Jurnal Rekayasa Sistem Informasi dan Teknologi
Volume 3, No 2 November 2025
e-ISSN : 3025-888X

PREDIKSI KECEPATAN RATA-RATA BERSEPEDA BERDASARKAN KONDISI TOPOGRAFI DAN FAKTOR CUACA MENGGUNAKAN XGBOOST DARI DATA STRAVA

Rifqy Ramdhani Hakim[1], Sam Farisa Chaerul Haviana[2]
[1,2]Teknik Informatika, Universitas Islam Sultan Agung, Indonesia
rifqyramdhanih@gmail.com[1], sam@unissula.ac.id[2]

Received: 05-10-2025	Revised: 25-10-2025	Approved: 05-10-2025

ABSTRAK

Meningkatnya minat bersepeda dan penggunaan aplikasi perekam data seperti Strava menuntut pemahaman mendalam mengenai faktor yang memengaruhi performa. Kecepatan rata-rata sangat dipengaruhi oleh variabel non-linear seperti kondisi topografi dan cuaca, sehingga memerlukan model prediksi yang akurat. Tujuan penelitian ini adalah mengimplementasikan Extreme Gradient Boosting (XGBoost) untuk membangun model prediksi kecepatan rata-rata bersepeda. Penelitian berfokus memodelkan hubungan antara data historis aktivitas Strava dengan variabel lingkungan. Metode penelitian dimulai dari pengumpulan data aktivitas pribadi (Juni 2024 - Agustus 2025) , mencakup fitur jarak, elevasi, cuaca, dan waktu tidur. Data mentah melalui pra-pemrosesan, termasuk normalisasi Min-Max Scaler. Data dibagi menjadi 80% data latih dan 20% data uji. Model XGBRegressor dilatih dengan hyperparameter seperti n_estimators=300 dan learning_rate=0.2. Kinerja model dievaluasi menggunakan Root Mean Squared Error (RMSE) dan R-squared (R^2). Hasilnya, model XGBoost mampu memberikan estimasi kecepatan dengan akurasi cukup baik. Model mencapai skor RMSE 1.240 km/jam , yang mengindikasikan rata-rata kesalahan prediksi. Selain itu, model memperoleh nilai R^2 sebesar 0.800. Nilai R^2 ini berarti model mampu menjelaskan 80% variasi data kecepatan. Kesimpulannya, model XGBoost terbukti representatif.

***Kata Kunci**: Sepeda, Strava, Topografi, Cuaca, XGBoost*

PENDAHULUAN

Kegiatan bersepeda merupakan kegiatan atau aktifitas yang cukup banyak diminati masyarakat sebagai pilihan olahraga. Sepeda adalah alat transportasi yang sangat umum dan luas penggunaannya di dunia yang digunakan oleh semua orang dari berbagai kalangan usia (Sitorus, 2024). Tidak hanya sebagai alat transportasi sepeda juga menjadi salah satu hobi yang banyak diminati oleh masyarakat dan sudah merupakan gaya hidup bagi sebagian masyarakat kota (Utomo, 2016). Saat ini olahraga sepeda menjadi sebuah kebutuhan utama dari setiap orang, bukan hanya untuk presatasi, namun juga untuk kepentingan kesehatan, untuk kepentingan hiburan atau rekreasi, untuk kepentingan mencari relasi dan berbagai kepentigan lainnya (Surojo et al., 2022).

Seiring meningkatnya minat terhadap gaya hidup sehat dan berbasis teknologi, aplikasi seperti Strava semakin banyak digunakan oleh pesepeda (Irawati et al., 2024). Strava memungkinkan penggunanya untuk merekam data aktivitas bersepeda secara rinci, seperti jarak tempuh, kecepatan, elevasi, serta waktu pelaksanaan (Imrit et al., 2024). Data ini membuka peluang besar untuk dianalisis secara ilmiah guna memahami faktor-faktor yang memengaruhi performa bersepeda (Rupaka et al., 2021).

Salah satu indikator performa pesepeda yang umum digunakan adalah kecepatan rata-rata. Nilai ini dapat dipengaruhi oleh berbagai faktor, di antaranya adalah kondisi topografi lintasan seperti elevasi dan kemiringan, serta cuaca, seperti suhu dan kelembapan udara (Aji & Bestari, 2025). Di samping itu, terdapat pula faktor-

faktor non-teknis seperti kualitas tidur, kondisi tubuh, dan permukaan jalan yang basah setelah hujan. Meskipun faktor non-teknis ini tidak tercatat secara eksplisit dalam data, pemahaman terhadapnya penting untuk memberi konteks tambahan terhadap hasil analisis performa (Widodo & Muhammad, 2023).

Untuk memahami dan memprediksi hubungan antara kondisi lingkungan dengan performa bersepeda, diperlukan pendekatan komputasional berbasis pembelajaran mesin (*machine learning*). Salah satu algoritma yang dapat digunakan adalah XGBoost (*Extreme Gradient Boosting*), yaitu metode prediksi berbasis pohon keputusan yang terkenal efektif dalam memodelkan hubungan non-linear serta menghasilkan akurasi tinggi (Abdullah, 2025). Dengan menggunakan algoritma ini, penelitian bertujuan untuk membangun model prediksi kecepatan rata-rata bersepeda berdasarkan data historis aktivitas dari Strava dan variabel lingkungan seperti topografi dan cuaca (Røsten & Rogstad, 2025).

Urgensi penelitian ini terletak pada pentingnya memahami bagaimana faktor-faktor seperti elevasi, kondisi cuaca, dan kualitas waktu tidur dapat memengaruhi kecepatan rata-rata bersepeda. Elevasi dan kemiringan jalur berhubungan langsung dengan tingkat kesulitan lintasan yang dapat memperlambat atau mempercepat kayuhan. Faktor cuaca, seperti hujan atau suhu udara, dapat memengaruhi kenyamanan, keamanan, serta efisiensi tenaga pesepeda. Sementara itu, kualitas waktu tidur yang baik berkontribusi terhadap kesiapan fisik dan daya tahan tubuh, yang pada akhirnya memengaruhi performa saat bersepeda. Oleh karena itu, analisis mendalam terhadap variabel-variabel tersebut tidak hanya penting untuk menghasilkan prediksi yang akurat, tetapi juga bermanfaat bagi pesepeda dalam merencanakan strategi latihan, mengatur pola tidur, serta menyesuaikan aktivitas dengan kondisi lingkungan untuk mencapai performa optimal .

Pada Penelitian Prediksi Angka Harapan Hidup Penduduk Menggunakan Model XGBoost menunjukkan performa prediksi yang sangat baik dengan tingkat akurasi mencapai 96,8%**.** Nilai Mean Absolute Error (MAE) sebesar 0,97 mengindikasikan bahwa rata-rata selisih antara hasil prediksi dan data aktual relatif kecil (Kurniawan and Indahyanti 2024).

Penelitian Sebelumnya Penerapan Algoritma *Extreme Gradient Boosting* (XGBoost) untuk Analisis Risiko Kredit Hasil analisis menggunakan algoritma XGBoost menunjukkan bahwa metode ini cukup efektif dalam melakukan klasifikasi pada data kredit. Pada tahap data mining, dilakukan pembangunan 10 model untuk memperoleh performa terbaik. Sebelum proses pemodelan, data terlebih dahulu diseimbangkan dengan SMOTE untuk mengatasi masalah ketidakseimbangan kelas. Dari hasil evaluasi diperoleh bahwa algoritma XGBoost dapat dijadikan pendekatan untuk membantu proses pengambilan keputusan dalam menerima atau menolak pengajuan kredit. Evaluasi juga memperlihatkan bahwa kinerja model meningkat setelah penerapan SMOTE, yang ditunjukkan dengan perbaikan nilai akurasi maupun AUC. Model terbaik diperoleh pada skenario pertama (90% data latih dan 10% data uji) dengan capaian akurasi sebesar 0,83 dan AUC 0,918.(Saputra et al., 2024).

Pada penelitian lain memanfaatkan XGBoost untuk prediksi penyakit stroke dan menghasilkan performa yang baik (akurasi 95,4%, precision 94,3%, recall 96,6%, F1-score 95,4%, AUC 95,4%). Meski hasilnya tinggi, performanya masih di bawah model Stacking dan Random Forest dari penelitian sebelumnya. Teknik SMOTE digunakan untuk mengatasi data imbalance, dan Bayesian Optimization untuk tuning hyperparameter. Peneliti menyarankan optimasi lebih lanjut karena ruang pencarian

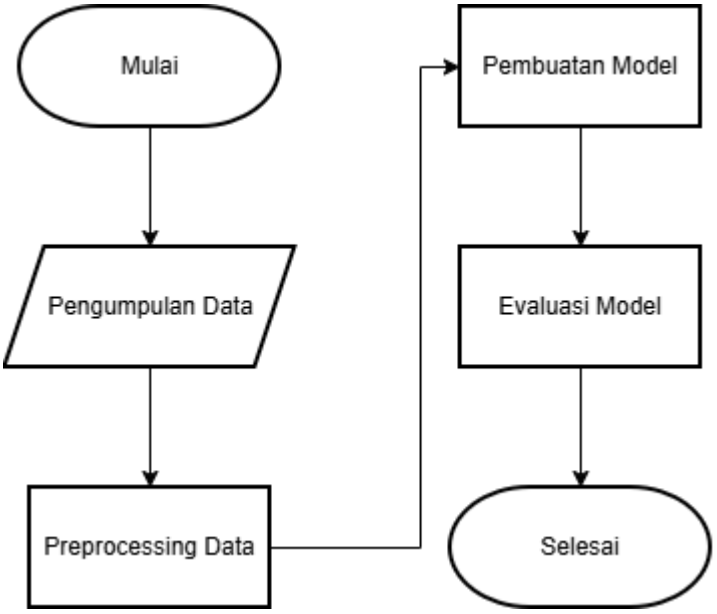
hyperparameter masih luas, membuka peluang peningkatan di penelitian selanjutnya (Murdiansyah, 2024).

Hasil penelitian lain menunjukkan bahwa algoritma Random Forest dan XGBoost tidak memiliki perbedaan signifikan dalam hal akurasi prediksi volume lalu lintas. Namun, nilai R^2 pada XGBoost lebih tinggi**,** sehingga kelayakan model ini dapat dikatakan lebih baik dibanding Random Forest. Perbedaan mencolok justru terlihat pada waktu pemrosesan**,** di mana XGBoost terbukti jauh lebih efisien, yaitu sekitar 532% lebih cepat daripada Random Forest. Berdasarkan hasil perbandingan tersebut, dapat disimpulkan bahwa XGBoost merupakan algoritma yang lebih tepat digunakan untuk membangun model prediksi kepadatan lalu lintas karena menawarkan akurasi tinggi sekaligus efisiensi waktu pemrosesan(Lisanthoni et al., 2023).

Penelitian Pemodelan ritme pembakaran kalori selama aktivitas bersepeda menggunakan *Feedforward Neural Network* (FFNN**)** menghasilkan konfigurasi optimal dengan 59 *neuron* pada lapisan tersembunyi. Model ini mampu memprediksi jumlah kalori yang terbakar (Calt) dengan tingkat error sekitar 7%**.** Hasil pengujian juga menunjukkan bahwa suhu memiliki pengaruh positif, di mana semakin tinggi suhu maka kalori yang terbakar juga meningkat. Secara keseluruhan, rancangan FFNN dalam penelitian ini dapat dikatakan cukup efektif dengan tingkat akurasi mencapai 93% dalam memprediksi kalori terbakar per detik (ICHWAN & ALFARISYI, 2024).

METODE PENELITIAN

Metode penelitian digunakan untuk merencanakan, melaksanakan, dan menganalisis penelitian. Metode penelitian ini dapat membantu dalam merancang prosedur yang tepat untuk menyusun data bermanfaat, ini adalah beberapa metode penelitian yang digunakan:



Gambar 1 Flowchart Penelitian

1. Pengumpulan Data

Sumber data pada penelitian ini menggunakan data aktivitas bersepeda pribadi yang diperoleh melalui aplikasi Strava, dengan rentang waktu mulai dari 2 Juni 2024 hingga 9 Agustus 2025. Data tersebut dikumpulkan oleh peneliti secara langsung dari Strava langsung diinput ke dalam Excel. Dataset yang digunakan berisi informasi aktivitas bersepeda seperti waktu mulai, durasi aktivitas, jarak

tempuh, total elevasi, kecepatan rata-rata, dan kecepatan maksimum. Data tersebut diekspor dalam format .csv, yang kemudian diolah dan dianalisis.

Tabel 1 Dataset

Tangal	Jam Mulai	Waktu Total	*Elevasi*	Jarak Tempuh	Kecpatan Rata-rata	Kecepatan Max	Cuaca	Waktu Tidur
2 Juni 24	15.42	40.13.23	80 m	15.67 km	25,5 km/jam	45.3 km/jam	0	7
9 Juni 24	5.41	2.45.12	723 m	29.12 km	21,2 km/jam	52,6 km/jam	0	6
15 Juni 24	5.33	1.24.35	275 m	33,72 km	26,3 km/jam	49,5 km/jam	1,2	4
23 Juni 24	15.53	1.14.45	57 m	25,3 km	18,7 km/jam	37,2 km/jam	3,5	7
1 Juli 24	5.45	1.40.50	427 m	36,6 km	22,4 km/jam	32,7 km/jam	1,7	7

2. Preprocessing Data

Tahapan ini adalah proses mengolah data mentah menjadi data yang siap digunakan untuk pemodelan dengan tujuan untuk menghasilkan data yang lebih akurat (Aprianto et al., 2025). Data *preprocessing* meliputi tahapan sebagai berikut :

a. Penanganan *Missing Value*

Untuk mengidentifikasi keberadaan nilai yang hilang secara sistematis, penelitian ini menggunakan pustaka pandas pada Python. Prosesnya adalah dengan menerapkan fungsi .isnull().sum(). Fungsi ini akan memindai seluruh *DataFrame* dan menghasilkan sebuah objek baru dengan ukuran yang sama, di mana setiap sel yang berisi nilai hilang akan ditandai sebagai True, dan yang berisi data akan ditandai sebagai False. Hasilnya adalah sebuah daftar yang menunjukkan nama setiap kolom beserta jumlah total nilai yang hilang di dalamnya.

b. Penanganan Data Duplikat

Tujuan dari tahap penanganan data duplikat ini adalah untuk memastikan bahwa setiap baris data dalam dataset bersifat unik. Keberadaan data duplikat dapat menyebabkan bias pada model *machine learning*, karena model akan memberikan bobot yang tidak semestinya pada data yang berulang. Hal ini juga dapat mengarah pada hasil evaluasi yang terlalu optimis dan tidak mencerminkan kinerja model di dunia nyata. Implementasi untuk menangani data duplikat dilakukan menggunakan pustaka pandas dengan fungsi .duplicated().sum().

c. Normalisasi Data

Normalisasi data merupakan teknik krusial dalam tahap pra-pemrosesan, terutama karena setiap variabel dalam dataset umumnya memiliki skala nilai yang berbeda. Tujuan dari normalisasi adalah untuk mengubah skala nilai dari berbagai fitur ke dalam rentang yang seragam. Dalam penelitian ini, peneliti menerapkan metode *min-max scaler* untuk melakukan normalisasi, yaitu dengan mengubah nilai-nilai data ke dalam rentang tertentu umumnya antara 0 hingga 1 agar setiap variabel berada pada skala yang sebanding dan dapat dibandingkan secara seimbang (Salsabil Muhammad et al., 2024). Transformasi ini dilakukan dengan menggunakan rumus matematis sebagai berikut:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Dimana :

X_{scaled} adalah nilai data setelah dinormalisasi.

X adalah nilai data asli.
X_{min} adalah nilai minimum dari seluruh data pada fitur tersebut.
X_{max} adalah nilai maksimum dari seluruh data pada fitur tersebut.

d. Split Data

Tahap ini merupakan prosedur fundamental dalam pengembangan model *machine learning* untuk mengevaluasi kinerja model secara objektif. Tujuan utama dari pembagian data adalah untuk melatih model pada satu set data (data latih) dan kemudian menguji seberapa baik model tersebut dapat melakukan generalisasi pada data baru yang belum pernah dilihat sebelumnya (data uji). Untuk data deret waktu (*time series*) seperti harga Ethereum, pembagian data tidak boleh dilakukan secara acak. Pembagian harus dilakukan secara kronologis untuk menjaga dependensi temporal data. Hal ini mensimulasikan skenario dunia nyata di mana kita menggunakan data masa lalu untuk memprediksi data di masa depan. Dalam penelitian ini, rasio pembagian yang digunakan adalah 80:20, dengan rincian yaitu 80% sebagai data latih (*training data*) dan 20% sebagai data uji (*testing data*).

3. Pembuatan Model

Tahap pembuatan model merupakan inti dari penelitian ini karena pada bagian ini dilakukan implementasi dua pendekatan pemodelan berbeda, yaitu model GRU (*Gated Recurrent Unit*) yang berbasis *deep learning* dan model XGBoost (*Extreme Gradient Boosting*) yang merupakan algoritma *machine learning* berbasis pohon keputusan. Tujuan dari tahap ini adalah membangun kedua model dengan arsitektur dan parameter yang sesuai .

a. Pembuatan Model XGBoost

Pada tahap ini merupakan langkah untuk pembuatan model *Extreme Gradient Boosting* (XGBoost) setelah proses preprocessing data. Berikut merupakan langkah pembuatan model XGBoost yaitu :

1) Persiapan Data untuk XGBoost: XGBoost memerlukan data dalam format tabular (baris dan kolom). Oleh karena itu, data deret waktu perlu diubah menjadi format *supervised learning*. Proses ini biasanya melibatkan pembuatan fitur (*feature engineering*) dari data masa lalu, seperti menggunakan harga penutupan hari sebelumnya (t-1, t-2, dst.) dan nilai indikator teknikal sebagai fitur input untuk memprediksi harga hari ini (t).
2) Inisialisasi Model XGBoost: Model XGBoost tidak dibangun lapis demi lapis seperti jaringan saraf. Sebagai gantinya, sebuah objek model diinisialisasi, yaitu *XGBRegressor*, karena tugasnya adalah prediksi regresi (memprediksi nilai kontinu).
3) Menentukan *Hyperparameter*: Sejumlah *hyperparameter* utama ditentukan untuk mengontrol kinerja dan kompleksitas model. Beberapa di antaranya adalah:
   - *n_estimators*: Jumlah total pohon keputusan (*decision tree*) yang akan dibangun.
   - *learning_rate*: Mengontrol laju belajar model untuk mencegah *overfitting*.
   - *max_depth*: Kedalaman maksimum dari setiap pohon untuk membatasi kompleksitas.
   - *objective*: Fungsi tujuan yang dioptimalkan, misalnya 'reg:squarederror' untuk masalah regresi.

4) Melatih model: Model dilatih menggunakan fungsi .fit() pada data latih. Seringkali, teknik *Early Stopping* digunakan selama pelatihan, di mana performa model dipantau pada data validasi. Pelatihan akan berhenti jika tidak ada peningkatan performa setelah beberapa iterasi tertentu untuk mendapatkan model yang optimal.

4. Evaluasi Model

Setelah model XGBoost berhasil dilatih menggunakan data latih (training data), tahap selanjutnya adalah melakukan evaluasi untuk mengukur kinerja dan akurasi prediksi kecepatan rata-rata bersepeda. Evaluasi ini bertujuan untuk mengetahui sejauh mana model mampu memprediksi dengan benar pada data yang belum pernah digunakan selama proses pelatihan. Data yang digunakan untuk evaluasi adalah data uji (test data), sehingga hasilnya dapat mencerminkan kemampuan generalisasi model terhadap data baru (Nugraha & Ariatmanto, 2025). Metrik evaluasi yang digunakan untuk mengukur seberapa baik model tersebut digunakan dalam memprediksi yaitu dengan menggunakan RMSE dan R^2Score. Dari hasil pengujian ini maka akan diberikan kesimpulan untuk masing-masing model yang di uji.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{\{i=1\}}^{n}(y_1 - \hat{y}_1)^2}$$

Dimana :

y_1: nilai aktual

$\hat{y}_1$: nilai prediksi

$\frac{1}{n}$: menghitung rata-rata dari kuadrat eror

$\sum$: menjumlahkan semua eror dari semua data

Dari perhitungan tersebut memberikan pengertian bahwa semakin kecil RMSE yang dihasilkan maka akurasi model yang digunakan semakin tinggi dan sebaliknya apabila hasil RMSE semakin besar maka nilai akurasi pada model semakin kecil. RMSE ini dipilih karena keunggulan yang dimiliki oleh RMSE dibandingkan dengan perhitungan akurasi yang lain. RMSE dinilai cocok untuk digunakan dalam pengolahan data yang besar (Meriani & Rahmatulloh, 2024).

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

Dimana :

n jumlah total data

y_i adalah nilai harga aktual pada data ke-I

$\hat{y}_i$ adalah nilai harga prediksi yang dihasilkan oleh model pada data ke-i

$\bar{y}_i$ adalah nilai rata-rata dari seluruh harga aktual

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ adalah jumlah kuadrat selisih antara nilai aktual dan prediksi (sama seperti pada perhitungan MSE)

$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ adalah jumlah kuadrat selisih total dati data aktual terhadap nilai rata-ratanya

Selain RMSE, penelitian ini juga menggunakan metrik *R-squared* (R^2) atau Koefisien Determinasi untuk mengukur kecocokan model (*goodness-of-fit*). Metrik ini berfungsi untuk mengukur seberapa besar proporsi variasi dari harga aktual Ethereum yang berhasil dijelaskan oleh model prediksi. Nilai R^2 berkisar antara 0 hingga 1(atau 0% hingga 100%), di mana skor yang lebih tinggi menunjukkan

performa yang lebih baik. Secara matematis, R^2 dihitung dengan membandingkan jumlah kuadrat kesalahan prediksi model terhadap jumlah kuadrat total dari data, sehingga memberikan gambaran komprehensif tentang kemampuan prediktif model (Sanjaya, 2024).

HASIL DAN PEMBAHASAN

1. Hasil Pengumpulan Data

Data diperoleh dari *export* aktivitas Strava dalam bentuk Excel/CSV, yang berisi informasi seperti jarak tempuh, durasi, kecepatan rata-rata, kecepatan maks, elevasi, dan kecepatan rata-rata. Data cuaca diambil berdasarkan tanggal dan waktu aktivitas menggunakan sumber *historical weather*. Data kondisi tidur diinput secara manual sesuai catatan distarava pesepeda.

Pengumpulan data dilakukan mulai 3 Juni 2024 hingga 9 Agustus 2025. Selama periode ini yg nantinya akan digunakan sebagai data training dan testing, setiap aktivitas bersepeda dicatat secara konsisten menggunakan perangkat GPS dan data cuaca diperoleh baik dari catatan manual maupun integrasi informasi cuaca pada aplikasi Strava.

Tabel 2 Dataset

Tangal	Jam Mulai	Waktu Total	*Elevasi*	Jarak Tempuh	Kecpatan Rata-rata	Kecepatan Max	Cuaca	Waktu Tidur
2 Juni 24	15.42	40.13.23	80 m	15.67 km	25,5 km/jam	45.3 km/jam	0	7
10 Juni 2025	05.20	2.00.12	723 m	29.12 km	21,2 km/jam	52,6 km/jam	0	6
3 Juli 2025	16.17	1.11.44	112 m	24,5 km	20,2 km/jam	49,5 km/jam	1,2	6
9 Agustus 2025	05.38	1.44.07	217 m	25,3 km	23,7 km/jam	97,0 km/jam	3,5	8
1 Juli 24	5.45	1.40.50	427 m	36,6 km	22,4 km/jam	32,7 km/jam	1,7	7

Fitur- Fitur yang diambil meliputi

- Tanggal : Aktivitas bersepeda dilakukan
- Jam mulai : Waktu pukul aktivitas bersepeda
- Waktu total : Total durasi bersepeda
- Elveasi : Total kenaikan ketinggan
- Jarak tempuh : Panjang total route yang ditempuh
- Kecepatan rata-rata : Nilai kecepatan rata-rata bersepeda
- Kecepatan maks : Kecepatan tertinggi yang dicapai selama aktivitas bersepeda
- Cuaca : Kondisi cuaca saat bersepeda, misalnya cerah, mendung, atau hujan.
- Waktu tidur : Jumlah jam tidur sebelum aktivitas bersepeda

2. Hasil Preprocessing Data

a. Mengecek Missing Value

Pengecekan *missing value* dilakukan dengan menghitung jumlah nilai kosong (NaN atau null) pada setiap kolom fitur yang digunakan, yaitu Tanggal, Jam Mulai, Waktu Total, Elevasi, Jarak Tempuh, Kecepatan Rata-rata, Kecepatan Max, Cuaca, Waktu Tidur. Hasil dari proses pengecekan disajikan pada tabel 2 di bawah ini.

Tabel 3 Handling Missing Value

Kolom	*Handling Missing Value*
Tanggal	0
Jam Mulai	0
Waktu Total	0
Elevasi	0
Jarak Tempuh	0
Kecepatan Rata-Rata	0
Kecepatan Max	0
Cuaca	0
Waktu Tidur	0

Berdasarkan Tabel 3, dapat dilihat bahwa hasil pengecekan menunjukkan angka 0 untuk semua kolom fitur. Hal ini mengindikasikan bahwa tidak ditemukan adanya nilai yang hilang dalam dataset penelitian. Dengan demikian, tidak diperlukan tindakan lebih lanjut seperti imputasi atau penghapusan data, dan dataset dapat dianggap lengkap serta siap untuk tahap pra-pemrosesan.

b. Mengecek Duplikasi Data

Tabel 4 Duplikasi Data

Kolom	*Duplikasi Data*
Tanggal	0
Jam Mulai	0
Waktu Total	0
Elevasi	0
Jarak Tempuh	0
Kecepatan Rata-Rata	0
Kecepatan Max	0
Cuaca	0
Waktu Tidur	0

Pada tahapan selanjutnya adalah melihat duplicate data dari setiap data. Tahap ini menghasilkan bahwa data yang akan digunakan tidak terdapat duplikasi data sehingga data tersebut bisa digunakan untuk proses selanjutnuya.

c. Normalisasi Data

Proses normalisasi data yang menggunakan metode min-max scaler dengan menggunakan library MinMaxScaler untuk merubah dataset ke dalam skala 0-1. Proses normalisasi data ini sangat penting karena membantu percepatan konvergensi dalam metode XGBoost sehingga pemrosesan dataset tidak membutuhkan waktu yang lama. Selain itu, normalisasi data ini juga berfungsi untuk stabilitas perhitungan sehingga nilai akurasi yang dihasilkan lebih stabil. Tahapan selanjutnya adalah merubah data menjadi data sequensial karena metode XGBoost ini tidak dapat memproses data biasa

Proses ini diharapkan dapat membantu XGBoost dalam menemukan pola hubungan antara topografi, cuaca, dan faktor fisik pengendara secara lebih akurat, serta meminimalkan pengaruh nilai ekstrim yang dapat mengganggu proses pembelajaran model

d. Split data

Tahapan ini digunakan untuk memisahkan data *training* dan data *cleaning* dengan tujuan untuk menghindari *overfitting* yang akan mengganggu pemrosesan data dalam model XGBoost. Pembagian data ini juga diperlukan untuk memastikan model yang sudah dibuat dapat mengikuti dengan baik dan memberikan hasil prediksi yang akurat

Tabel 5 Hasil *Splitting Data*

Data/Kategori	Data *Splitting*	
	Data *Training*	Data *testing*
Elevasi	0.8	0.2
Cuaca	0.8	0.2
Waktu Tidur	0.8	0.2

Pada tabel 4 memberikan informasi untuk pembagian data yang akan digunakan dalam pemrosesan data yaitu dengan rasio 80:20 dimana 80% merupakan data *training* dan 20% digunakan untuk data testing.

3. Hasil Pembuatan *Extreme Gradient Boosting* (XGBoost)

Tabel 6 *Hyperparameter* model XGBoost

Model	Hyperparameter				
	n_estimator	learning rate	Max depth	sub sample dan colsample_bytree	Random state
XGBoost	300	0.2	3	0.8/0.8	42

Model XGBoost diinisialisasi sebagai *XGBRegressor* karena tugas yang diemban adalah *regresi* untuk memprediksi nilai kontinu, dengan *objective* diatur ke *'reg:squarederror'* agar model secara eksplisit mengoptimalkan *Mean Squared Error* (RMSE). Untuk membangun model ini, jumlah pohon keputusan (*n_estimators*) ditetapkan sebanyak 300, dengan laju belajar (*learning_rate*) sebesar 0.2 yang memungkinkan model belajar secara perlahan namun lebih robust. Guna mengontrol kompleksitas dan mencegah *overfitting*, kedalaman maksimum setiap pohon (*max_depth)* dibatasi hingga 3. Selain itu, teknik regularisasi lebih lanjut diterapkan dengan mengatur parameter subsample dan *colsample_bytree* ke 0.8. Hal ini berarti setiap pohon hanya menggunakan 80% sampel data dan 80% fitur secara acak, sehingga menambah keragaman dan kemampuan generalisasi model. Terakhir, *random_state* diatur ke 42 untuk memastikan bahwa hasil pelatihan dapat diulang kembali dengan konsisten.

4. Hasil Evaluasi Model

Tabel 8 Hasil evaluasi model XGBoost

Model	RMSE	R^2
XGBoost	1.240 km/h	0.800

Dalam penelitian ini, metrik evaluasi yang digunakan adalah *Root Mean Squared Eror* (RMSE) dan R-Squared (R^2). Hasil evaluasi menunjukan bahwa model menghasilkan nilai RMSE sebesar 1.240 km/h. Nilai ini menunjukan bahwa rata-rata kesalahan prediksi kecepatan model sekitar 1.240 km/h dari nilai aktual. Semakin kecil nilai RMSE yang dihasilkan maka akurasi model semakin tinggi, dan sebaliknya jika RMSE besar maka tingkat kesalahan prediksi juga semakin besar.

Selain RMSE, penelitian ini juga menggunakan metrik *R-squared* (R^2) atau Koefisien Determinasi untuk mengukur kecocokan model (*goodness-of-fit*). Hasil evaluasi menunjukkan nilai R^2 sebesar 0.800 atau 80%, yang berarti model XGBoost mampu menjelaskan sekitar 80% variasi data kecepatan rata-rata bersepeda berdasarkan faktor-faktor masukan seperti topografi, cuaca, dan waktu tidur. Nilai R^2 yang tinggi ini menandakan bahwa model memiliki performa yang baik dan dapat digunakan untuk melakukan prediksi dengan tingkat keandalan yang cukup tinggi.

KESIMPULAN

Berdasarkan hasil penelitian dan implementasi yang telah dilakukan, tujuan utama penelitian untuk mengimplementasikan algoritma *Extreme Gradient Boosting* (XGBoost) dalam membangun model prediksi kecepatan rata-rata bersepeda telah berhasil dicapai. Model XGBoost yang dikembangkan terbukti mampu memberikan hasil estimasi kecepatan rata-rata dengan tingkat akurasi yang cukup baik. Hal ini ditunjukkan secara kuantitatif oleh nilai performa model, yaitu *Root Mean Squared Error* (RMSE) sebesar 1.240 km/jam dan nilai *R-squared* (R^2) sebesar 0.800. Nilai RMSE yang relatif kecil menunjukkan bahwa rata-rata kesalahan prediksi model terhadap nilai kecepatan aktual tidak signifikan. Sementara itu, nilai R^2 sebesar 0.800 menegaskan bahwa model mampu menjelaskan 80% variasi data kecepatan rata-rata berdasarkan variabel-variabel masukan, yang menandakan model ini cukup representatif.

DAFTAR PUSTAKA

Abdullah, M. A. (2025). Prediksi Jumlah Pasien Medical Check Up Berdasarkan Time Series Forecasting Menggunakan Algoritma XGBoost. *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, *6*(2), 488–497.

Aji, T., & Bestari, S. (2025). ANALISIS FAKTOR-FAKTOR YANG MEMENGARUHI WAKTU TEMPUH PEMBALAP BENTANG JAWA 2024 DENGAN METODE REGRESI RIDGE. *Jurnal Statistika Industri Dan Komputasi*, *10*(2), 15–22.

Aprianto, K., Mahdiyah, U., & Wulanningrum, R. (2025). Analisis Perbandingan Model PSO-LSTM dan LSTM Konvensional untuk Prediksi Harga Bitcoin di Market Cryptocurrency 1*. In *INOTEK* (Vol. 9).

ICHWAN, M., & ALFARISYI, S. F. (2024). Pemodelan Ritme Kalori Terbakar Setiap Waktu Selama Bersepeda dengan Feedforward Neural Network. *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, *12*(1), 13. https://doi.org/10.26760/elkomika.v12i1.13

Imrit, A. A., Fischer, J., Chan, T. C. Y., Saxe, S., & Bonsma-Fisher, M. (2024). A Street-Specific Analysis of Level of Traffic Stress Trends in Strava Bicycle Ridership and its Implications for Low-Stress Bicycling Routes in Toronto. *Transport Findings*, *2024*. https://doi.org/10.32866/001c.92109

Irawati, A. F., Irwin, & Anshar, A. M. (2024). Kepopuleran Olahraga Sepeda Sebagai Bagian Dari Pola Hidup Sehat. *Jurnal Ilmiah Global Education*, *5*(3), 2038–2043.

Kurniawan, W., & Indahyanti, U. (2024). Prediksi Angka Harapan Hidup Penduduk Menggunakan Metode XGBoost. *Indonesian Journal of Applied Technology*, *1*(2), 18. https://doi.org/10.47134/ijat.v1i2.3045

Lisanthoni, A., Sari, F. I., Gunawan, E. L., & Adhigiadany, C. A. (2023). Model Prediksi Kepadatan Lalu Lintas: Perbandingan Algoritma Random Forest dan XGBoost. *Prosiding Seminar Nasional Sains Data*, *3*(1), 296–303. https://doi.org/10.33005/senada.v3i1.126

Meriani, A. P., & Rahmatulloh, A. (2024). PERBANDINGAN GATED RECURRENT UNIT (GRU) DAN ALGORITMA LONG SHORT TERM MEMORY (LSTM) LINEAR REFRESSION DALAM PREDIKSI HARGA EMAS MENGGUNAKAN MODEL TIME SERIES. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(1). https://doi.org/10.23960/jitet.v12i1.3808

Murdiansyah, D. T. (2024). Prediksi Stroke Menggunakan Extreme Gradient Boosting. *JIKO (Jurnal Informatika Dan Komputer)*, *8*(2), 419.

https://doi.org/10.26798/jiko.v8i2.1295

Nugraha, D. M., & Ariatmanto, D. (2025). *Meningkatkan Akurasi Prediksi Harga Bitcoin dengan Algoritma GRU-LSTM Hibrida*. *11*(1). https://journal.fkom.uniku.ac.id/index.php/buffer

Røsten, S., & Rogstad, E. T. (2025). Ride, record, and share? A study of elite cyclists' sharing practices on Strava. *European Journal for Sport and Society*. https://doi.org/10.1080/16138171.2025.2532278

Rupaka, A. P., Sulistyo, A. B., & Punia, D. (2021). Pemetaan Perilaku Pesepeda Pra dan Pasca Pandemi Covid-19 di Provinsi Bali Menggunakan Data Strava Metro. *Jurnal Teknologi Transportasi Dan Logistik*, *2*(2), 119–126. https://doi.org/10.52920/jttl.v2i2.32

Salsabil Muhammad, Azizah Lutvi Nuril, & Ade Eviyanti. (2024). Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost. *Jurnal Ilmiah Komputasi*, *23*(1). https://doi.org/10.32409/jikstik.23.1.3507

Sanjaya, K. D. (2024). *PREDIKSI HARGA RUMAH DENGAN METODE REGRESI LINEAR DAN SUPPORT VECTOR REGRESSION DI DAERAH TEBET JAKARTA SELATAN* (Vol. 19, Issue 2). Versi Cetak.

Saputra, A. A., Sari, B. N., Rozikin, C., Singaperbangsa, U., & Abstrak, K. (2024). Penerapan Algoritma Extreme Gradient Boosting (Xgboost) Untuk Analisis Risiko Kredit. *Jurnal Ilmiah Wahana Pendidikan*, *10*(7), 27–36.

Sitorus, B. (2024). MENUMBUHKAN MINAT MASYARAKAT BERSEPEDA SEBAGAI KEBIASAAN BARU DI KOTA BEKASI. *Jurnal Penelitian Sekolah Tinggi Transportasi Darat*, *14*(2), 108–115. https://doi.org/10.55511/jpsttd.v14i2.656

Surojo, S. S., Widiyatmoko, F., & Kresnapati, P. (2022). Survei Antusiasme Dan Ketertarikan Masyarakat Dalam Bersepeda Di Kota Semarang. *Journal of Sport Science and Fitness*, *8*(1), 63–68. https://doi.org/10.15294/jssf.v8i1.58379

Utomo, A. W. (2016). 済無No Title No Title No Title. *Angewandte Chemie International Edition, 6(11), 951–952.*, *1*(0), 1–23.

Widodo, A. P., & Muhammad. (2023). Profil Kondisi Fisik Atlet Balap Sepeda Jalana Raya Puslatcab Issi Surabaya Dalam Rangka Persiapan Porprov Jawa Timur 2022. *Jurnal Prestasi Olahraga*, *6*(1), 30–35.