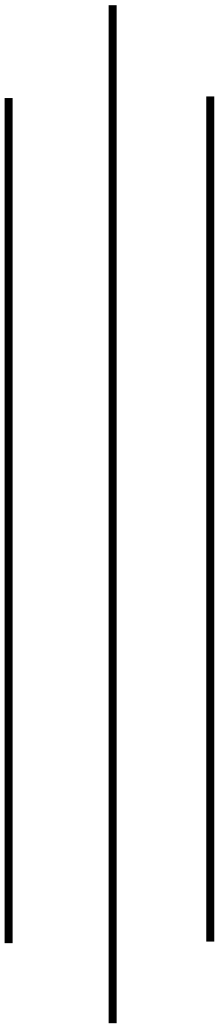
Time Series Forecasting Rata Rata Kecepatan Kendaraan Dengan Menggunakan Algoritma XGBoost



Mohamad Maulana Firdaus Ramadhan
Rizqika Mulia Pratama
Ikhwan Al Hakim

DAFTAR ISI

I. PENDAHULUAN

A. Latar Belakang

Time series forecasting adalah proses analisis data historis untuk memperkirakan nilai-nilai masa depan dari suatu variabel yang berubah seiring waktu. *Time series forecasting* dapat digunakan untuk berbagai tujuan, seperti perencanaan bisnis, peramalan cuaca, peramalan permintaan produk atau jasa, dan peramalan kondisi lalu lintas. Dengan menggunakan time series forecasting, pengelola transportasi dapat mengetahui pola-pola perubahan kecepatan rata-rata kendaraan di suatu jalan pada waktu-waktu tertentu, sehingga dapat merencanakan strategi-strategi untuk mengurangi kemacetan atau meningkatkan kelancaran lalu lintas.

Untuk melakukan *time series forcasting*, diperlukan data historis yang berkualitas dan cukup banyak untuk merepresentasikan karakteristik dari variabel yang diprediksi. Data historis kecepatan rata-rata kendaraan di suatu jalan dapat diperoleh dari berbagai sumber, seperti sensor lalu lintas, kamera CCTV, GPS, atau aplikasi navigasi online. Data tersebut kemudian dapat dianalisis dengan menggunakan berbagai metode *machine learning*, matematika, atau statistik untuk menghasilkan model *forcasting* yang akurat dan andal.

B. Tujuan

Penelitian ini memiliki tujuan untuk membuat model pembelajaran mesin berupa *time series forcasting* tentang kecepatan rata-rata kendaraan di suatu jalan dengan menggunakan algoritma XGBoost.

C. Manfaat

Penelitian ini memiliki manfaat diantaranya :

1. Mendukung pilihan rute perjalanan: Dengan mengetahui kecepatan rata-rata kendaraan di jalan tertentu, pengguna jalan dapat memilih rute yang paling optimal hemat waktu, dan nyaman. Hal ini dapat mengurangi kemacetan, stres, dan biaya transportasi. Salah satu contoh aplikasi yang menggunakan *time series forcasting* untuk kecepatan rata-rata kendaraan adalah Google Maps.
2. Membantu mengambil keputusan dalam hal manajemen lalu lintas, perencanaan transportasi, dan mitigasi polusi udara: Dengan mengetahui kecepatan rata-rata kecepatan rata-rata kendaraan di jalan tertentu, pihak berwenang dapat mengatur arus lalu lintas, mengalokasikan sumber daya, dan merancang kebijakan yang sesuai dengan kondisi jalan. Hal ini dapat meningkatkan efisiensi, keselamatan, dan kualitas lingkungan.

II. STUDI LITERATUR

A. XGBoost

XGBoost adalah salah satu algoritma pembelajaran mesin yang paling populer dan kuat untuk tugas-tugas seperti klasifikasi dan regresi. Ini adalah bentuk perbaikan dari algoritma gradient boosting, yang bekerja dengan membangun serangkaian pohon keputusan yang lemah secara berurutan, di mana setiap pohon berfokus pada mengoreksi kesalahan prediksi pohon sebelumnya. *XGBoost* mengambil langkah lebih jauh dengan menggabungkan keunggulan dari beberapa teknik, seperti regularisasi untuk mencegah *overfitting*, penanganan nilai yang hilang, dan kemampuan bekerja dengan berbagai jenis data.

B. RandomizedSearchCV

RandomizedSearchCV adalah teknik yang digunakan untuk mengoptimalkan performa model dengan mengevaluasi berbagai kombinasi *hyperparameter*. Saat kita membangun model pembelajaran mesin, terdapat berbagai parameter yang harus diatur, seperti kedalaman maksimum pohon keputusan atau tingkat pembelajaran. Alih-alih mencoba semua kombinasi yang mungkin, *RandomizedSearchCV* mengambil pendekatan acak dengan mengambil sampel nilai-nilai *hyperparameter* dari distribusi yang ditentukan sebelumnya. Ini memungkinkan kita untuk mengeksplorasi variasi *hyperparameter* dengan lebih cepat dan efisien, mencapai konfigurasi yang lebih baik dalam waktu yang lebih singkat.

C. Data Encoding

Data Encoding merupakan tahap pra-pemrosesan yang penting dalam *machine learning*. Ini merujuk pada proses mengubah data kategori atau teks menjadi format numerik, agar dapat digunakan sebagai masukan untuk algoritma-algoritma untuk diproses. Alasan dilakukannya pengodean adalah bahwa sebagian besar algoritma *machine learning* bekerja dengan angka dan bukan dengan teks atau variabel kategorikal. Dengan mengkodekan data, kita memungkinkan algoritma untuk menganalisis dan membuat prediksi berdasarkan informasi yang terkandung dalam data kategori atau teks.

D. Root Mean Square Error

RMSE adalah metrik penting dalam mengevaluasi kualitas model regresi. Ini mengukur sejauh mana prediksi model dari nilai sebenarnya, dan memberikan bobot lebih besar pada kesalahan yang lebih besar. Dengan mengkuadratkan setiap selisih antara prediksi dan nilai

sebenarnya, RMSE menyoroti kesalahan besar dan membantu mengidentifikasi potensi prediksi yang buruk atau outlier. Dengan nilai RMSE yang lebih rendah, model cenderung memiliki performa prediksi yang lebih baik.

III. PROSES DAN HASIL EKSPERIMEN

A. Exploratory Data Analysis

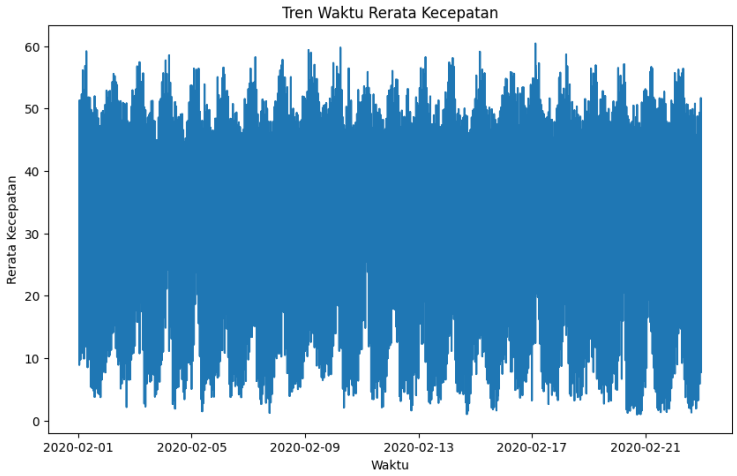
Dalam melakukan penelitian ini, digunakan dataset yang diperoleh dari Kaggle yang sudah disediakan, Data ini berisi rata-rata kecepatan kendaraan pada setiap jam dalam rentang 1 Februari 2020—29 Februari 2020. Original dataset ini berisi dua bagian, yaitu data train dan test. Data train terdiri dari 398648 baris dan 5 kolom yang digunakan untuk melatih model, sedangkan data tes terdiri dari dari 127489 baris dan 5 kolom yang digunakan untuk menghasilkan prediksi yang memiliki komponen sebagai berikut.

- waktu_setempat - waktu pencatatan kejadian
- id_jalan - id unik jalan saat perekaman data
- id_titik_mulai - id unik titik awal perekaman data
- id_titik_akhir - id unik titik akhir perekaman data
- rerata_kecepatan - target prediksi dari *task* ini, satuannya adalah km/jam

Analisis eksplorasi awal pada dataset ini mengungkapkan beberapa nilai unik untuk setiap detail:

- Jumlah nilai unik pada waktu_setempat: 527.
- Jumlah nilai unik pada id_jalan: 20.
- Jumlah nilai unik pada id_titik_mulai: 488.
- Jumlah nilai unik pada id_titik_akhir: 488.
- Jumlah nilai unik pada rerata_kecepatan: 29.023.

Selain itu, kami juga melakukan analisis distribusi pada variabel rerata_kecepatan pada setiap interval waktu:



Gambar 1. Distribusi rerata_kecepatan

B. Data Preprocessing

Tahap pra-pemrosesan data dimulai dengan menambahkan beberapa kolom baru, yaitu highway, maxspeed, latitude, dan longitude, yang diambil dari sumber eksternal, yakni OpenStreetMap. Penambahan data eksternal ini dilakukan karena semua entri 'id_jalan' mengandung atribut-atribut tersebut dari sumber tersebut.

Meskipun data 'highway' dan 'maxspeed' tidak memiliki nilai yang kosong, kolom 'latitude' dan 'longitude' memiliki beberapa nilai kosong. Untuk mengatasi kekosongan ini, nilai-nilai kosong tersebut diisi dengan menemukan titik terjauh yang tersedia . Berikut adalah penjelasan mengenai setiap kolom yang terdapat dalam data:

- 'highway': Kunci utama yang digunakan untuk mengidentifikasi berbagai jenis jalan, jalan raya, atau jalur.
- 'maxspeed': Batas kecepatan maksimum untuk suatu jalan.
- 'latitude': Koordinat garis bujur.
- 'longitude': Koordinat garis lintang.

Setelah langkah awal ini, data menjalani proses lebih lanjut. Pertama-tama, untuk memanfaatkan informasi temporal dalam kolom 'waktu_setempat', data tanggal dan waktu diubah menjadi format yang lebih terstruktur menggunakan fungsi pd.to_datetime(). Dari sini, berbagai atribut terkait waktu seperti hari, jam, dan bahkan hari dalam seminggu diekstraksi. Pemperkayaan ini memperkenalkan dimensi tambahan pada data, yang mungkin mengungkapkan pola atau tren berdasarkan periode waktu tertentu, seperti pola harian atau per jam.

Selanjutnya, pada kolom 'highway' dan 'maxspeed_km_jam', mengubah tipe data menjadi kategori dapat membantu dalam menghemat penggunaan memori dan mengoptimalkan efisiensi. Pendekatan ini memberikan manfaat pengurangan beban komputasi dan penggunaan memori yang lebih efisien.

C. Feature Engineering

Pada tahap Feature Engineering, fokus beralih untuk mempersiapkan data agar siap untuk dimasukkan ke dalam model. Pertama-tama, dilakukan Label Encoding pada kolom 'highway' dan 'maxspeed_km_jam' menggunakan LabelEncoder. Proses ini secara cerdas mengubah nilai-nilai kategori menjadi representasi numerik, memungkinkan model untuk mengolahnya dalam bentuk yang lebih mudah diinterpretasi.

Selanjutnya, dilakukan transformasi yang disebut one-hot encoding pada kolom-kolom kategorikal seperti 'id_jalan', 'month', dan 'day_of_week'. Pendekatan ini menghindari potensi bias yang mungkin muncul jika kita menggunakan representasi numerik untuk kategori. Representasi ini juga memungkinkan model untuk memahami perbedaan dan hubungan antara kategori dengan lebih baik.

Lebih lanjut, fitur-fitur masukan yang dibutuhkan untuk melatih model ditentukan dengan menghapus kolom target 'rerata_kecepatan'. Nama-nama kolom fitur disimpan dalam variabel 'input_features'. Sebaliknya, kolom 'rerata_kecepatan' ditetapkan sebagai variabel target yang akan menjadi fokus utama dalam prediksi.

D. Modeling

Dalam eksperimen ini, kami telah mengadopsi algoritma XGBoost sebagai model utama untuk memprediksi variabel target. Algoritma XGBoost dipilih karena memiliki keunggulan dalam hal akurasi, efisiensi, dan interpretabilitas dibandingkan dengan algoritma lain. Model pembelajaran mesin ini diharapkan dapat memberikan prediksi kecepatan rata-rata kendaraan yang akurat dan andal untuk membantu pengambilan keputusan dalam hal manajemen lalu lintas, perencanaan transportasi, dan mitigasi polusi udara. Algoritma ini dikenal efektif dalam tugas regresi dan klasifikasi berkat kemampuannya dalam mengatasi masalah non-linear dan kompleks. Selain itu, untuk memaksimalkan kinerja model, kami memanfaatkan teknik *hyperparameter tuning* menggunakan alat yang disebut RandomizedSearchCV.

Kami telah dengan cermat mendefinisikan rentang nilai spesifik untuk sejumlah *hyperparameter* yang relevan, seperti jumlah estimator, tingkat pembelajaran, kedalaman maksimum pohon, dan berat minimum untuk setiap *leaf* dalam pohon. Selain itu, kami telah

memperhitungkan *hyperparameter* subsampel dan beberapa *hyperparameter* lainnya dengan mendefinisikan distribusi yang sesuai.

Langkah selanjutnya dalam eksperimen ini melibatkan pembuatan model XGBoost menggunakan kelas XGBRegressor. Kami juga telah menetapkan nilai untuk parameter `random_state` guna memastikan hasil yang dapat direproduksi secara konsisten.

Untuk melakukan pencarian *hyperparameter*, kami menerapkan objek RandomizedSearchCV. Dalam objek ini, kami mengintegrasikan model XGBoost yang telah dibuat sebelumnya dan berbagai kombinasi *hyperparameter* dari `param_dist` yang telah ditentukan sebelumnya. Selain itu, kami menentukan parameter lain seperti jumlah iterasi (`n_iter`), skor evaluasi (`scoring`), dan skema validasi silang (`cv`).

Pendekatan ini memberi kami peluang untuk secara otomatis menemukan konfigurasi *hyperparameter* yang paling optimal untuk model XGBoost kami. Dengan memadukan algoritma XGBoost yang canggih dan pendekatan RandomizedSearchCV yang berfokus pada pencarian yang efisien, kami mampu meningkatkan kemampuan prediktif model kami secara signifikan. Dalam proses ini, kami berupaya menghasilkan model yang dapat menghadapi tantangan data yang kompleks dan memberikan hasil prediksi yang akurat.

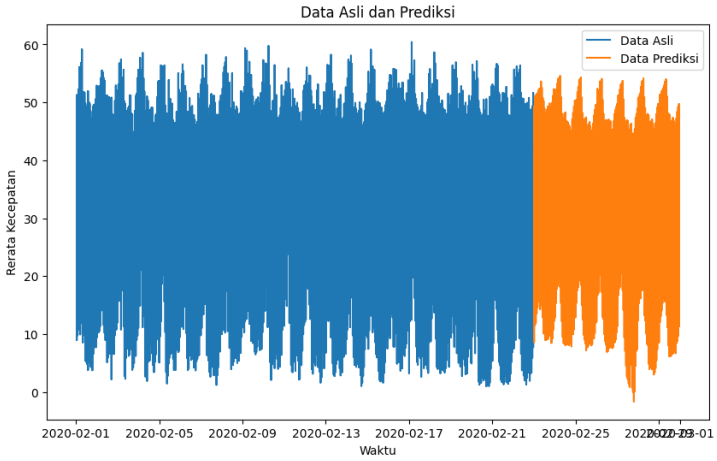
E. Evaluasi

Dalam tahap evaluasi, kami menggunakan metode validasi silang (cross validation) untuk mengukur performa model ramalan kami dengan lebih akurat. Metode ini membagi dataset menjadi beberapa bagian (fold) yang berbeda, dengan tujuan untuk melatih dan menguji model pada subset data yang berbeda-beda. Dalam konteks ini, metrik evaluasi yang kami pilih adalah RMSE.

RMSE adalah metrik evaluasi yang umum digunakan dalam konteks ramalan. Ia mengukur perbedaan antara nilai aktual dan nilai yang diprediksi oleh model. Nilai RMSE yang lebih rendah menunjukkan bahwa model cenderung lebih akurat dalam melakukan prediksi. Namun, penting untuk diingat bahwa interpretasi nilai RMSE perlu mempertimbangkan skala data asli. Semakin rendah RMSE relatif terhadap skala variabel target, semakin baik performa model tersebut.

Dalam eksperimen ini, setelah menerapkan cross validation dengan model XGBoost yang telah dikonfigurasi sebelumnya, kami memperoleh nilai RMSE sebesar 2.99. Ini berarti rata-rata perbedaan antara nilai ramalan dan nilai aktual adalah sekitar 2.99 pada skala variabel target yang relevan. Hasil ini memberikan gambaran tentang akurasi umum model kami dalam

meramalkan data yang belum pernah dilihat sebelumnya. Nilai RMSE yang lebih rendah adalah tujuan yang diinginkan dalam usaha untuk meningkatkan ketepatan dan keakuratan ramalan model. Berikut merupakan plot hasil prediksi.



Gambar 2. Plot Hasil Prediksi

F. Analisis Hasil Prediksi

Dalam analisis kami terhadap hasil prediksi yang dihasilkan oleh model kami, kami berhasil mengidentifikasi berbagai wawasan yang sangat berharga dari prediksi serta karakteristik dari model yang telah kami gunakan.

Salah satu aspek yang sangat menonjol adalah akurasi prediksi yang signifikan yang dicapai oleh model XGBoost yang kami terapkan. Dengan nilai RMSE sebesar 2.99, kami dapat menyimpulkan bahwa prediksi yang kami hasilkan memiliki perbedaan yang rendah dengan nilai aktual pada variabel target. Hal ini menunjukkan bahwa model kami memiliki kemampuan yang kuat dalam meramalkan data yang belum pernah dilihat sebelumnya. Keakuratan ini memberikan landasan yang kokoh untuk mengandalkan prediksi model dalam pengambilan keputusan yang lebih baik.

Selanjutnya, kami mengakui pentingnya proses penalaan *hyperparameter* dalam meningkatkan kinerja model kami. Dengan menerapkan teknik penalaan *hyperparameter* menggunakan RandomizedSearchCV, kami berhasil menemukan kombinasi *hyperparameter* yang optimal untuk model XGBoost kami. Hasil ini mencerminkan bagaimana pengoptimalan *hyperparameter* memiliki dampak positif yang signifikan pada kemampuan prediksi model kami. Teknik ini memungkinkan kami untuk mengesampingkan kemungkinan konfigurasi *hyperparameter* yang suboptimal, dan akhirnya membantu memaksimalkan performa model dalam meramalkan data dengan akurasi tertinggi.

Dengan demikian, analisis kami tidak hanya memperkuat keyakinan terhadap kapabilitas model dalam melakukan prediksi yang akurat, tetapi juga menegaskan pentingnya pendekatan yang cermat terhadap proses penalaan *hyperparameter*. Kesimpulan ini mengilustrasikan dedikasi kami dalam mencapai prediksi yang handal dan bermanfaat melalui pendekatan yang holistik dan terarah.

IV. KESIMPULAN

Dalam penelitian ini, kami berhasil menghasilkan prediksi dalam bentuk time series forecasting untuk rerata kecepatan menggunakan algoritma XGBoost, dengan mencapai tingkat akurasi yang dapat diukur menggunakan skor RMSE sebesar 2.99.

V. REFERENSI

[1]
Y. Zhang, Y. Zhang, and A. Haghani, “A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model,” *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 65–78, Jun. 2014, doi: https://doi.org/10.1016/j.trc.2013.11.011.

[2]
Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4. https://cran.ms.unimelb.edu.au/web/packages/xgboost/vignettes/xgboost.pdf

[3]
T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, pp. 785–794, 2016, doi: https://doi.org/10.1145/2939672.2939785.