi, algoritma *XGBoost* adalah pilihan yang tepat dalam membangun model prediksi kepadatan lalu lintas.
Kata kunci: Regresi Supervised Learning, Prediksi Kepadatan lalu lintas, Random Forest, XGBoost

I. PENDAHULUAN

Kepadatan lalu lintas sudah menjadi permasalahan yang semakin umum di masa sekarang. Berdasarkan *The 2022 Revision of World Population Prospects* tercatat sebesar 8 miliar jiwa pada 15 November 2022 [1]. Dari data *2023 World Population by Country (Live)* tercatat bahwa populasi Indonesia pada 2022 mencapai 275 juta jiwa dengan luas daratan 1,9 juta km^2, itu berarti populasi masyarakat Indonesia mencapai 135 juta jiwa/km^2 [2].

Kepadatan lalu lintas tentunya didasari oleh beberapa aspek salah satunya adalah cuaca [3]. Dengan berbagai permasalahan tersebut, adanya prediksi tentang kepadatan lalu lintas menjadi sangat penting. Dengan adanya prediksi kepadatan lalu lintas diharapkan dapat membantu dalam menghindari bahkan mengatasi peristiwa kemacetan. Prediksi yang akurat akan memberikan hasil yang bagus dan juga bermanfaat bagi pengguna jalan. Cara untuk memprediksi dengan nilai akurasi yang bagus tentunya menggunakan teknik machine learning seperti regresi linier, *random forest*, dan *XGBoost* [4]. Pada penelitian sebelumnya, untuk memprediksi kepadatan lalu lintas menggunakan metode *random forest algorithm* dibuktikan bahwa *random forest algorithm* merupakan algoritma terbaik untuk memprediksi lalu lintas berdasarkan data [5]. Kemudian metode-metode tersebut akan dievaluasi kinerjanya menggunakan matrik evaluasi seperti *Root Mean Square Error* (RMSE), koefisien determinasi (R^2), dan juga *Mean Absolute Error* (MAE). Dari hasil evaluasi ini nantinya

akan membantu menentukan metode manakah yang paling efektif dalam memprediksi kepadatan lalu lintas.

Tujuan daripada pembuatan artikel ini yaitu untuk mencari model terbaik dalam melakukan *traffic forecasting* (prediksi kepadatan lalu lintas). Dalam perencanaan transportasi, mengetahui informasi arus lalu lintas sangat krusial. Oleh sebab itu, dibutuhkan model yang mampu memproyeksikan dengan tepat dan akurat volume lalu lintas di masa mendatang. Dalam artikel ini, akan dibandingkan dua model yaitu *random forest* dan *XGBoost* dalam rangka menemukan model optimal yang dapat digunakan dalam *traffic forecasting*. Kedua model ini telah banyak digunakan dalam berbagai aplikasi *machine learning* dan memiliki kelebihan masing-masing. Dengan membandingkan kedua model ini, diharapkan dapat ditemukan model terbaik yang dapat digunakan dalam perencanaan transportasi untuk memprediksi volume lalu lintas di masa depan.

II. KAJIAN PUSTAKA

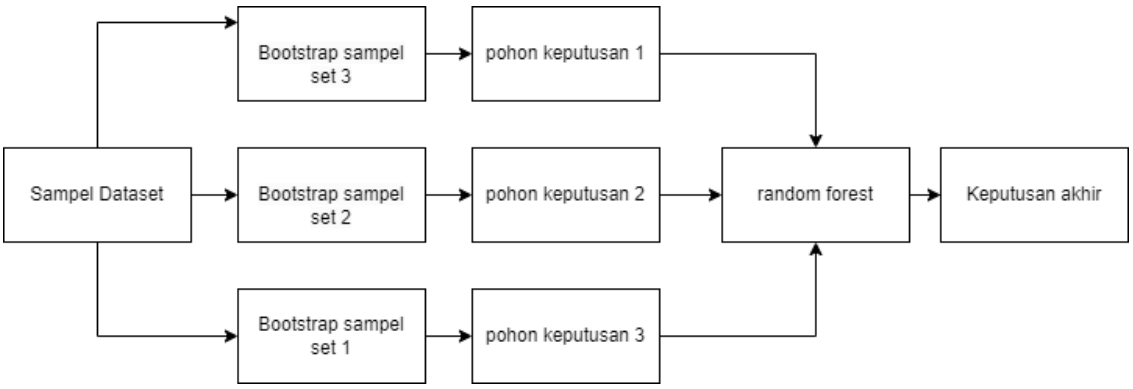
2.1. Traffic Forecasting

Traffic forecasting adalah salah satu jenis model *supervised machine learning* yang memiliki tujuan utama memprediksi kepadatan lalu lintas pada waktu dekat di masa depan berdasarkan data kepadatan lalu lintas sekarang dan data di masa lampau sehingga model akan mempelajari pola data input [6]. Prediksi *traffic forecasting* yang akurat sangat membantu pengambilan keputusan di masa depan agar dapat merencanakan strategi dengan tepat dan lebih efisien dalam mengurangi kepadatan lalu lintas, meningkatkan kualitas udara, serta mobilitas [7].

Traffic forecasting akan memprediksi volume kemacetan dimana volume kemacetan yang dimaksud adalah jumlah kendaraan yang melintas pada jalan yang diobservasi. Volume kemacetan dipengaruhi oleh beberapa faktor diantaranya cuaca dan hari libur. ketika hari libur dan suhu meningkat maka mobilitas akan ikut meningkat cukup signifikan; ketika curah hujan dan jumlah tingkat salju meningkat, maka mobilitas menurun cukup signifikan; serta, jika proposir langit yang tertutup oleh awan menaik, maka mobilitas meningkat meskipun tidak secara signifikan [8]. Berdasarkan penelitian [9], kepadatan lalu lintas dipengaruhi oleh hari libur, suhu, curah hujan, jumlah tingkat salju, dan proporsi langit yang tertutup oleh awan sebesar 93.3%.

2.2. Random Forest Algorithm

Algoritma *Random Forest* dapat digunakan untuk melakukan regresi maupun klasifikasi tergantung tujuan analisis yang diinginkan. Jika digunakan untuk melakukan regresi, maka hasil yang ditampilkan adalah rata - rata dari pohon yang berbeda [10].

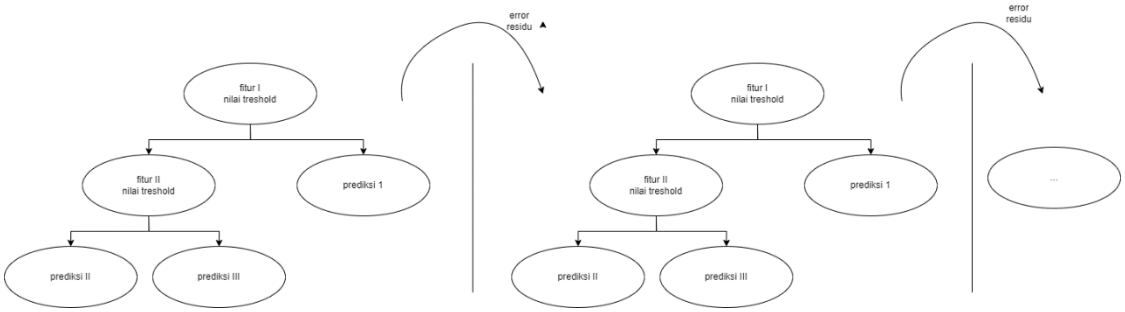


Gambar 1. Proses algoritma *Random Forest*

Gambar 1 menunjukkan ilustrasi algoritma *Random Forest* berjalan. Data akan diproses melalui metode bootstrap untuk menemukan k sampel data yang berbeda. Di setiap *bootstrap,* terdapat pohon keputusan yang dilakukan secara berulang untuk memecah data menjadi dua kelompok dengan

kriteria tertentu. Setiap pohon akan memberikan keputusan hingga didapatkan k keputusan yang berbeda. Kemudian, diambil satu keputusan sebagai keputusan akhir berdasarkan suara voting terbanyak [11].

2.3. XGBoost Algorithm



Gambar 2. Proses algoritma *XGBoost*

eXtreme Gradient Boosting adalah kepanjangan dari *XGBoost*. Algoritma ini mengimplementasikan pengembangan dari pohon keputusan dan menerapkan teknik ansambel artinya model akan terus diperbarui untuk memperbaiki kesalahan pada model sebelumnya. Algoritma ini dibuat dalam rangka meningkatkan efisien waktu pemroses termasuk automasi menangani data hilang, melakukan pohon keputusan secara paralel, serta dapat melakukan training data terus - menerus untuk meningkatkan hasil akurasi [12]. Ilustrasi proses algoritma *XGBoost* ditampilkan pada gambar 2 [13]. Algoritma ini dirumuskan sebagai berikut:

$$obj(\theta) = \sum_i L(\hat{y}_i, y_i) + \sum_i L\Omega(f_k), \quad f_k \epsilon F \quad (1)$$

L menyatakan perbedaan standar deviasi antara nilai prediksi $\hat{y}_i$ dan nilai asli y_i. Ω menyatakan fungsi regularisasi kompleksitas model untuk menghindari overfitting. f menyatakan fungsi dalam ruang fungsional F dan F menyatakan himpunan semua pohon yang dibuat [14].

2.4. Evaluasi Parameter

Untuk melakukan evaluasi terhadap model, terdapat tiga parameter yang akan dibandingkan yaitu koefisien determinasi (R^2) yang digunakan untuk menafsir kelayakan model dengan rentang nilai $0 \leq R^2 \leq 1$. *Root Mean Square Error* (RMSE) yaitu selisih nilai prediksi dengan nilai sebenarnya dan akar dari *Mean Square Error* (MSE), serta *Mean Absolute Error* (MAE) yaitu rata - rata kesalahan mutlak antara nilai prediksi dengan nilai yang sebenarnya. Ketiga parameter dirumuskan sebagai berikut [14] [15]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(\hat{y}_1))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

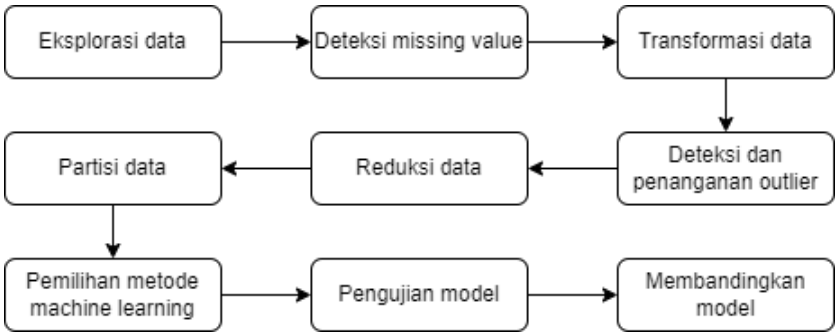
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (3)$$

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

III. METODE PENELITIAN

Dalam studi ini, dataset yang digunakan peneliti adalah Metro Interstate Traffic Volume, yang disediakan oleh repositori data UCI. Dataset ini terdiri 48204 baris dan 9 kolom, dimana data ini berisi informasi volume lalu lintas Interstate 94 Westbound untuk MN DoT ATR station 301, berkisar antara Minneapolis dan St Paul, MN dari tahun 2012-2018 dengan berbagai kondisi, seperti cuaca, hari libur, dan suhu. Sehingga pada penelitian ini, peneliti ingin memprediksi kepadatan lalu lintas berdasarkan beberapa kondisi informasi sebagai berikut:

- holiday yaitu hari libur nasional Amerika Serikat kategorikal ditambah hari libur regional, Minnesota State Fair
- temp yaitu numerik rata-rata suhu dalam kelvin
- rain_1h yaitu numeric jumlah dalam mm hujan yang terjadi dalam satu jam
- snow_1h yaitu jumlah numerik dalam mm dari salju yang terjadi dalam satu jam
- cloud_all yaitu persentase numerik tutupan awan
- weather_main yaitu kategori deskripsi tekstual singkat tentang cuaca saat ini
- weather_description yaitu kategorikal deskripsi tekstual yang lebih panjang tentang cuaca saat ini
- date_time yaitu tanggal, waktu, dan jam dari data yang dikumpulkan dalam waktu CST lokal
- traffic_volume yaitu numeric hourly I-94 ATR 301 melaporkan volume lalu lintas arah barat



Gambar 3. Diagram alur pengolahan data

Gambar 3 menunjukkan alur metode dalam pengolahan data. Tahap *preprocessing* data merupakan langkah penting dalam penelitian yang melibatkan analisis data. Tahapan *preprocessing* data yang dilakukan oleh peneliti meliputi eksplorasi data, penanganan *missing value*, transformasi data, deteksi dan penanganan outlier, reduksi data, dan partisi data. Sedangkan algoritma yang digunakan peneliti adalah *Random Forest* dan *XGBoost*. *Random Forest* dan *XGBoost* merupakan algoritma machine learning yang populer dan sering digunakan karena kemampuannya dalam mengatasi overfitting, meningkatkan akurasi prediksi serta dapat menangani masalah regresi. Basis algoritma *Random Forest* dan *XGBoost* menggunakan *decision tree*, oleh sebab itu peneliti melakukan perbandingan keakuratan dan efisiensi terhadap dataset yang digunakan.

IV. HASIL DAN PEMBAHASAN

Penelitian ini melakukan analisa dengan dua algoritma, yaitu *Random Forest* dan *XGBoost*. Data dipisahkan menjadi dua bagian yaitu data training sebanyak 70% dan data testing sebanyak 30% dengan alat bantu Machine Learning bahasa pemrograman Python.

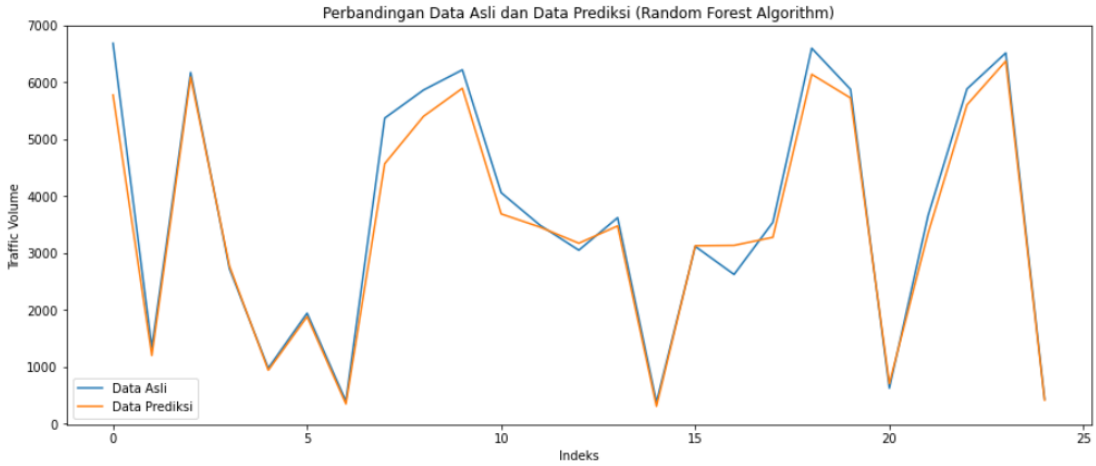
Tabel 1. Perbedaan volume lalu lintas data asli dan prediksi menggunakan *random forest* terhadap 25 data teratas

Indeks	Data Asli	Prediksi
0	6685	5776.19
1	1336	1195.81
2	6172	6092.79
3	2721	2766.86
4	973	936.28
5	1939	1878.79
6	393	343.77
7	5370	4568.07
8	5864	5401.22
9	6218	5894.28
Indeks	**Data Asli**	**Prediksi**

E-ISSN 2808-5841
P-ISSN 2808-7283

10	4058	3684.97
11	3486	3453.21
12	3047	3169.11
13	3619	3473.41
14	374	302.30
15	3113	3125.47
16	2620	3131.39
17	3538	3273.73
18	6599	6139.82
19	5875	5723.51
20	621	701.40
21	3660	3350.04
22	5883	5602.97
23	6516	6372.90
24	429	410.83

Tabel 1 merupakan perbedaan volume lalu lintas dari data asli dengan data prediksi pada 25 data teratas. Gambar 4 merupakan visualisasi perbedaan volume lalu lintas dari data asli dan data prediksi yang terdapat pada Tabel 1. Sumbu x pada grafik di Gambar 4 menunjukkan indeks dari data, sedangkan sumbu y menunjukkan volume lalu lintas. Kurva biru menunjukkan volume lalu lintas dari data asli sedangkan kurva biru menunjukkan volume lalu lintas dari data prediksi menggunakan algoritma *Random Forest*.



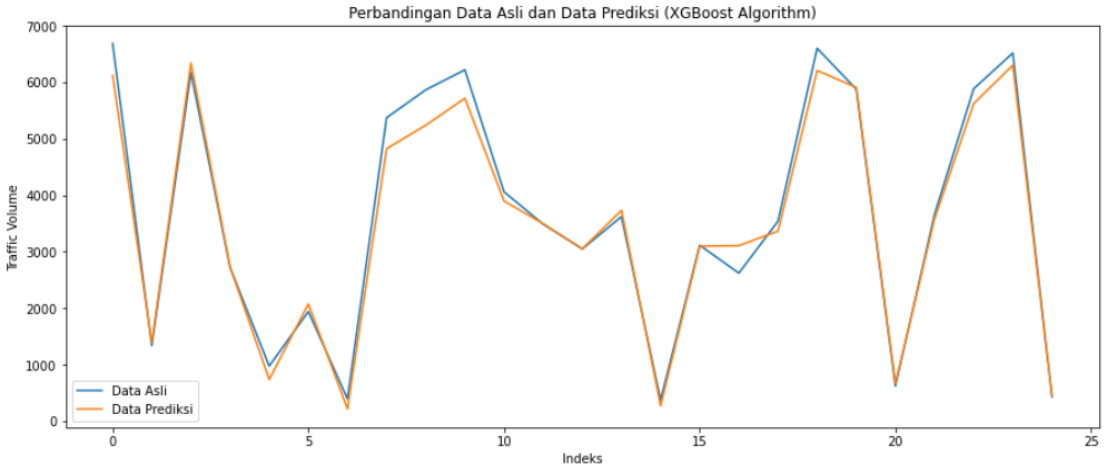
Gambar 4. Visualisasi grafik perbedaan volume lalu lintas data asli dan data prediksi dengan *random forest* terhadap 25 data teratas

Tabel 2. Perbedaan volume lalu lintas data asli dan prediksi menggunakan *XGBoost* terhadap 25 data teratas

Indeks	**Data Asli**	**Prediksi**
0	6685	6121.34
1	1336	1384.57
2	6172	6339.75
3	2721	2731.25
4	973	735.72
5	1939	2073.62
6	393	215.20
7	5370	4823.57
8	5864	5237.60
9	6218	5714.29
10	4058	3896.49
11	3486	3493.68

12	3047	3050.46
Indeks	**Data Asli**	**Prediksi**
13	3619	3730.57
14	374	271.96
15	3113	3096.56
16	2620	3105.85
17	3538	3362.94
18	6599	6206.52
19	5875	5906.25
20	621	666.80
21	3660	3580.76
22	5883	5622.24
23	6516	6300.70
24	429	446.37

Tabel 2 merupakan perbedaan volume lalu lintas dari data asli dengan data prediksi pada 25 data teratas. Gambar 5 merupakan visualisasi perbedaan volume lalu lintas dari data asli dan data prediksi yang terdapat pada Tabel 2. Sumbu x pada grafik di Gambar 5 menunjukkan indeks dari data, sedangkan sumbu y menunjukkan volume lalu lintas. Kurva biru menunjukkan volume lalu lintas dari data asli sedangkan kurva biru menunjukkan volume lalu lintas dari data prediksi menggunakan algoritma *XGBoost*. Dari Gambar 2 dan 5 dapat disimpulkan bahwa hasil prediksi *XGBoost* lebih baik daripada *Random Forest* karena perbedaan antara data asli dan data prediksi yang lebih kecil sesuai dengan nilai RMSE.



Gambar 5. Visualisasi grafik perbedaan volume lalu lintas data asli dan data prediksi dengan *XGBoost* terhadap 25 data teratas

Dalam penelitian ini, evaluasi parameter dilakukan dengan menggunakan koefisien determinasi (R^2), *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan waktu yang dibutuhkan saat pemrosesan setiap algoritma. Hasil pengujian evaluasi parameter antara dua algoritma yang digunakan, yaitu *Random Forest* dan *XGBoost* ditampilkan pada tabel 3

Tabel 3. Perbandingan evaluasi parameter dan waktu pemrosesan antara *Random Forest* dan *XGBoost*

Algoritma	R2	RMSE	MAE	Waktu pemrosesan (s)
Random Forest	95.53	420.09	226.87	19.91
XGBoost	95.92	401.02	236.38	3.15

Berdasarkan tabel 3, didapat nilai akurasi algoritma *Random Forest* sebesar 95.53% sedangkan untuk algoritma *XGBoost* sebesar 95.92%. Nilai RMSE algoritma *Random Forest* sebesar 420.09, sedangkan XGBoost sebesar 401.02 dan nilai MAE *Random Forest* sebesar 226.87, sedangkan

XGBoost sebesar 236.38.

Nilai R2 *XGBoost* yang lebih tinggi, menunjukkan bahwa kelayakan model lebih bagus. Nilai RMSE *XGBoost* yang lebih rendah menunjukkan bahwa perbedaan nilai prediksi dengan nilai sebenarnya lebih kecil sehingga lebih bagus. Nilai MAE *Random Forest* yang lebih rendah menunjukkan bahwa rata - rata kesalahan mutlak antara nilai prediksi dengan nilai yang sebenarnya lebih rendah. Hasil R2, RMSE, dan MAE dari kedua algoritma ini tidak berbeda secara signifikan. Namun, untuk perbedaan nilai waktu pemrosesan antara *Random Forest* dan *XGBoost* memiliki perbedaan yang signifikan. *XGBoost* memiliki waktu pemrosesan yang sangat cepat, yaitu 532% lebih cepat dibanding *Random Forest*.

V. KESIMPULAN

Berdasarkan penelitian yang dilakukan dengan menggunakan algoritma *Random Forest* dan *XGBoost* tidak menunjukkan perbedaan yang signifikan mengenai prediksi data volume lalu lintas. Nilai R2 pada algoritma *XGBoost* yang lebih tinggi, menunjukkan bahwa kelayakan model lebih bagus. Nilai RMSE *XGBoost* yang lebih rendah menunjukkan bahwa perbedaan nilai prediksi dengan nilai sebenarnya lebih kecil sehingga lebih bagus. Sedangkan nilai MAE *Random Forest* yang lebih rendah menunjukkan bahwa rata - rata kesalahan mutlak antara nilai prediksi dengan nilai yang sebenarnya lebih rendah. Hasil R2, RMSE, dan MAE dari kedua algoritma ini tidak berbeda secara signifikan. Namun, untuk perbedaan nilai waktu pemrosesan antara *Random Forest* dan *XGBoost* memiliki perbedaan yang sangat signifikan. *XGBoost* memiliki waktu pemrosesan yang sangat cepat, yaitu 532% lebih cepat dibanding *Random Forest*. Dari hasil perbandingan, algoritma *XGBoost* adalah pilihan yang tepat dalam membangun model prediksi kepadatan lalu lintas karena memiliki akurasi tinggi dan efisiensi dalam waktu.

UCAPAN TERIMA KASIH

REFERENSI

1. U. nations, "Data Portal Population Division," population.un.org, [Online]. Available: https://population.un.org/dataportal/home. [Accessed 1 April 2023].
2. w. p. review, "2023 World Population by Country (Live)," worldpopulationreview.com, [Online]. Available: https://worldpopulationreview.com/. [Accessed 1 April 2023].
3. M. Hudzaifah and A. A. Rismayadi, "PERAMALAN ARUS LALU LINTAS BERDASARKAN WAKTU TEMPUH DAN CUACA MENGGUNAKAN METODE TIME SERIES DECOMPOSITION," *JURNAL RESPONSIF,* vol. III, no. 2, 2021.
4. B. ALAOUI, D. BARI and Y. GHABBAR, "Surface Weather Parameters Forecasting Using Analog Ensemble the Main Airports of Morocco," *Journal of Meteorological Research,* vol. XXXVI, 2022.
5. A. C. N and D. H. V. Kumaraswamy, "Traffic Prediction using Random Forest Machine Learning Algorithms," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* vol. XI, no. 4, pp. 55 - 59, 2022.
6. J. S. ANGARITA-ZAPATA, A. D. MASEGOSA and I. TRIGUERO, "A Taxonomy of Traffic Forecasting Regression Problems From a Supervised Learning Perspective," *IEEE Access,* vol. VII, 2019.
7. J. Liu, N. Wu, Y. Qiao and Z. Li, "A scientometric review of research on traffic forecasting in transportation," *IET Intell Transp Syst,* vol. XV, no. 1, pp. 1-16, 2021.
8. P. Lin, Y. He, M. Pei and R. Yang, "Data-driven spatial-temporal analysis of highway traffic volume considering weather and festival impacts," *Travel Behaviour and Society,* vol. XXIX, pp. 95 - 112, 2022.

9. H. J. Nascimento, "Statistical estimation of traffic volume for the Minneapolis-St Paul Metropolitan area," *International Journal of Advanced Engineering Research and Science (IJAERS),* vol. IX, no. 11, pp. 392 - 397, 2022.
10. I. Alam, D. M. Farid and R. J. F. Rossetti, "The Prediction of Traffic Flow with Regression Analysis," in *Emerging Technologies in Data Mining and Information Security*, 2018.
11. Y. Liu and H. Wu, "Prediction of Road Traffic Congestion Based on Random Forest," in *10th International Symposium on Computational Intelligence and Design*, Hangzhou, China, 2017.
12. J. Brownlee, XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn, Machine Learning Mastery, 2016.
13. A. I. A. Osman, A. N. Ahmed, M. F. Chow, Y. F. Huang and A. El-Shafie, " Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Engineering Journal,* vol. XII, no. 2, pp. 1545 - 1556, 2021.
14. J. Luo, Z. Zhang, Y. Fu and F. Rao, "Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms," *Results in Physics,* vol. XXVII, 2021.
15. A. E. Putra and A. Juarna, "Prediksi Produksi Daging Sapi Nasional dengan Metode Regresi Linier dan Regresi Polinomial," *Jurnal Ilmiah KOMPUTASI,* vol. XX, no. 2, pp. 209 - 215, 2021.