

PREDIKSI TREN KASUS COVID-19 DI AMERIKA SERIKAT BERDASARKAN DATA MOBILITAS MANUSIA MENGGUNAKAN LINEAR REGRESSION DAN RANDOM FOREST

Henrie Rafi Ardiyanto, Ferdianto, Rifki Alkazim, Winda Kurnia Sari, Ken Ditha Tania

Sistem Informasi, Universitas Sriwijaya

Jalan Palembang-Prabumulih,

KM 32 Indralaya, Kabupaten Ogan Ilir, Sumatera Selatan, Indonesia

09031282328031@student.unsri.ac.id

ABSTRAK

Pandemi COVID-19 memberikan dampak signifikan terhadap berbagai aspek kehidupan, termasuk kesehatan dan mobilitas masyarakat. Perubahan mobilitas manusia memiliki keterkaitan erat dengan peningkatan atau penurunan kasus COVID-19, sehingga diperlukan metode prediksi yang akurat untuk mengetahui tren kasus dalam jangka pendek. Permasalahan dalam penelitian ini adalah bagaimana memanfaatkan data mobilitas manusia untuk memprediksi perubahan jumlah kasus COVID-19 secara efektif. Penelitian ini bertujuan untuk membangun dan membandingkan model prediksi menggunakan metode regresi. Metode yang digunakan adalah pendekatan kuantitatif dengan membangun dua model, yaitu *Linear Regression* dan *Random Forest Regressor*, menggunakan data time series kasus harian COVID-19 dan mobilitas manusia di Amerika Serikat. Evaluasi dilakukan menggunakan RMSE, MAE, dan R^2 . Hasil penelitian menunjukkan bahwa *Random Forest Regressor* memiliki kinerja lebih baik dibandingkan *Linear Regression* dalam memprediksi tren jangka pendek kasus COVID-19.

Kata kunci : covid-19, linear regression, mobilitas manusia, prediksi, random forest regressor

1. PENDAHULUAN

Pandemi COVID-19 menjadi salah satu krisis kesehatan global punya dampak yang signifikan pada berbagai aspek kehidupan manusia. Penyebaran virus yang cepat menyebabkan gangguan besar pada sistem kesehatan, aktivitas ekonomi, serta pola kehidupan masyarakat di berbagai negara. Pandemi juga memicu perubahan dalam aktivitas harian seperti bekerja, belajar dan kegiatan sosial. Perubahan tersebut tidak hanya berpengaruh pada perilaku individu, namun juga berdampak pada dinamika mobilitas manusia dan aktivitas perkotaan secara keseluruhan.[1]

Aktivitas Masyarakat yang berubah selama pandemi juga mempunyai dampak langsung pada pola mobilitas berbagai wilayah. Diterapkannya kebijakan pembatasan sosial untuk pengendalian penyebaran COVID-19 menyebabkan perubahan signifikan dalam aktivitas transportasi dan pergerakan masyarakat di wilayah perkotaan. Penelitian sebelumnya menunjukkan bahwa selama periode pembatasan aktivitas terjadi penurunan mobilitas transportasi serta pola perjalanan masyarakat di berbagai kota. Kondisi ini menunjukkan bahwa salah satu indikator penting untuk memahami dinamika aktivitas masyarakat selama pandemi serta dampaknya terhadap sistem perkotaan.[2]

Dengan adanya perkembangan teknologi digital, ketersediaan data mobilitas manusia telah membuka peluang baru dalam analisis penyebaran COVID-19. Data mobilitas tersebut memungkinkan peneliti untuk mengamati pola pergerakan populasi secara lebih rinci dan menganalisis hubungan mobilitas manusia dengan dinamika penyebaran COVID-19 di suatu wilayah.

Penelitian sebelumnya menunjukkan bahwa data mobilitas dapat digunakan untuk memodelkan pola interaksi manusia serta membantu memahami mekanisme penyebaran pandemi pada skala spasial dan temporal yang berbeda.[3]

Penelitian mengenai prediksi perkembangan COVID-19 terus berkembang dengan memanfaatkan berbagai pendekatan pemodelan berbasis data. Model prediksi biasanya menggunakan data kasus harian untuk menganalisis pola temporal penyebaran COVID-19 sehingga dapat memperkirakan jumlah kasus pada periode waktu mendatang. Penelitian sebelumnya menunjukkan faktor mobilitas manusia dan karakteristik sosial demografis memiliki pengaruh penting dalam memprediksi penyebaran COVID-19.[4] Sehingga, integrasi data mobilitas dalam model prediksi dapat membantu meningkatkan pemahaman mengenai dinamika penyebaran COVID-19.

Penelitian sebelumnya menunjukkan bahwa berbagai pendekatan *machine learning* telah banyak digunakan untuk memprediksi perkembangan kasus COVID-19 berdasarkan data historis yang tersedia.[5] Pendekatan tersebut memanfaatkan pola dari data kasus harian untuk memperkirakan tren penyebaran COVID-19 pada periode waktu tertentu. Namun penelitian yang mengintegrasikan data mobilitas manusia sebagai variabel tambahan dalam model prediksi masih perlu dikembangkan. Penelitian ini bertujuan menganalisis hubungan mobilitas manusia dan tren penyebaran COVID-19 serta membandingkan kinerja model prediksi *linear Regression* dan *Random Forest* dalam memprediksi jumlah kasus harian selama tujuh hari ke depan di Amerika Serikat.

2. TINJAUAN PUSTAKA

2.1. Studi Literatur

Terjadi pergeseran pada Penelitian terkait pemodelan pandemi secara signifikan dari model tradisional menuju pendekatan berbasis data. Dalam studi sebelumnya berbagai model *machine learning* digunakan untuk memprediksi penyebaran COVID-19 di Amerika Serikat dengan memanfaatkan data historis. Dari hasil penelitian menunjukkan bahwa integrasi variabel dinamis ke dalam model statistik mampu menurunkan nilai *Root Mean Square Error*(RMSE) secara signifikan dibandingkan dengan model transmisi klasik. [6] Studi ini menegaskan bahwa pendekatan berbasis data lebih adaptif dalam menangkap fluktuatif kasus harian yang tidak menentu.

Evaluasi performa berbagai model *Machine Learning* seperti *Linear regression* dan *Random Forest* dalam memprediksi kasus harian COVID-19 telah dilakukan melalui studi komparatif.[7] Dari penelitian menunjukkan bahwa *Linear Regression* memiliki hasil yang kompetitif dan memiliki keunggulan dalam hal interpretabilitas jangka pendek. Hasil penelitian juga menunjukkan *Random Forest* cenderung memberikan akurasi yang lebih tinggi karena kemampuannya dalam menangani variabel data yang lebih kompleks.

2.2. Machine Learning

Machine Learning adalah salah satu cabang utama dari kecerdasan buatan yang memungkinkan sistem komputer mempelajari pola pola rumit dari data historis secara otomatis. *Machine Learning* berguna untuk melakukan prediksi atau klasifikasi tanpa perlu instruksi pemrograman yang eksplisit. Algoritma *Machine Learning* punya keunggulan fleksibilitas dan kemampuan mengolah variabel-variabel kompleks yang memiliki hubungan non-linier dibanding dengan model statistik konvensional.[8] Dalam pemodelan Pandemi Teknik ini sangat efektif dalam mengidentifikasi indikator-indikator krusial dari dataset yang besar agar menghasilkan estimasi risiko tren lonjakan kasus dengan presisi yang tinggi.

2.3. Linear Regression

Linear Regression adalah metode analisis statistika yang digunakan memodelkan hubungan fungsional antara variabel dependen dengan satu atau lebih variabel independen melalui pendekatan *linear*. Secara sistematis *Linear Regression* berupaya menemukan garis estimasi terbaik yang meminimalkan jumlah kuadrat selisih antara nilai observasi dan nilai prediksi.[9] Metode ini menjadi landasan dalam analisis prediktif karena struktur modelnya yang transparan sehingga memungkinkan peneliti secara menginterpretasikan secara langsung koefisien dan variabel prediktor terhadap variabel respon secara efisien.

2.4. Contoh Persamaan Matematika

Random Forest Regression adalah algoritma *Machine Learning* berbasis *Ensemble Learning* yang bekerja dengan cara membangun sekumpulan besar pohon keputusan (*decision trees*) secara mandiri selama fase pelatihan. Kekuatan utama algoritma ini adalah pada kemampuannya dalam menggabungkan prediksi dari banyak pohon melalui mekanisme rata-rata guna meningkatkan akurasi dan stabilitas model secara keseluruhan. Dalam pemodelan pandemi, *Random Forest* sangat handal karena mampu menangkap hubungan non-linear yang kompleks antara berbagai variabel prediktor dan memitigasi risiko *overfitting* yang sering terjadi pada model pohon keputusan tunggal, sehingga menghasilkan estimasi yang lebih presisi pada data yang memiliki variansi tinggi.

3. METODE PENELITIAN

3.1. Pendekatan Penelitian

Penelitian ini menggunakan metode kuantitatif dengan pendekatan eksperimen komparatif. Metode kuantitatif digunakan karena penelitian ini berfokus pada pengolahan dan analisis data numerik untuk menghasilkan prediksi yang terukur dan objektif. Pendekatan eksperimen komparatif dilakukan dengan membangun dan membandingkan beberapa model *machine learning* untuk mengetahui model yang memiliki kinerja terbaik dalam melakukan prediksi terhadap data yang digunakan.

Dalam penelitian ini dilakukan proses pembangunan model, pelatihan model, serta pengujian model untuk melihat kemampuan masing-masing algoritma dalam memprediksi perubahan tren kasus COVID-19. Hasil dari setiap model kemudian dibandingkan berdasarkan performa yang dihasilkan untuk menentukan algoritma yang paling efektif dalam menghasilkan prediksi.

3.2. Data Penelitian

Data yang digunakan dalam penelitian ini berupa *data time series tabular numerik* yang berkaitan dengan jumlah kasus COVID-19 serta data mobilitas manusia di Amerika Serikat. Data time series dipilih karena memiliki karakteristik yang tersusun berdasarkan urutan waktu sehingga memungkinkan analisis terhadap pola perubahan kasus COVID-19 dari waktu ke waktu.

Dataset yang digunakan mencakup beberapa variabel utama, seperti jumlah kasus harian COVID-19 serta indikator mobilitas manusia yang menggambarkan aktivitas pergerakan masyarakat. Data mobilitas tersebut digunakan untuk melihat kemungkinan hubungan antara tingkat mobilitas manusia dengan perubahan jumlah kasus COVID-19.

3.3. Target Prediksi

Tujuan utama penelitian ini adalah memprediksi persentase perubahan rata-rata kasus harian COVID-19 dalam 7 hari berikutnya. Target prediksi tersebut

digunakan untuk melihat apakah tren kasus COVID-19 diperkirakan mengalami peningkatan atau penurunan dalam jangka waktu pendek.

Nilai persentase perubahan digunakan sebagai variabel target dalam proses pemodelan. Jika nilai prediksi bernilai positif, maka hal tersebut menunjukkan bahwa kasus COVID-19 diperkirakan mengalami peningkatan. Sebaliknya, jika nilai prediksi bernilai negatif, maka hal tersebut menunjukkan bahwa tren kasus COVID-19 diperkirakan mengalami penurunan.

3.4. Model Prediksi

Dalam penelitian ini digunakan dua algoritma regresi dalam machine learning, yaitu Linear Regression dan Random Forest Regressor. Linear Regression digunakan untuk memodelkan hubungan linear antara variabel independen dan variabel target sehingga dapat digunakan untuk memprediksi nilai numerik berdasarkan pola hubungan yang terbentuk dalam data.

Selain itu, penelitian ini juga menggunakan algoritma Random Forest Regressor yang merupakan metode ensemble learning berbasis decision tree. Algoritma ini bekerja dengan membangun banyak pohon keputusan dan menggabungkan hasil prediksi dari setiap pohon untuk menghasilkan prediksi akhir yang lebih stabil dan akurat. Penggunaan kedua algoritma ini bertujuan untuk membandingkan performa model sederhana dengan model yang lebih kompleks dalam memprediksi tren perubahan kasus COVID-19.

3.5. Evaluasi Model

Setelah proses pelatihan dan pengujian model dilakukan, tahap selanjutnya adalah evaluasi performa model untuk mengetahui tingkat akurasi prediksi yang dihasilkan. Evaluasi dilakukan dengan menggunakan beberapa metrik pengukuran yang umum digunakan dalam permasalahan regresi, yaitu Root Mean Square Error (RMSE), Mean Absolute Error (MAE), dan Squared Correlation (R^2).

Metrik evaluasi tersebut digunakan untuk mengukur seberapa baik model dalam memprediksi nilai target dibandingkan dengan nilai aktual pada dataset. Hasil evaluasi kemudian digunakan untuk membandingkan performa antara algoritma Linear Regression dan Random Forest Regressor sehingga dapat diketahui model yang memiliki kinerja terbaik dalam memprediksi tren perubahan kasus COVID-19.

4. HASIL DAN PEMBAHASAN

Pada bagian ini disajikan hasil *preprocessing* data, pembentukan target prediksi, pengujian model *Linear Regression* dan *Random Forest Regressor*, serta analisis hubungan antar variabel menggunakan matriks korelasi. Kinerja model dievaluasi menggunakan *root mean squared error* (RMSE), *mean absolute error* (MAE), dan *squared correlation*.

4.1. Hasil Preprocessing Data

Pada tahap awal penelitian, data kasus COVID-19 dan mobilitas manusia terlebih dahulu melalui proses *preprocessing* agar siap digunakan dalam pemodelan. Tahap ini penting karena penelitian terdahulu menunjukkan bahwa indikator mobilitas memiliki keterkaitan dengan dinamika kasus COVID-19 dan dapat berperan sebagai indikator awal dalam pemodelan prediktif, sehingga kualitas struktur data, urutan waktu, dan pemilihan atribut menjadi sangat menentukan hasil analisis.[6] [8] [11] Oleh karena itu, *preprocessing* pada penelitian ini dilakukan untuk menghasilkan dataset yang lebih bersih, terstruktur, dan sesuai untuk analisis *time series* berbasis regresi.

4.2. Workflow Processing



Gambar 1. Workflow processing

Proses dimulai dari pengambilan dataset, kemudian dilakukan penyaringan data berdasarkan wilayah penelitian, seleksi atribut yang relevan, agregasi data harian, pengurutan berdasarkan tanggal, penanganan missing value, dan penamaan ulang atribut. Langkah-langkah ini dilakukan untuk menjaga konsistensi deret waktu serta memastikan bahwa variabel mobilitas yang digunakan benar-benar dapat dianalisis bersama data kasus harian. Pendekatan ini sejalan dengan temuan penelitian sebelumnya yang menekankan bahwa hubungan antara mobilitas dan kasus COVID-19 perlu dibaca dalam kerangka temporal yang terstruktur, karena pengaruh mobilitas sering muncul dengan pola yang tidak sederhana dan dapat melibatkan efek jangka pendek maupun jangka panjang.[7] [11]

4.3. Preprocessing Data

Row No.	date	daily_cases	grocery_p	parks	residential	retail_m	transit	workplaces	day_of_week
1	Jan 1, 2021	0	-31.329	2.957	24.088	-45.538	-53.024	-71.429	4
2	Jan 2, 2021	52190	-15.240	-25.280	10.955	-30.731	-36.119	-24.833	5
3	Jan 3, 2021	35909	-16.256	-15.583	8.238	-29.098	-30.128	-18.889	6
4	Jan 4, 2021	40721	-11.280	-31.932	13.873	-26.462	-39.349	-35.143	0
5	Jan 5, 2021	37284	-7.408	-15.800	12.837	-24.577	-36.419	-34.393	1
6	Jan 6, 2021	39904	-13.143	-25.844	14.143	-29.558	-40.605	-34.838	2
7	Jan 7, 2021	40420	-12.720	-19.844	14.102	-29.431	-39	-34.768	3
8	Jan 8, 2021	50071	-14.080	-26	14.531	-32.250	-40.977	-32.893	4
9	Jan 9, 2021	47229	-11.479	-18.380	9.250	-30.615	-37.310	-17.352	5
10	Jan 10, 2021	42402	-13.822	-9.917	7.738	-28.365	-36	-18.352	6
11	Jan 11, 2021	40624	-12.940	-19.884	12.286	-29.078	-38.744	-32.036	0
12	Jan 12, 2021	45883	-12.020	-22.209	13.306	-30.176	-38.326	-34.107	1
13	Jan 13, 2021	41674	-13.694	-16.690	12.939	-29.902	-39.488	-34.268	2
14	Jan 14, 2021	40715	-14.100	-17.744	12.816	-30.404	-39.163	-34.768	3
15	Jan 15, 2021	39102	-14.694	-17.761	13.020	-32.269	-38.762	-33.286	4
16	Jan 16, 2021	37587	-12.380	-2.796	7.750	-29.615	-33.220	-19.537	5
17	Jan 17, 2021	32703	-13.556	-18.773	6.071	-24.500	-31.684	-19.370	6
18	Jan 18, 2021	25383	-13.122	-16.762	13.653	-23.885	-39	-47.446	0
19	Jan 19, 2021	16104	-13.493	-20.517	-13.304	-16.673	-17.376	-14.444	1

Gambar 2. Preprocessing data

Setelah *preprocessing* dilakukan, diperoleh dataset yang lebih terstruktur dan siap untuk tahap pembentukan target prediksi. Data yang semula masih berada pada level *county* berhasil diagregasi menjadi data harian pada level negara bagian, sehingga setiap baris merepresentasikan satu tanggal pengamatan.

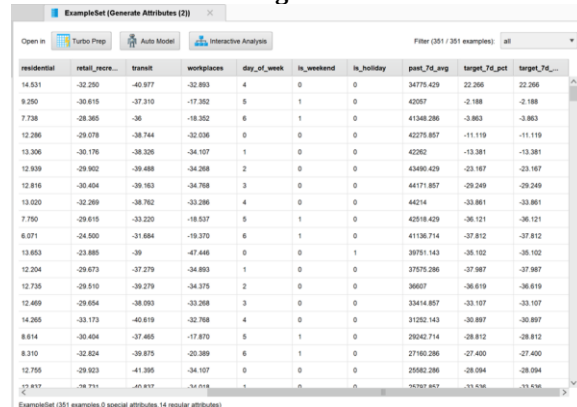
Selain itu, penanganan *missing value* dan penamaan ulang atribut menghasilkan data yang lebih rapi dan mudah diinterpretasikan pada tahap analisis berikutnya. Hasil *preprocessing* ini penting karena penelitian terdahulu menunjukkan bahwa kualitas representasi data masukan, termasuk indikator mobilitas dan kovariat temporal, berpengaruh terhadap kemampuan model dalam mempelajari pola penyebaran COVID-19 secara lebih akurat.[6] [8]

Pada tahap *preprocessing*, data difilter pada level negara bagian, kemudian dilakukan agregasi harian, pengurutan berdasarkan tanggal, penanganan *missing value*, dan penamaan ulang atribut. Hasil *preprocessing* menghasilkan dataset time series harian yang siap digunakan untuk pembentukan target prediksi. Setelah diperoleh dataset hasil *preprocessing* yang telah bersih dan terstruktur, tahap selanjutnya adalah pembentukan atribut target yang merepresentasikan persentase perubahan rata-rata kasus harian COVID-19 untuk 7 hari berikutnya.

4.4. Pembentukan Target Prediksi

Setelah diperoleh dataset hasil *preprocessing* yang telah bersih dan terstruktur, tahap selanjutnya adalah pembentukan target prediksi. Pada penelitian ini, target tidak didefinisikan sebagai jumlah kasus harian mentah, melainkan sebagai persentase perubahan rata-rata kasus harian COVID-19 untuk 7 hari berikutnya. Pendekatan ini dipilih karena penelitian terdahulu menunjukkan bahwa hubungan antara mobilitas dan kasus COVID-19 tidak selalu muncul secara langsung pada hari yang sama, tetapi sering terlihat sebagai perubahan tren dalam horizon waktu tertentu, sehingga pendekatan berbasis jendela waktu lebih relevan untuk analisis prediktif.[8] [11]

4.5. Pembentukan Target



residential	retail_merc...	transit	workplaces	day_of_week	is_weekend	is_holiday	past_7d_avg	target_7d_avg	target_7d_p...
14.531	-32.250	-40.977	-32.893	4	0	0	34775.429	22.266	22.266
8.250	-30.616	-37.310	-17.352	5	1	0	42087	-2.188	-2.188
7.738	-28.365	-36	-18.352	6	1	0	41348.286	-3.863	-3.863
12.286	-29.078	-38.744	-32.036	0	0	0	42270.857	-11.119	-11.119
13.306	-30.176	-38.326	-34.107	1	0	0	42262	-13.381	-13.381
12.939	-29.802	-39.468	-34.268	2	0	0	43480.429	-23.167	-23.167
12.816	-30.494	-39.163	-34.768	3	0	0	44171.857	-29.249	-29.249
13.020	-32.289	-38.762	-32.286	4	0	0	44214	-33.861	-33.861
7.790	-29.616	-33.220	-18.507	5	1	0	42518.429	-36.121	-36.121
6.071	-24.500	-31.884	-19.370	6	1	0	41136.714	-37.812	-37.812
13.653	-23.885	-39	-47.446	0	0	1	39751.143	-38.102	-38.102
12.204	-29.673	-37.279	-34.893	1	0	0	37575.286	-37.987	-37.987
12.735	-29.510	-39.279	-34.375	2	0	0	36607	-36.619	-36.619
12.469	-29.654	-38.093	-33.268	3	0	0	33414.857	-33.107	-33.107
14.265	-33.173	-40.619	-32.768	4	0	0	31252.143	-30.897	-30.897
8.614	-30.404	-37.485	-17.870	5	1	0	29242.714	-28.812	-28.812
8.310	-32.824	-39.875	-20.389	6	1	0	27160.286	-27.400	-27.400
12.756	-29.923	-41.395	-34.107	0	0	0	25682.286	-28.094	-28.094
27.817	-19.711	-40.817	-14.016	1	0	0	16787.857	-13.536	-13.536

Gambar 3. Pembentukan target

Pembentukan target dilakukan dengan terlebih dahulu menghitung *past_7d_avg*, yaitu rata-rata kasus harian selama 7 hari sebelumnya, serta *future_7d_avg*, yaitu rata-rata kasus harian selama 7 hari berikutnya. Selanjutnya, target prediksi dihitung sebagai persentase perubahan antara rata-rata kasus pada 7 hari berikutnya terhadap rata-rata 7 hari sebelumnya. Definisi ini digunakan agar model dapat menangkap

arah dan besarnya perubahan tren kasus, bukan sekadar memprediksi nilai kasus harian tunggal yang cenderung lebih fluktuatif. Pendekatan semacam ini sejalan dengan literatur yang menekankan bahwa indikator mobilitas dan dinamika epidemi lebih tepat dianalisis dalam kerangka temporal yang mempertimbangkan efek jeda dan perubahan tren, bukan hanya hubungan linier harian yang bersifat langsung.[7] [8] [11]

Target penelitian dibentuk sebagai persentase perubahan rata-rata kasus harian COVID-19 selama 7 hari berikutnya terhadap rata-rata 7 hari sebelumnya. Pendekatan ini digunakan agar model tidak hanya memprediksi jumlah kasus mentah, tetapi juga arah serta besar perubahan tren kasus. Setelah atribut target berhasil dibentuk, dataset kemudian digunakan pada tahap pemodelan dengan membandingkan kinerja *Linear Regression* dan *Random Forest Regressor*.

4.6. Hasil Pengujian Linear Regression

Setelah atribut target berhasil dibentuk, tahap selanjutnya adalah melakukan pengujian model prediksi menggunakan algoritma *Linear Regression*. Model ini digunakan sebagai baseline karena memiliki mekanisme yang sederhana, mudah diinterpretasikan, dan dapat menunjukkan arah hubungan antarvariabel terhadap target prediksi. Pada penelitian ini, evaluasi model dilakukan menggunakan metrik *root mean squared error* (RMSE), *absolute error* (MAE), dan *squared correlation* untuk melihat kemampuan model dalam memprediksi persentase perubahan rata-rata kasus harian COVID-19 untuk 7 hari berikutnya.

4.7. Workflow Linear Regression



Gambar 4. Workflow linear regression

Setelah atribut target berhasil dibentuk, tahap selanjutnya adalah melakukan pengujian model prediksi menggunakan algoritma *Linear Regression*. Model ini digunakan sebagai baseline karena memiliki mekanisme yang sederhana, mudah diinterpretasikan, dan dapat menunjukkan arah hubungan antarvariabel terhadap target prediksi. Pada penelitian ini, evaluasi model dilakukan menggunakan metrik *root mean squared error* (RMSE), *absolute error* (MAE), dan *squared correlation* untuk melihat kemampuan model dalam memprediksi persentase perubahan rata-rata kasus harian COVID-19 untuk 7 hari berikutnya.

4.8. Hasil Evaluasi Linear Regression



PerformanceVector
PerformanceVector:
root_mean_squared_error: 38.089 +/- 0.000
absolute_error: 27.343 +/- 26.516
squared_correlation: 0.882

Gambar 5. Hasil evaluasi linear regression

Berdasarkan hasil pengujian, model *Linear Regression* menghasilkan RMSE sebesar 70,929, MAE sebesar 53,012, dan *squared correlation* sebesar 0,045. Nilai RMSE dan MAE yang relatif besar menunjukkan bahwa rata-rata kesalahan prediksi model masih cukup tinggi. Sementara itu, nilai *squared correlation* yang sangat rendah menunjukkan bahwa model linear hanya mampu menjelaskan sebagian kecil variasi target prediksi. Hasil ini mengindikasikan bahwa hubungan antara variabel mobilitas, riwayat kasus, dan target perubahan tren kasus 7 hari kedepan belum dapat ditangkap dengan baik oleh model linear sederhana.

Rendahnya performa *Linear Regression* pada penelitian ini dapat dipahami karena hubungan antara mobilitas manusia dan dinamika kasus COVID-19 tidak selalu bersifat linear maupun langsung. Beberapa penelitian menunjukkan bahwa indikator mobilitas sering berperan sebagai leading indicators dengan efek jeda waktu tertentu, serta dapat dipengaruhi oleh hubungan timbal balik antara perubahan mobilitas dan kondisi epidemiologis di masyarakat.[7] [8] [11] Dalam kondisi seperti ini, model linear cenderung memiliki keterbatasan karena mengasumsikan hubungan yang relatif sederhana, padahal pola penyebaran kasus COVID-19 dapat terbentuk dari interaksi yang lebih kompleks dan non-linear.

Oleh karena itu, pada tahap berikutnya dilakukan pengujian menggunakan *Random Forest Regressor* untuk melihat apakah model non-linear berbasis *ensemble* mampu memberikan hasil prediksi yang lebih baik dibandingkan *Linear Regression*.

4.9. Hasil Pengujian Random Forest

Setelah dilakukan pengujian menggunakan *Linear Regression*, tahap berikutnya adalah evaluasi model *Random Forest Regressor*. Model ini dipilih karena mampu menangani hubungan non-linear dan interaksi antar variabel yang lebih kompleks dibandingkan model linear sederhana. Dalam konteks prediksi COVID-19, beberapa penelitian menunjukkan bahwa pendekatan berbasis *Random Forest* atau model non-linear lain dapat memberikan hasil yang lebih baik ketika data dipengaruhi banyak faktor dan pola temporal yang tidak sederhana.[12] [15]

4.10. Workflow Random Forest



Gambar 6. Workflow random forest

Subproses pelatihan dan pengujian model *Random Forest* menggunakan *Split Validation* pada RapidMiner. Pada tahap pelatihan, data training digunakan untuk membentuk model *Random Forest*, sedangkan pada tahap pengujian model diterapkan pada data *testing* menggunakan operator *Apply Model*.

Selanjutnya, hasil prediksi dievaluasi melalui *operator Performance (Regression)* untuk memperoleh nilai RMSE, MAE, dan *squared correlation*. Prosedur ini dilakukan agar kinerja model dapat diukur pada data yang tidak digunakan selama proses pembelajaran.

4.11. Hasil Evaluasi Random Forest



Gambar 7. Hasil evaluasi random forest

Setelah dilakukan penanganan *outlier* pada target melalui proses *capping*, performa *Random Forest* meningkat secara signifikan. Model terbaik menghasilkan RMSE sebesar 25,291, MAE sebesar 15,776, dan *squared correlation* sebesar 0,968. Nilai tersebut menunjukkan bahwa model mampu memprediksi target dengan kesalahan yang jauh lebih kecil dibandingkan pengujian awal, serta memiliki kemampuan yang sangat tinggi dalam menjelaskan variasi target pada data uji. Hasil ini memperlihatkan bahwa kestabilan rentang target berpengaruh besar terhadap performa model regresi, terutama ketika target didefinisikan sebagai persentase perubahan yang rentan menghasilkan nilai ekstrim.

4.12. Correlation Matrix

Setelah dilakukan *preprocessing* dan pembentukan target prediksi, analisis dilanjutkan dengan melihat hubungan antar variabel menggunakan *correlation matrix*. Analisis ini dilakukan untuk memberikan gambaran awal mengenai kekuatan dan arah hubungan linear antara variabel mobilitas, riwayat kasus, dan target prediksi. Langkah ini relevan karena berbagai penelitian menunjukkan bahwa mobilitas manusia merupakan salah satu komponen penting dalam analisis dinamika COVID-19, meskipun hubungan tersebut tidak selalu sederhana dan tidak selalu bersifat langsung pada hari yang sama.[6] [7] [8]

4.13. Correlation Matrix

Attributes	daily_cases	grocery_pharmacy	parks	residential	retail_recreation	transit	workplaces	past_7d_avg	target_7d_pct_cap
daily_cases	1	-0.294	-0.370	0.398	-0.445	-0.504	-0.188	0.833	-0.099
grocery_pharmacy	-0.294	1	0.574	-0.453	0.844	0.556	0.021	-0.473	0.344
parks	-0.370	0.574	1	-0.491	0.663	0.737	0.041	-0.405	0.200
residential	0.398	-0.453	-0.491	1	-0.622	-0.823	-0.814	0.395	-0.193
retail_recreation	-0.445	0.844	0.663	-0.622	1	0.750	0.201	-0.589	0.380
transit	-0.504	0.556	0.737	-0.823	0.750	1	0.496	-0.560	0.273
workplaces	-0.188	0.021	0.041	-0.814	0.201	0.496	1	-0.128	-0.045
past_7d_avg	0.833	-0.473	-0.405	0.395	-0.589	-0.560	-0.128	1	-0.260
target_7d_pct_cap	-0.099	0.344	0.200	-0.193	0.380	0.273	-0.045	-0.260	1

Gambar 8. Correlation matrix

Matriks korelasi Pearson antara variabel utama pada dataset hasil preprocessing. Berdasarkan hasil tersebut, *daily_cases* memiliki korelasi positif yang kuat dengan *past_7d_avg*, yaitu sebesar 0,833. Hasil ini menunjukkan bahwa riwayat kasus pada periode

sebelumnya memiliki keterkaitan yang tinggi dengan kondisi kasus saat ini. Selain itu, beberapa variabel mobilitas seperti *retail_recreation* dan *grocery_pharmacy* menunjukkan korelasi positif terhadap target, masing-masing sebesar 0,380 dan 0,344. Variabel transit juga memiliki korelasi positif sebesar 0,273, sedangkan *residential* menunjukkan korelasi negatif lemah sebesar -0,193. Sementara itu, *workplaces* hanya menunjukkan korelasi yang sangat lemah terhadap target, yaitu -0,045.

Berdasarkan perbandingan hasil, *Random Forest* unggul pada seluruh matrik evaluasi. Korelasi antar variabel juga menunjukkan bahwa tidak semua fitur memiliki hubungan linear yang kuat dengan target, sehingga pendekatan non-linear lebih sesuai dibandingkan model linear sederhana.

Hasil *correlation matrix* menunjukkan bahwa tidak semua variabel memiliki hubungan linear yang kuat dengan target prediksi. Temuan ini penting karena membantu menjelaskan mengapa model *Linear Regression* pada penelitian ini menghasilkan performa yang lebih rendah dibandingkan *Random Forest*. Literatur sebelumnya menunjukkan bahwa hubungan antara mobilitas dan kasus COVID-19 sering muncul sebagai pola yang tertunda, saling mempengaruhi, dan dapat berubah antar periode, sehingga pendekatan linear sederhana sering kali tidak cukup untuk merepresentasikan keseluruhan pola hubungan tersebut.[7] [8] Di sisi lain, korelasi yang lebih tinggi pada beberapa variabel mobilitas tetap menunjukkan bahwa informasi mobilitas memiliki kontribusi dalam menjelaskan perubahan tren kasus, sehingga tetap relevan digunakan sebagai variabel prediktor.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa data mobilitas manusia memiliki potensi yang baik untuk digunakan sebagai variabel prediktor dalam memprediksi tren jangka pendek kasus COVID-19 di Amerika Serikat dengan pendekatan regresi pada data time series tabular numerik. Tahapan *preprocessing* yang meliputi seleksi atribut, agregasi data harian pada level negara bagian, penanganan *missing value*, pembentukan atribut target, serta penyesuaian nilai ekstrim pada target, berhasil menghasilkan dataset yang sesuai untuk proses pemodelan. Hasil pengujian menunjukkan bahwa *Random Forest Regressor* memberikan performa yang lebih unggul dibandingkan *Linear Regression*. *Linear Regression* menghasilkan RMSE sebesar 70,929, MAE sebesar 53,012, dan *squared correlation* sebesar 0,045, sedangkan *Random Forest* terbaik menghasilkan RMSE sebesar 25,291, MAE sebesar 15,776, dan *squared correlation* sebesar 0,968. Temuan ini menunjukkan bahwa pola hubungan antara variabel mobilitas dan perubahan tren kasus COVID-19 cenderung lebih tepat direpresentasikan menggunakan model non-linear dibandingkan model linear sederhana.

Berdasarkan capaian penelitian ini, pengembangan model prediksi berbasis data mobilitas manusia dengan algoritma *Random Forest Regressor* layak untuk dilanjutkan pada penelitian berikutnya. Pengembangan tersebut dapat diarahkan pada perluasan cakupan data, integrasi variabel pendukung yang relevan, serta eksplorasi metode prediksi lain sebagai pembanding untuk memperoleh model yang semakin optimal, adaptif, dan aplikatif. Dengan demikian, hasil penelitian ini tidak hanya memberikan kontribusi pada pengembangan metode prediksi berbasis machine learning, tetapi juga membuka peluang pemanfaatan model sebagai instrumen analisis tren kasus penyakit menular dalam mendukung pengambilan keputusan berbasis data.

DAFTAR PUSTAKA

- [1] D. Li and J. Lasenby, "Investigating impacts of COVID-19 on urban mobility and emissions," *Cities*, vol. 135, no. July 2022, p. 104246, 2023, doi: 10.1016/j.cities.2023.104246D. Li and J. Lasenby, "Investigating impacts of COVID-19 on urban mobility and emissions," *Cities*, vol. 135, p. 104246, 2023, doi: 10.1016/j.cities.2023.104246.
- [2] M. A. Bouzaghrane, H. Obeid, M. González, and J. Walker, "Human mobility reshaped? Deciphering the impacts of the Covid-19 pandemic on activity patterns, spatial habits, and schedule habits," *EPJ Data Science*, vol. 13, art. no. 24, 2024, doi: 10.1140/epjds/s13688-024-00463-4.
- [3] F. Delussu, M. Tizzoni, and L. Gauvin, "The limits of human mobility traces to predict the spread of COVID-19: A transfer entropy approach," *PNAS Nexus*, vol. 2, no. 10, pp. 1–12, 2023, doi: 10.1093/pnasnexus/pgad302.
- [4] M. Qiao and B. Huang, "COVID-19 spread prediction using socio-demographic and mobility-related data," *Cities*, vol. 138, p. 104360, 2023, doi: 10.1016/j.cities.2023.104360.
- [5] Y. Cheng, R. Cheng, T. Xu, X. Tan, and Y. Bai, "Machine Learning Techniques Applied to COVID-19 Prediction: A Systematic Literature Review," *Bioengineering*, vol. 12, no. 5, p. 514, 2025, doi: 10.3390/bioengineering12050514.
- [6] F. Benita, "Human mobility behavior in COVID-19: A systematic literature review and bibliometric analysis," *Sustainable Cities and Society*, vol. 70, p. 102916, 2021, doi: 10.1016/j.scs.2021.102916.
- [7] D. Boto-García, "Investigating the two-way relationship between mobility flows and COVID-19 cases," *Economic Modelling*, vol. 118, p. 106083, 2023, doi: 10.1016/j.econmod.2022.106083.
- [8] W.-H. Chiang, X. Liu, and G. Mohler, "Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates," *International Journal of Forecasting*, vol. 38, no.

- 2, pp. 505–520, 2022, doi: 10.1016/j.ijforecast.2021.07.001.
- [9] S. Ibarra-Espinosa, E. D. de Freitas, K. Ropkins, F. Dominici, and A. Rehbein, “Negative-Binomial and quasi-poisson regressions between COVID-19, mobility and environment in São Paulo, Brazil,” *Environmental Research*, vol. 204, p. 112369, 2022, doi: 10.1016/j.envres.2021.112369.
- [10] P. Wulandari, Andika, and H. Halide, “Determining significant factors associated with daily COVID-19 cases within three social distancing regimes,” *Gaceta Sanitaria*, vol. 35, suppl. 2, pp. S455–S458, 2021, doi: 10.1016/j.gaceta.2021.07.024.
- [11] F. Zargari, N. Aminpour, M. A. Ahmadian, A. Samimi, and S. Saidi, “Impact of mobility on COVID-19 spread – A time series analysis,” *Transportation Research Interdisciplinary Perspectives*, vol. 13, p. 100567, 2022, doi: 10.1016/j.trip.2022.100567.
- [12] J. Galasso, D. M. Cao, and R. Hochberg, “A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data,” *Chaos, Solitons & Fractals*, vol. 156, p. 111779, 2022, doi: 10.1016/j.chaos.2021.111779.
- [13] R. Li, Y. Song, H. Qu, M. Li, and G.-P. Jiang, “A data-driven epidemic model with human mobility and vaccination protection for COVID-19 prediction,” *Journal of Biomedical Informatics*, vol. 149, p. 104571, 2024, doi: 10.1016/j.jbi.2023.104571.
- [14] H. Du, E. Dong, H. S. Badr, M. E. Petrone, N. D. Grubaugh, and L. M. Gardner, “Incorporating variant frequencies data into short-term forecasting for COVID-19 cases and deaths in the USA: a deep learning approach,” *eBioMedicine*, vol. 89, p. 104482, 2023, doi: 10.1016/j.ebiom.2023.104482.
- [15] F. Özen, “Random forest regression for prediction of Covid-19 daily cases and deaths in Turkey,” *Heliyon*, vol. 10, no. 4, p. e25746, 2024, doi: 10.1016/j.heliyon.2024.e25746.