

RISTEK DATATHON 2023

Forecast Kecepatan Kendaraan pada Waktu Tertentu



Bintang Riset SC

Nailfaaz (22/498485/PA/21523)

Delfia Nur Anrianti Putri (21/473882/PA/20438)

Raden Bagus Muhammad Adryan Putra Adhy Wijaya (21/473083/PA/20366)

**DEPARTEMEN ILMU KOMPUTER DAN ELEKTRONIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS GADJAH MADA**

Daftar Isi

Daftar Isi.....	2
1 Pendahuluan	3
2 Metodologi	3
2.1 Usulan Pendekatan	3
2.2 Exploratory Data Analysis	4
2.2.1 Dataset.....	4
2.2.2 Daily Traffic Trend.....	4
2.2.3 Weekly Traffic Trend.....	5
2.2.4 Segment-wise Traffic Trend.....	6
2.2.5 Kesimpulan Analisa.....	8
2.3 Data Preprocessing	8
2.3.1 Interpolasi Data	8
2.3.1 Pemisahan Data Per Segmen Jalan.....	8
2.4 Feature Engineering.....	8
2.4.1 Jam.....	8
2.4.2 Hari	8
2.4.3 Weekend.....	9
2.4.4 Kelas Kelancaran.....	9
2.5 Modeling.....	9
2.6 Evaluasi	10
3 Analisis Hasil Prediksi	11
4 Kesimpulan.....	13
Daftar Pustaka	14

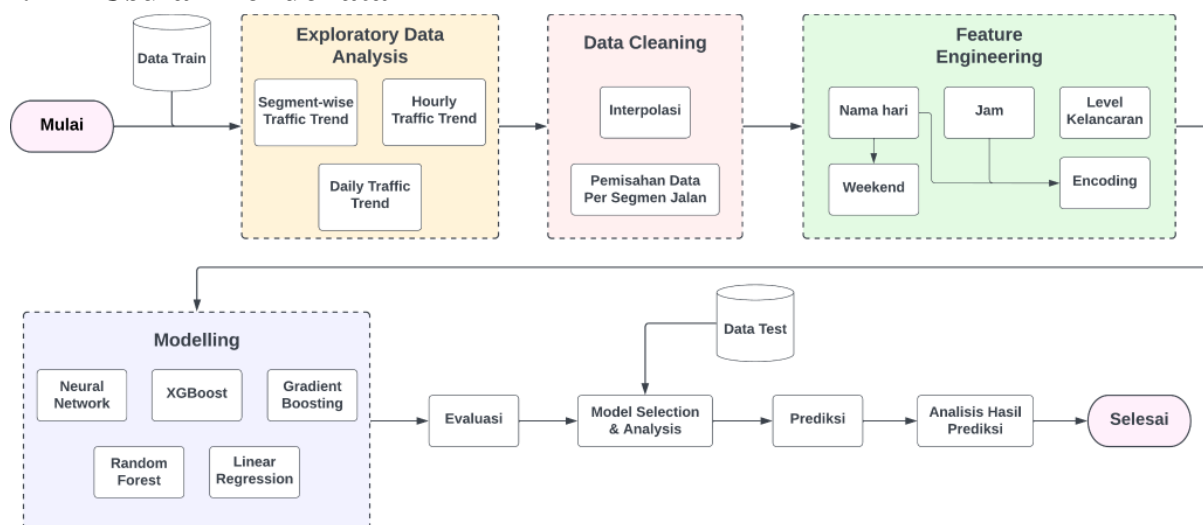
1 Pendahuluan

Kemacetan di jalan raya kota besar menyebabkan masalah dan kerugian yang signifikan. Menurut sebuah studi dari Universitas Indonesia, kerugian ekonomi akibat kemacetan lalu lintas di Jakarta diperkirakan mencapai Rp 28,1 triliun (US\$ 3,25 miliar) per tahun [1]. Di Bandar Lampung, studi [2] memperkirakan bahwa biaya yang ditimbulkan akibat kecelakaan, termasuk biaya pengobatan dan biaya hilangnya produktivitas, mencapai Rp 50,8 miliar dalam satu tahun. Angka-angka ini menunjukkan dampak ekonomi yang signifikan dari kemacetan di Indonesia. Untuk mengatasi masalah kemacetan jalan di kota, analisa terhadap data kecepatan kendaraan pada suatu jalan dapat dilakukan untuk memprediksi kecepatan rata-rata di jalan tersebut di waktu selanjutnya. Informasi ini dapat membantu para perencana kota dan otoritas transportasi untuk membuat kebijakan mengenai strategi pengelolaan lalu lintas atau meningkatkan pilihan transportasi publik, serta membantu pengguna jalan memutuskan rute perjalanan yang dapat meminimalisir kemacetan.

Dalam kompetisi ini, kami menggunakan data kecepatan rata-rata dari waktu ke waktu pada suatu segmen jalan selama 3 minggu untuk melakukan prediksi rata-rata kecepatan pada 7 hari berikutnya. Objektif dari proses pengolahan data ini adalah untuk: (1) menemukan tren lalu lintas, (2) menganalisa faktor-faktor kemacetan atau kelancaran jalan, (3) membangun model *forecasting* yang optimal untuk memprediksi rerata kecepatan kendaraan di suatu jalan pada waktu tertentu, dan (4) menganalisa hasil prediksi model *forecasting*.

2 Metodologi

2.1 Usulan Pendekatan



Gambar 1. Usulan Pendekatan Masalah Forecast Kecepatan Kendaraan pada Waktu Tertentu.

Penelitian kami tersusun dalam serangkaian tahapan yang saling terkait yang dapat dilihat pada **Gambar 1** Dimulai dengan *Exploratory Data Analysis* (EDA) pada data *train*, kami mengeksplorasi dan mengidentifikasi *insight* penting untuk tahapan berikutnya. Temuan dari EDA memandu proses pembersihan data. Dengan fondasi data yang lebih bersih, kami beralih ke *feature engineering* dimana kami merancang fitur-fitur tambahan yang bertujuan mengoptimalkan hasil prediksi model. Setelah mempersiapkan data, kami melatih dan membandingkan beberapa model *machine learning*. Evaluasi dari setiap model yang telah terlatih memungkinkan kami untuk memilih yang paling optimal, yang kemudian dianalisis lebih lanjut. Model yang telah dianalisis kami gunakan untuk memprediksi data *test*. Terakhir, kami melakukan analisis terhadap hasil prediksi untuk menilai kualitas dan performa model secara visual dan kualitatif.

2.2 *Exploratory Data Analysis*

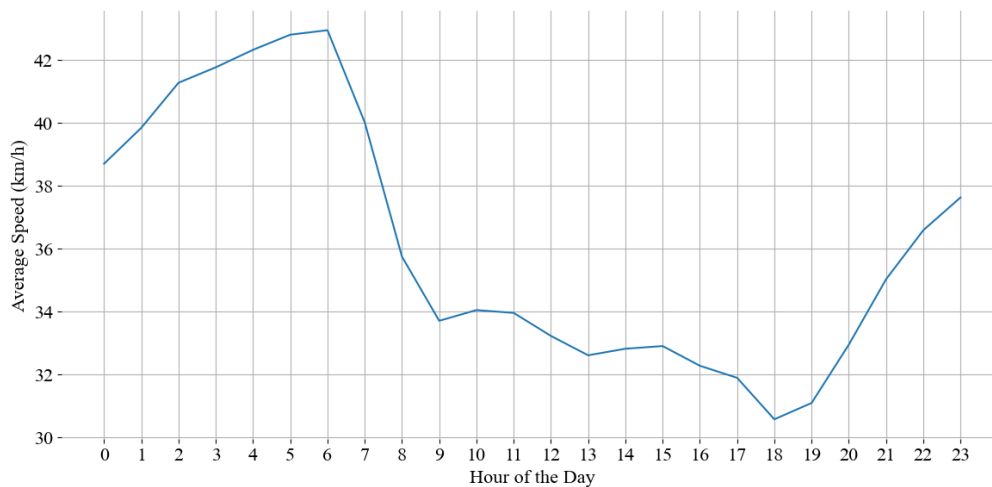
2.2.1 Dataset

Dataset yang diberikan berisi data tentang kecepatan rata-rata pada berbagai segmen jalan raya pada waktu tertentu dengan fitur berupa *waktu_setempat*, *indeks_jalan*, *id_titik_mulai*, *id_titik_akhir*, dan label berupa *rerata_kecepatan*. Jumlah baris dalam data *train* adalah 398648 baris dan pada data *testing* sebanyak 127489 baris.

Pada dataset ini, variabel *waktu_setempat* berisikan tanggal dan jam pada zona waktu lokal London, Inggris dari 1 Februari 2020 pukul 01:00:00 hingga 22 Februari 2020 pukul 23:00:00. Berdasarkan total nilai unik dari *id_titik_mulai* dan *id_titik_akhir*, didapatkan 934 segmen jalan raya berbeda. Meskipun tidak terdapat *missing values* eksplisit dalam data, kami melihat bahwa hanya terdapat 3 segmen jalan yang memiliki *record* rerata kecepatan yang utuh atau memiliki tepat 527 *data point*. Jumlah *data point* paling sedikit yaitu 233 yang dimiliki oleh segmen jalan dengan *id_jalan* 47010584, *id_titik_mulai* 196049 dan *id_titik_akhir* 1030634577. Secara keseluruhan, terdapat 207 data dengan *missing values* di atas 30%.

2.2.2 *Daily Traffic Trend*

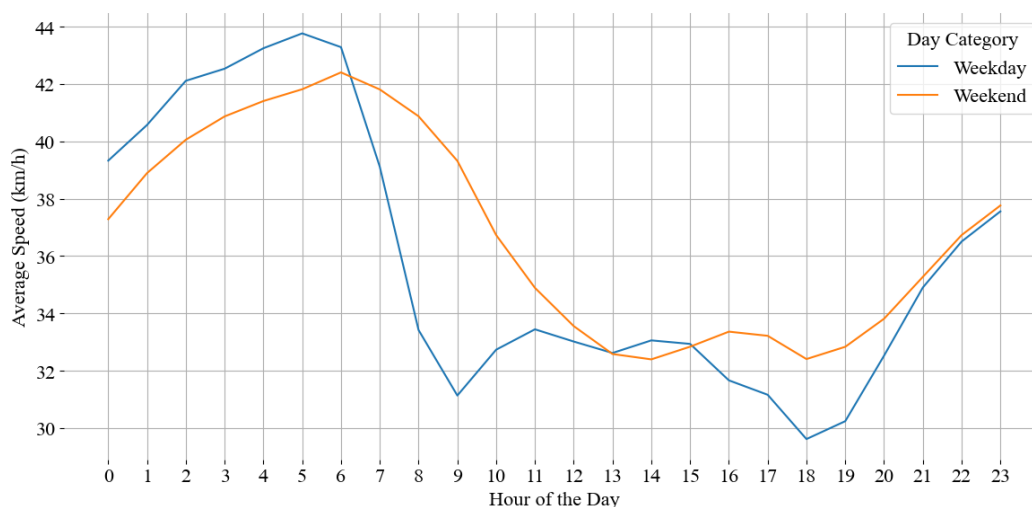
Menggunakan grafik pada **Gambar 2**, kami menemukan pola jam sibuk/*rush hour* yang berguna dalam memperkirakan tingkat kemacetan maupun kelancaran. Kami menganalisa pola rerata kecepatan pada setiap jam dalam satu hari dan menemukan bahwa rerata kecepatan menurun pada jam 6:00-9:00 serta 17:00-19:00 waktu setempat, dimana waktu-waktu ini berhubungan dengan jam sibuk pagi dan sore hari di Kota London. Selain itu, rerata kecepatan yang tinggi yang mengindikasikan kelancaran lalu lintas berada pada tengah malam hingga pukul 6:00, dimana waktu tersebut merupakan waktu tidur dari masyarakat kota berdasarkan survei [3] .



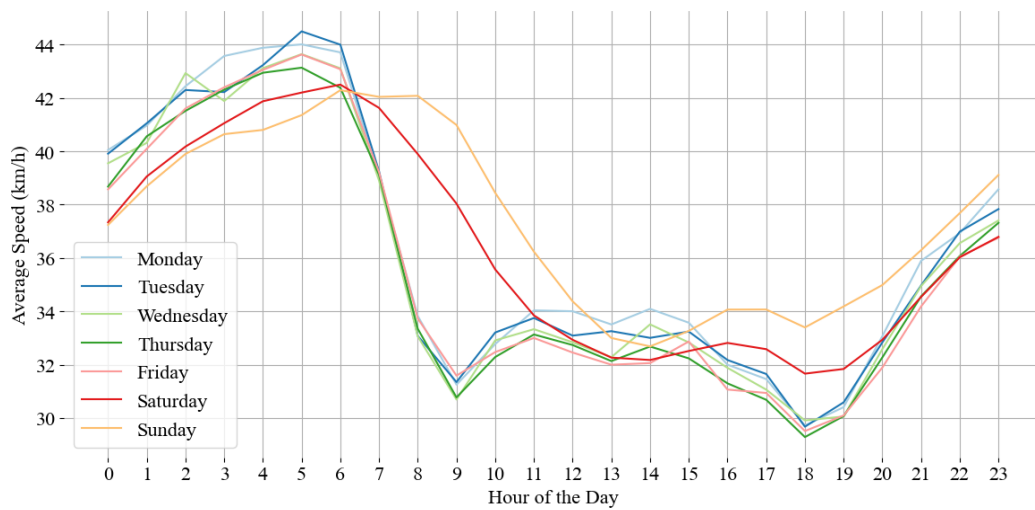
Gambar 2. Kecepatan Rata-rata Per Jam.

2.2.3 Weekly Traffic Trend

Kami membandingkan rata-rata kecepatan mingguan menggunakan plot pada **Gambar 3** dan **Gambar 4**. Pada hari kerja (Senin-Jumat) maupun akhir pekan (Sabtu-Minggu), terjadi penurunan dalam rata-rata kecepatan pukul 6:00 hingga 9:00. Meskipun demikian, penurunan kecepatan pada hari kerja lebih tajam, mengindikasikan adanya jam sibuk pagi yang lebih intens selama hari kerja serta pada akhir minggu, masyarakat cenderung beristirahat lebih lama dibandingkan pada hari kerja [4]. Pola jam sibuk juga terlihat di sore hari di hari kerja, sementara rata-rata kecepatan akhir pekan tetap relatif stabil sepanjang siang dan sore hari. Terdapat peningkatan kecil dalam rata-rata kecepatan pada akhir pekan setelah tengah malam, mengindikasikan adanya potensi aktivitas larut malam yang lebih banyak di akhir pekan dibandingkan hari kerja.



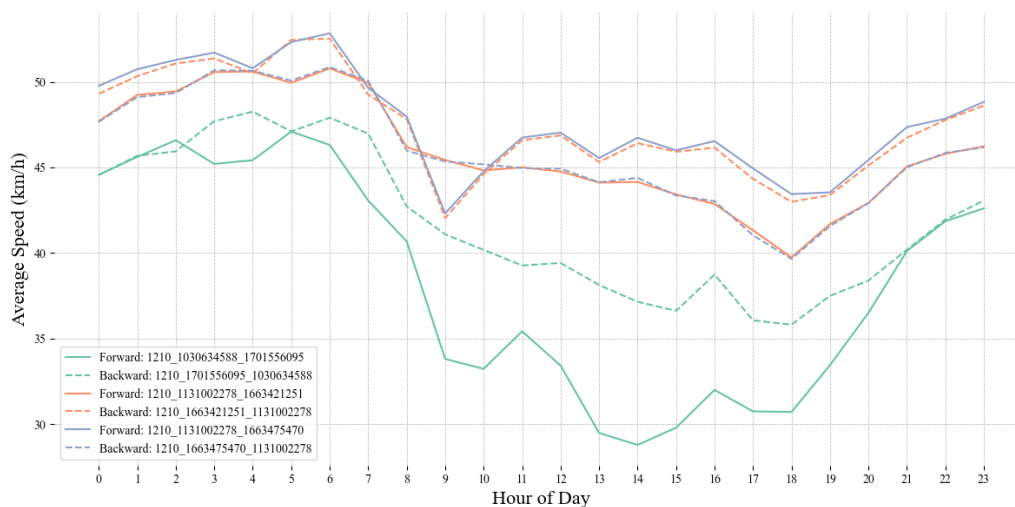
Gambar 3. Kecepatan Rata-rata per Jam: Hari Kerja vs. Akhir Pekan.



Gambar 4. Kecepatan Rata-rata per Jam untuk Berbagai Hari dalam Seminggu.

2.2.4 Segment-wise Traffic Trend

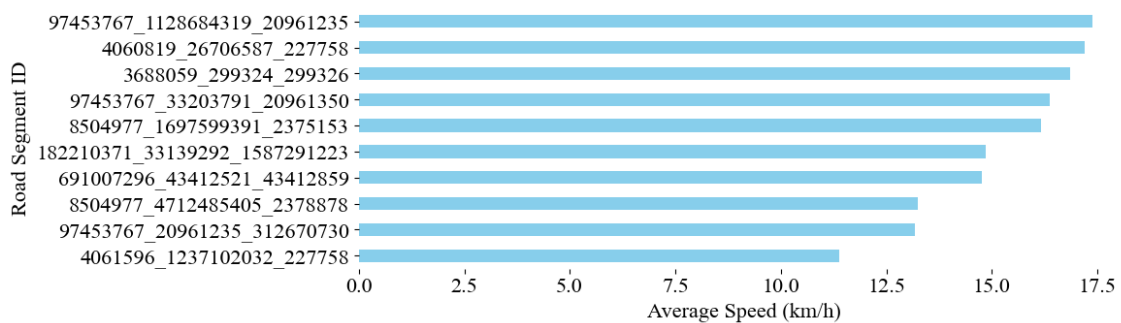
Dalam data, untuk setiap segmen jalan¹ dari titik mulai ke titik akhir, terdapat segmen jalan dari arah sebaliknya, yang mengindikasikan bahwa jalan pada data merupakan jalan dua arah. Kami mengamati tren *rerata_kecepatan* dari *forward lane* dan *backward lane* dengan plot pada **Gambar 5**. ditemukan perbedaan signifikan pada *rerata_kecepatan* di jam-jam tertentu meskipun merupakan bagian dari jalan raya yang sama pada beberapa segmen jalan yang berarti ada kemungkinan bahwa satu jalur lebih rawan mengalami kemacetan dibandingkan jalur lainnya pada jam sibuk. Selain itu, tren *rerata_kecepatan* yang serupa dapat ditemukan pada segmen-segmen jalan yang terkoneksi langsung pada satu jalur seperti pada segmen jalan "1210_1663421251_1131002278" dan "1210_1131002278_1663475470" yang menunjukkan pematatan di jam sibuk pagi serta jalur sebaliknya pada jam sibuk sore.



Gambar 5. Perbandingan Tren Rerata Kecepatan Pada Sampel Segmen Jalan Raya.

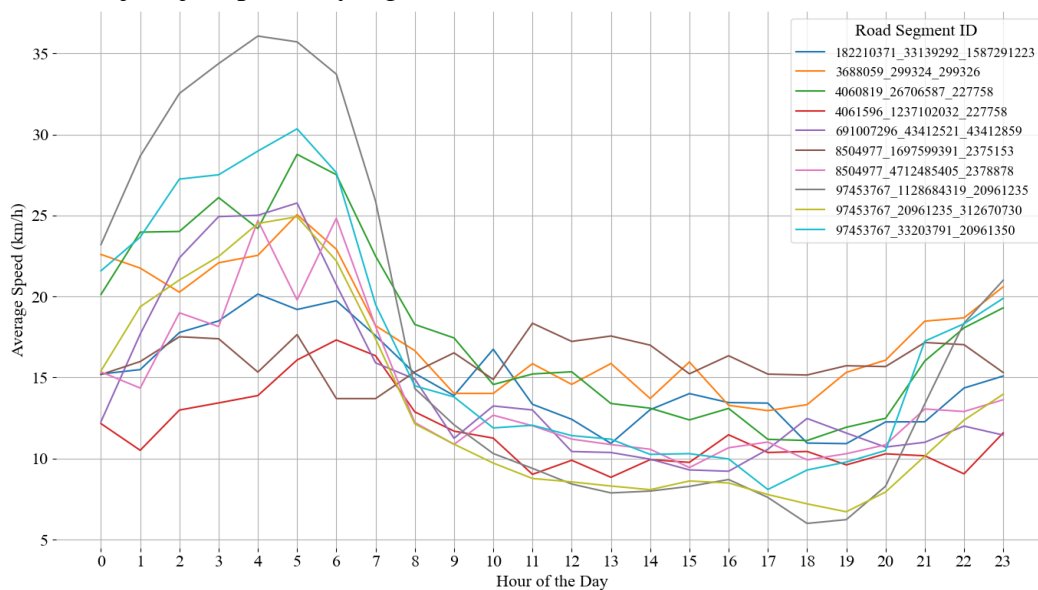
¹ Pada makalah ini, "{id_jalan}_{id_titik_mulai}_{id_titik_akhir}" digunakan sebagai format penamaan segmen jalan.

Mengacu pada plot pada **Gambar 6**, kami juga menganalisa pola mingguan dari 10 segmen jalan yang memiliki rerata_kecepatan yang terendah. Di antara segmen-segmen dengan rerata_kecepatan terendah tersebut, tiga di antaranya merupakan bagian dari jalan raya North End Road "97453767" yang memiliki rerata kecepatan terendah, yaitu 29.2 km/h. Hal ini dapat dikarenakan oleh aktivitas masyarakat di pasar North End Road yang terkenal juga memiliki nilai historis yang tinggi [5].



Gambar 6. 10 Segmen Jalan Terpadat.

Dari tren harian pada **Gambar 7**, kami menyadari bahwa terdapat segmen-segmen jalan seperti "4061596_1237102032_227758" yang menunjukkan kecepatan yang relatif rendah sepanjang hari. Hal ini dapat mengindikasikan adanya kemacetan yang konsisten atau kemungkinan adanya masalah infrastruktur. Beberapa segmen mengalami penurunan yang signifikan selama jam-jam sibuk, sementara segmen lain memiliki kecepatan yang lebih konsisten serta jam-jam puncak yang berbeda.



Gambar 7. Kecepatan Rata-rata per Jam untuk 10 Jalan Terpadat.

2.3 Data Preprocessing

2.3.1 Interpolasi Data

Missing value pada beberapa interval waktu kami imputasi dengan teknik interpolasi. Menggunakan *library* `numpy`, kami memanfaatkan fungsi `numpy.interp()` untuk melakukan interpolasi linear. Dengan metode interpolasi linear, nilai rerata kecepatan pada suatu jam yang hilang diestimasi berdasarkan posisinya relatif terhadap garis yang menghubungkan nilai rerata kecepatan pada data jam sebelum dan sesudahnya.

2.3.1 Pemisahan Data Per Segmen Jalan

Kami mengidentifikasi adanya 20 `id_jalan` yang masing-masing memiliki beragam `id_titik_mulai` dan `id_titik_akhir`. Selain itu, kami juga menemukan bahwa setiap `id_jalan` memiliki dua ruas jalan yang berbeda pada setiap rantai *node* (`id_titik_mulai` dan `id_titik_akhir`). Untuk memastikan analisis yang lebih tepat, kami membagi setiap segmen jalan ke dalam *dataframe* terpisah. Langkah ini diambil agar saat memasukkan data ke dalam model, setiap segmen jalan memiliki trennya sendiri dan tidak mempengaruhi segmen lain, sehingga menghindari adanya potensi bias.

2.4 Feature Engineering

2.4.1 Jam

Kami menciptakan fitur tambahan bernama `jam` yang merupakan turunan dari kolom `waktu_setempat`. Fitur ini tipe data integer yang menampilkan jam dengan rentang 0 hingga 23 pada setiap baris data. Penambahan fitur ini didasarkan pada temuan bahwa terdapat *hourly pattern* yang serupa untuk segmen-segmen jalan dalam setiap harinya.

2.4.2 Hari

Selain mengekstrak fitur `jam`, kami juga mengembangkan fitur `hari` dengan tipe data object yang didapatkan dari fitur `waktu_setempat`. Kemudian, berdasarkan kalender bulan Februari 2020, kami mengkonversi tanggal tersebut ke dalam nama hari yang sesuai. Sejalan dengan ekstraksi fitur `jam`, fitur ini kami rancang guna mendapati pendekatan *time series* pada model.

2.4.3 Weekend

Fitur `weekend` merupakan turunan dari fitur `hari`. Fitur ini bertipe *boolean*, di mana nilai akan menjadi `true` jika dalam baris yang sama fitur `hari` bernilai `Saturday` atau `Sunday` dan `false` untuk hari-hari lainnya. Kami memutuskan untuk menambahkan fitur ini

dikarenakan pada proses EDA terdapat perbedaan tren kecepatan rerata antara hari kerja dan akhir pekan. Dengan fitur ini, model akan lebih akurat dalam menemukan pola tersebut.

2.4.4 Kelas Kelancaran

Berdasarkan **Bagian 2.2.2**, kami mengamati adanya perbedaan tren pada jam-jam spesifik di tiap segmen jalan. Perbedaan ini disebabkan oleh fenomena *rush hour* di kota London. Untuk mengatasi hal ini, kami memperkenalkan fitur `kelas_kelancaran`. Fitur ini merupakan hasil pengelompokan `rerata_kecepatan` ke dalam empat kelas berdasarkan interval kuartil *mean rerata_kecepatan* di suatu jam tertentu. Jika nilai *mean rerata_kecepatan* di suatu jam tertentu di bawah nilai Q1 termasuk kelas 0, diantara nilai Q1 dan nilai Q2 akan digolongkan kelas 1, diantara nilai Q2 dan nilai Q3 akan dikelompokkan sebagai kelas 2, dan yang di atas nilai Q3 termasuk kelas 3. Distribusi nilai kuartil didapatkan dengan mengambil nilai rata-rata `rerata_kecepatan` di tiap jam yang sama di suatu segmen jalan tertentu. Dengan pendekatan ini, setiap segmen jalan mendapat pembagian kelas `kelas_kelancaran` yang unik, sehingga memastikan tidak ada bias dalam pengklasifikasian kelas antar segmen jalan yang memiliki tren berbeda-beda. Harapannya, fitur ini dapat membuat model lebih akurat dalam menganalisis tren harian `rerata_kecepatan` pada waktu dan segmen jalan tertentu.

2.5 Modeling

Berdasarkan data test, prediksi akan dilakukan untuk berbagai segmen jalan pada waktu yang berbeda-beda. Karena segmen jalan satu dengan lainnya memiliki karakteristik tren rerata kecepatan yang berbeda, kami memutuskan untuk membuat satu model per segmen jalan atau total 934 model. Meskipun jumlah model yang dilatih tergolong sangat banyak, setiap model hanya menerima 527 baris data sehingga proses pelatihan tergolong cepat. Keunggulan metode ini adalah model dapat berfokus dalam menganalisa satu segmen jalan spesifik dalam proses pelatihan dan juga inferensi. Selain itu, penyesuaian atau pembaruan pada model lebih mudah dilakukan jika terjadi perubahan dalam karakteristik lalu lintas di suatu segmen jalan tertentu.

Kami bereksperimen dengan 5 jenis model yang berbeda, yakni *Linear Regression* (LR), *Gradient Boosting Regressor* (GB), *Random Forest Regressor* (RF), *XGBoost Regressor* (XGB) dan *Deep Neural Network* (DNN). LR dapat mengolah kumpulan data dalam jumlah yang sangat besar dengan komputasi yang sangat efisien dan model yang dihasilkan lebih mudah untuk dipahami dibandingkan model lainnya [6]. GB dapat mengatasi permasalahan dataset yang mengandung nilai *outlier* dan dapat menemukan hubungan antar variabel yang non-linear [7]. RF dapat mengolah data yang terdapat *missing values* didalamnya, cocok untuk fitur kategorikal, dan juga dapat memberikan informasi berupa fitur-fitur yang paling penting untuk melakukan prediksi [8]. XGB merupakan sebuah metode yang dapat mengolah data dalam skala yang besar dengan lebih efisien dan dapat membuat hasil prediksi yang *general* sehingga dapat menghindari *overfitting* [9]. DNN dapat mencari pola-pola yang kompleks dalam

data sehingga dapat menghasilkan performa yang lebih baik [10]. Dalam memilih metode yang paling tepat, kelebihan dari masing-masing metode tersebut harus dipertimbangkan untuk menghasilkan performa yang paling optimal untuk menyelesaikan permasalahan *forecasting* ini.

2.6 Evaluasi

Kami melakukan perbandingan performa model dengan melakukan *fit model* terhadap masing-masing 934 bagian jalan dan melakukan *testing* dengan data *test public* yang disediakan oleh panitia. Model akan dites dengan metrik sMAPE [11]. Metrik tersebut dipilih dikarenakan metrik sMAPE sangat optimal dalam membandingkan kinerja prediksi model regresi pada kumpulan data dengan skala nilai yang berbeda [12].

Tabel 1. Nilai sMAPE Error untuk Setiap Model

Model	sMAPE Error
Random Forest Regressor	8.18209
XGB Regressor	8.23000
Gradient Boosting Regressor	8.58011
Deep Neural Network	8.70645
Linear Regression	9.45250

Dari **Tabel 1**, didapatkan informasi bahwa model RF, GB dan XGB memberikan performa yang terbaik dibandingkan oleh model lainnya. Hal ini dikarenakan model-model tersebut dapat menemukan pola data yang lebih kompleks dibandingkan model lainnya, seperti menemukan pola dan hubungan dalam fitur-fitur yang berbentuk *binary*. DNN memberikan performa yang kurang optimal dikarenakan jumlah data yang terbatas sehingga pola-pola data yang kompleks tidak dapat ditemukan oleh model ini. Meskipun model LR merupakan model yang dapat melakukan prediksi dengan waktu sangat efisien, model ini memberikan performa yang paling buruk dibandingkan model lainnya.

Berdasarkan evaluasi yang telah kami lakukan, kami memutuskan untuk menggunakan model *Random Forest Regressor* (RF) untuk memprediksi data *test*. Model RF sebagai model berbasis *bagging tree based*, sangat sesuai dalam situasi di mana terdapat banyak data hasil *encoding* yang bersifat biner. Setelah menetapkan keputusan untuk menggunakan RF, kami melanjutkan dengan menganalisis fitur-fitur yang memiliki peran signifikan dalam model RF ini, untuk memahami lebih lanjut kontribusi masing-masing fitur terhadap prediksi.

Tabel 2. Rangkuman Feature Importance Tiap Segmen Jalan

Fitur	Mean_importance	Count_top5
kelas_kelancaran	0.602288	934
weekend	0.049492	769
hari_Sunday	0.025749	451
hari_Monday	0.022672	347
hari_Friday	0.022817	333

Kami juga melakukan ekstraksi untuk menentukan besaran pentingnya setiap fitur pada masing-masing segmen jalan. Kesimpulan dari analisis ini dapat dilihat pada **Tabel 2**. Dalam tabel tersebut, kolom *Mean_importance* menunjukkan rerata pentingnya fitur di semua model, sementara kolom *Count_top5* menunjukkan berapa kali suatu fitur muncul di antara lima fitur terpenting dalam model tersebut. Menariknya, fitur *kelas_kelancaran* muncul 934 kali dalam *Count_top5*, yang berarti fitur ini selalu berada di lima besar dalam setiap model, mengingat total jumlah segmen jalan adalah 934. Selain *kelas_kelancaran*, fitur-fitur yang sangat berpengaruh meliputi *weekend*, *hari_Sunday*, *hari_Monday*, dan *hari_Friday*. Semua fitur ini, yang berkaitan dengan hari-hari tertentu, terutama di sekitar akhir pekan, menunjukkan pengaruh yang lebih tinggi dalam prediksi. Hal ini mungkin disebabkan oleh perbedaan antara *weekday* dan *weekend*.

Tabel 3. Rangkuman Feature Importance Fitur Jam Tiap Segmen Jalan

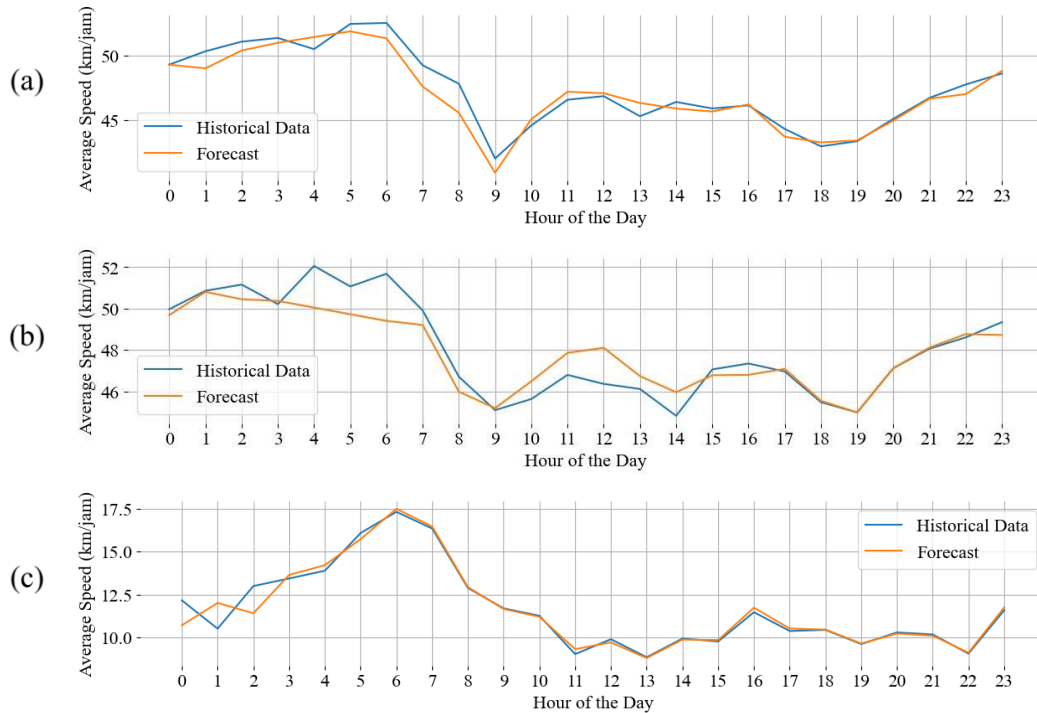
Fitur	Mean_importance	Count_top5
jam_9	0.019293	299
jam_8	0.016186	189
jam_7	0.010205	59
jam_18	0.011150	79

Selain itu, kami juga memfokuskan analisis pada fitur *jam*. Berdasarkan **Tabel 3**, fitur *jam* yang menonjol meliputi *jam_9*, *jam_8*, *jam_7*, dan *jam_18*. Jam-jam tersebut merupakan bagian dari jam sibuk pagi dan sore, sehingga, kami dapat menyimpulkan bahwa periode jam sibuk memegang peranan penting dalam prediksi model ini.

3 Analisis Hasil Prediksi

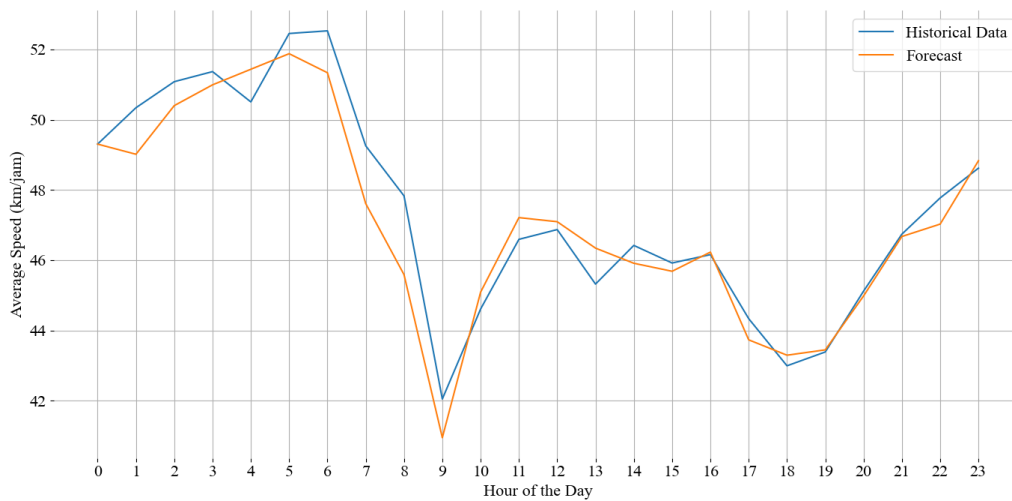
Kami membandingkan hasil *forecasting* dari model dengan data historis untuk melihat apakah model dapat mempelajari pola data dengan baik. Analisa pertama dilakukan dengan memvisualisasikan hasil prediksi pada beberapa segmen jalan seperti yang tertera pada **Gambar 8**. Tren dari data *forecast* tampak mengikuti tren umum yang pada data historis. Namun, yang menarik adalah adanya beberapa penurunan signifikan pada beberapa bagian yang mana sudah sejalan dengan apa yang diharapkan dari prediksi model. Meskipun demikian, model kami tampaknya masih enggan memprediksi nilai yang sangat rendah, sebagaimana yang

ada pada beberapa bagian di data *train*. Hal ini dapat dikarenakan oleh adanya outlier handling pada model *Random Forest* [13].



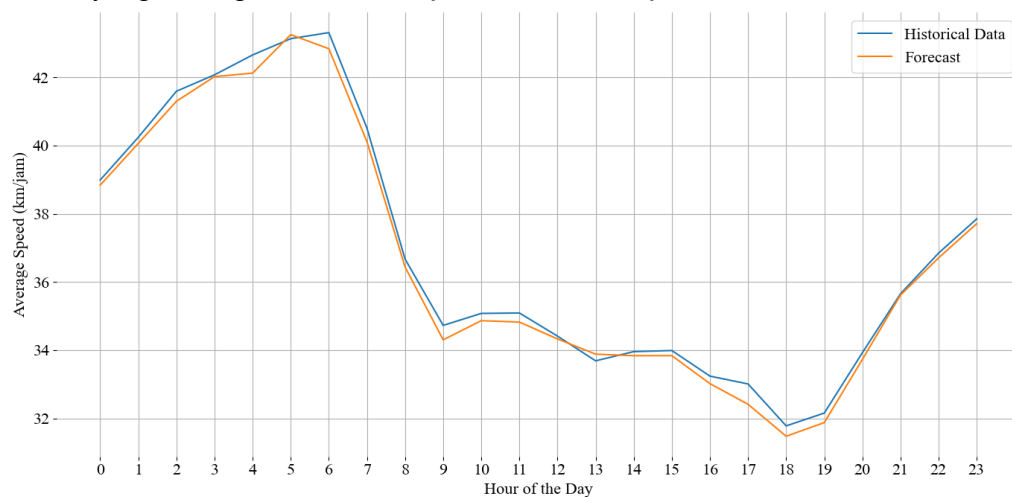
Gambar 8. Perbandingan rerata_kecepatan Semua Tanggal Data Historis dan *Forecast* dengan RF Pada Sampel Segmen (a) "1210_1663421251_1131002278", (b) "3688059_25631788_299320", dan (c) "4061596_1237102032_227758".

Selain mengamati visualisasi tren data secara keseluruhan, kami juga memfokuskan analisis pada tren harian. Berdasarkan **Gambar 9**, tampak bahwa model prediksi mampu mengenali tren kenaikan dan penurunan yang berlangsung setiap jam. Meski model tidak selalu sempurna dalam menyesuaikan setiap detail kenaikan dan penurunan, secara garis besar, tren yang dihasilkan model sudah menggambarkan pola yang ada. Hal ini penting untuk memastikan model tidak *overfit* terhadap data.



Gambar 9. Perbandingan rerata_kecepatan Harian Data Historis dan Forecast dengan RF Pada Sampel Segmen "1210_1663421251_1131002278".

Kemudian berdasar dari **Gambar 10**, model dapat menemukan dan menggeneralisasi pola dari tren rerata_kecepatan harian dengan baik. Hasil prediksi dengan data *median* dari data sebelumnya cukup mirip tetapi tidak saling tumpang tindih, sehingga dapat disimpulkan bahwa model yang dibangun sudah *well-fit* dan tidak *overfit*.



Gambar 10. Perbandingan rerata_kecepatan Harian Data Historis dan Forecast dengan RF Pada Seluruh Segmen Jalan Menggunakan Nilai Median.

4 Kesimpulan

Berdasarkan analisa data, terdapat pola waktu lalu lintas yang berkontribusi signifikan terhadap kemacetan atau kelancaran jalan, yaitu jam sibuk dan jenis hari (hari kerja atau akhir pekan). Pola ini juga dapat diamati pada level jalan dan segmen jalan dengan tren rentang kecepatan yang berbeda-beda, menandakan bahwa setiap jalan maupun segmen jalan memiliki ciri khas tersendiri. Informasi ini menjadi acuan dalam melakukan *forecasting* kecepatan kendaraan di masa mendatang, sehingga, untuk menyelesaikan permasalahan ini, pemodelan dilakukan berdasarkan segmen jalan dan dibuat fitur-fitur turunan dari data seperti hari dan jam dari fitur waktu_setempat, kelas_kelancaran yang mengelompokkan kelancaran segmen jalan di jam tertentu berdasarkan rerata_kecepatan, serta weekend yang mengindikasikan waktu akhir pekan. Fitur-fitur ini memberikan pengaruh dalam melakukan prediksi rata-rata kecepatan, terutama seperti hari_Sunday, hari_Monday, hari_Friday, serta jam_9, jam_8, jam_7, dan jam_18 yang merupakan jam-jam sibuk di jalan-jalan tersebut.

Eksperimen telah dilakukan menggunakan berbagai model *machine learning* untuk mendapatkan jenis model yang paling optimal untuk melakukan prediksi rerata_kecepatan. Dari eksperimen yang telah dilakukan, model *Random Forest Regressor* (RF) cocok dengan data dengan *encoded feature* berbentuk biner dan ordinal sehingga menghasilkan performa prediksi yang lebih baik dibandingkan model lainnya dengan nilai *error* sMAPE sebesar 8.18181-8.18209 pada data *testing public* yang disediakan oleh panitia. Hasil *forecasting* menggunakan model ini juga menunjukkan bahwa model dapat menggeneralisasi pola dengan baik.

Daftar Pustaka

- [1] F. A. Pinagara and S. Khamtanet, “Will Jakarta Still Have Traffic Congestion after MRT?,” presented at the International Conference on Business and Management Research (ICBMR 2017), Atlantis Press, Nov. 2017, pp. 311–321. doi: 10.2991/icbmr-17.2017.29.
- [2] R. Sulistyorini, “HOW MUCH MONEY WILL LOSSES CAUSE OF ACCIDENT AND TRAFFIC JAM IN BANDAR LAMPUNG INDONESIA,” 2015. Accessed: Aug. 10, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/HOW-MUCH-MONEY-WILL-LOSSES-CAUSE-OF-ACCIDENT-AND-IN-Sulistyorini/135ea8b7acf848ccf20e9c8f4cef4e46d4872dff>
- [3] “The YouGov Sleep Study: Part one - Sleeping patterns | YouGov.” <https://yougov.co.uk/topics/health/articles-reports/2022/06/29/yougov-sleep-study-part-one-sleeping-patterns> (accessed Aug. 10, 2023).
- [4] J. Kim, J.-W. Noh, A. Kim, and Y. D. Kwon, “The Impact of Weekday-to-Weekend Sleep Differences on Health Outcomes among Adolescent Students,” *Children (Basel)*, vol. 9, no. 1, p. 52, Jan. 2022, doi: 10.3390/children9010052.
- [5] “North End Road Market,” *LBHF*, Nov. 25, 2015. <https://www.lbhf.gov.uk/business/markets/north-end-road-market> (accessed Aug. 10, 2023).
- [6] M. A. Iqbal, “Application of Regression Techniques with their Advantages and Disadvantages,” vol. 4, pp. 11–17, Sep. 2021.
- [7] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front Neurobot*, vol. 7, p. 21, Dec. 2013, doi: 10.3389/fnbot.2013.00021.
- [8] R. Couronné, P. Probst, and A.-L. Boulesteix, “Random forest versus logistic regression: a large-scale benchmark experiment,” *BMC Bioinformatics*, vol. 19, no. 1, p. 270, Jul. 2018, doi: 10.1186/s12859-018-2264-5.
- [9] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD ’16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [10] M. M. Taye, “Understanding of Machine Learning with Deep Learning: Architectures,

Workflow, Applications and Future Directions,” *Computers*, vol. 12, no. 5, Art. no. 5, May 2023, doi: 10.3390/computers12050091.

- [11] K. Taunk, P. Singh, and R. K. Behera, “Suicide Trend Analysis and Prediction in India using Facebook Prophet,” in *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2021, pp. 670–675.
- [12] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” *PeerJ Comput Sci*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.
- [13] R. Meenal, P. A. Michael, D. Pamela, and E. Rajasekaran, “Weather prediction using random forest machine learning model,” *IJECS*, vol. 22, no. 2, p. 1208, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1208-1215.