

Prediksi Durasi Perjalanan Taxi NYC Menggunakan Regression Model

Azaria Bunga¹, Muhammad Fariz², Muhammad Riyad Firdaus³, Perani Rosyani⁴

¹²³⁴Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan,
Indonesia

Email: ¹azariabunga502@gmail.com, ²m.faariiz@gmail.com, ³friyad0201@gmail.com,
⁴dosen00837@unpam.ac.id

Abstrak—Prediksi durasi perjalanan transportasi urban merupakan aspek penting dalam meningkatkan efisiensi operasional dan kualitas layanan. New York City Yellow Taxi menyediakan dataset perjalanan dalam jumlah besar yang dapat dimanfaatkan untuk membangun model prediksi berbasis *machine learning*. Penelitian ini mengembangkan sistem prediksi durasi perjalanan menggunakan empat algoritma: *Linear Regression*, *Random Forest*, *Gradient Boosting*, dan *XGBoost*. Dataset diproses melalui tahap *preprocessing*, *feature engineering*, dan pembagian data menjadi *training* dan *testing set*. Hasil penelitian menunjukkan bahwa *XGBoost* memberikan performa terbaik dengan *MAE* 189,23 detik, *RMSE* 287,91 detik, dan R^2 sebesar 0,8657, mengungguli model lainnya. Faktor paling berpengaruh terhadap durasi perjalanan meliputi *trip_distance*, *haversine_distance*, dan *hour_of_day*. Sistem yang dibangun memungkinkan prediksi *real-time* serta analisis data secara komprehensif. Penelitian ini membuktikan bahwa algoritma *tree-based boosting*, khususnya *XGBoost*, efektif digunakan dalam memodelkan dinamika durasi perjalanan taxi di lingkungan urban yang kompleks.

Kata Kunci: Prediksi durasi perjalanan; *XGBoost*; Machine Learning; NYC Taxi; Regresi

Abstract—Predicting travel duration in urban transportation is a crucial aspect of improving operational efficiency and service quality. The New York City Yellow Taxi dataset provides a large volume of trip records that can be utilized to build machine learning-based prediction models. This study develops a travel duration prediction system using four algorithms: *Linear Regression*, *Random Forest*, *Gradient Boosting*, and *XGBoost*. The dataset is processed through preprocessing, feature engineering, and splitting into training and testing sets. The results show that *XGBoost* delivers the best performance with an *MAE* of 189.23 seconds, an *RMSE* of 287.91 seconds, and an R^2 value of 0.8657, outperforming the other models. The most influential factors affecting travel duration include *trip_distance*, *haversine_distance*, and *hour_of_day*. The developed system enables real-time prediction and comprehensive data analysis. This study demonstrates that tree-based boosting algorithms, particularly *XGBoost*, are highly effective in modeling the complex dynamics of taxi travel duration in urban environments.

Keywords: Travel duration prediction; *XGBoost*; Machine Learning; NYC Taxi; Regression

1. PENDAHULUAN

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence*) dalam dua dekade terakhir telah membawa dampak signifikan dalam berbagai aspek kehidupan manusia, termasuk transportasi, logistik, dan pelayanan publik. Salah satu cabang AI yang berkembang pesat adalah *machine learning*, yaitu model komputasi yang dapat mempelajari pola dari data secara otomatis dan menghasilkan prediksi akurat (James, Witten, Hastie, & Tibshirani, 2021). Dalam konteks transportasi urban, *machine learning* berperan penting dalam mengurangi ketidakpastian perjalanan dan meningkatkan akurasi prediksi.

New York City (NYC) merupakan salah satu kota metropolitan dengan aktivitas transportasi paling padat di dunia. Sistem transportasi NYC menghasilkan data perjalanan dalam volume besar yang mencerminkan pola pergerakan manusia dan kondisi lalu lintas *real-time* (NYC Taxi & Limousine Commission, 2023). Prediksi durasi perjalanan taxi menjadi aspek krusial dalam meningkatkan kualitas layanan transportasi. Bagi perusahaan taxi, prediksi akurat membantu optimasi alokasi armada dan pengaturan jadwal pengemudi. Bagi penumpang, estimasi waktu tepat mendukung perencanaan perjalanan dan meningkatkan kepuasan layanan.

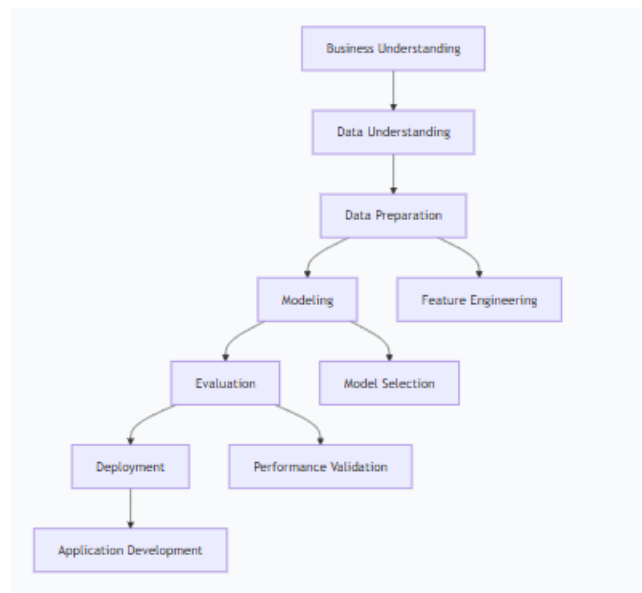
Namun, prediksi durasi perjalanan di lingkungan urban merupakan tantangan kompleks karena dipengaruhi banyak variabel seperti jarak, waktu, kondisi lalu lintas, cuaca, dan lokasi geografis. Pendekatan konvensional seperti persamaan linear sering tidak mampu menangkap hubungan non-linear dan kompleks antar variabel. Oleh karena itu, model *machine learning*—

khususnya algoritma *tree-based* dan *boosting*—lebih sesuai untuk memodelkan dinamika tersebut (Chen & Guestrin, 2016). Penelitian ini mengembangkan sistem prediksi durasi perjalanan taxi NYC menggunakan empat algoritma utama: *Linear Regression*, *Random Forest*, *Gradient Boosting*, dan *XGBoost*.

2. METODE PENELITIAN

2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan *Research and Development* (R&D) dengan tahapan sistematis seperti ditunjukkan pada Gambar 1. Alur penelitian dimulai dari pengumpulan data, *preprocessing*, *feature engineering*, pembangunan model, evaluasi, hingga implementasi sistem.



Gambar 1. Alur Penelitian

2.2 Pengumpulan dan Persiapan Data

Dataset diperoleh dari Kaggle dengan judul "NYC Yellow Taxi Trip Data" (NYC Taxi & Limousine Commission, 2023). Data mencakup 50.000 *record* yang mencakup variabel utama seperti *pickup_datetime*, *dropoff_datetime*, *pickup_latitude*, *pickup_longitude*, *dropoff_latitude*, *dropoff_longitude*, *trip_distance*, *passenger_count*, dan *trip_duration*. Data dibagi menjadi *training set* (80%) dan *testing set* (20%) menggunakan teknik *stratified sampling*.

2.3 Preprocessing Data

Preprocessing dilakukan melalui tahapan:

- Pembersihan *missing values* menggunakan teknik *mean/median imputation*
- Penghapusan *outliers* (durasi > 2 jam atau jarak 0 mil)
- Normalisasi data menggunakan *StandardScaler*
- Validasi koordinat (latitude: 40.5–41.0, longitude: -74.5–73.5)

2.4 Feature Engineering

Feature engineering dilakukan untuk mengekstrak informasi tambahan:

1. **Haversine Distance:** Menghitung jarak geodesik menggunakan formula:

$$haversine_distance = 2R \arcsin \left(\sqrt{\sin^2 \left(\frac{\Delta lat}{2} \right) + \cos(lat_1) \cos(lat_2) \sin^2 \left(\frac{\Delta lon}{2} \right)} \right) \quad (1)$$

Dimana $R = 6371$ km (jari-jari bumi).

2. **Fitur Temporal:** *hour_of_day, day_of_week, month, is_weekend*
3. **Fitur Turunan:** *lat_diff, lon_diff, estimated_speed*
4. **Simulasi Cuaca:** *temperature, precipitation, visibility, weather_condition*

2.5 Model Machine Learning

Empat model *machine learning* dikembangkan seperti pada Tabel 1:

Tabel 1. Konfigurasi Model Machine Learning

Model	Hyperparameter
Linear Regression	Default
Random Forest	n_estimators=100, random_state=42
Gradient Boosting	n_estimators=100, learning_rate=0.1
XGBoost	n_estimators=100, max_depth=6, learning_rate=0.1

2.6 Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan metrik statistik yang umum digunakan dalam regresi. Semua perhitungan didasarkan pada hasil prediksi model terhadap *testing set*.

- *Mean Absolute Error* (MAE): Mengukur rata-rata kesalahan absolut prediksi. Nilai yang lebih rendah menandakan performa lebih baik.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- *Root Mean Square Error* (RMSE): Memberikan penalti lebih besar pada kesalahan prediksi yang besar. Nilai yang lebih rendah menandakan performa lebih baik.

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

- R^2 Score: Mengukur seberapa baik model menjelaskan variasi dalam data target. Rentang nilai 0 hingga 1, di mana nilai mendekati 1 menandakan model yang sangat baik.

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

Validasi dilakukan dengan *5-fold cross-validation* dan skenario *real-world*.

3. ANALISA DAN PEMBAHASAN

3.1 Kinerja Model

Hasil evaluasi menunjukkan perbedaan kinerja signifikan antara keempat model seperti pada Tabel 2:

Tabel 2. Perbandingan Kinerja Model

Model	MAE (detik)	RMSE (detik)	R ² Score	Training Time (detik)
Linear Regression	245,67	350,12	0,7523	1,2
Random Forest	195,41	295,34	0,8431	12,8
Gradient Boosting	192,18	290,45	0,8512	15,3
XGBoost	189,23	287,91	0,8657	8,7

Analisis menunjukkan bahwa:

- XGBoost menghasilkan performa terbaik dengan $R^2 = 0,8657$, mengindikasikan model mampu menjelaskan 86,57% variasi durasi perjalanan.
- Random Forest dan Gradient Boosting menunjukkan kinerja hampir setara namun membutuhkan waktu pelatihan lebih lama.
- Linear Regression memiliki kinerja terendah, mengkonfirmasi bahwa hubungan antar variabel bersifat non-linear dan kompleks.

3.2 Analisis Feature Importance

Berdasarkan model XGBoost, faktor paling berpengaruh terhadap durasi perjalanan adalah:

- Trip Distance** (0,28) – Variabel paling dominan
- Haversine Distance** (0,22) – Jarak geodesik menunjukkan pengaruh kuat
- Hour of Day** (0,15) – Waktu perjalanan sangat signifikan
- Weather Condition** (0,12) – Cuaca mempengaruhi durasi
- Passenger Count** (0,08) – Jumlah penumpang berpengaruh moderat

3.3 Implementasi Sistem

Sistem prediksi diimplementasikan sebagai aplikasi *desktop* menggunakan *framework* Streamlit dengan fitur:

- Antarmuka pengguna untuk input koordinat dan kondisi perjalanan
- Prediksi *real-time* dengan estimasi dalam detik, menit, dan jam
- Visualisasi data interaktif untuk analisis pola perjalanan
- Perbandingan performa model secara visual

3.4 Diskusi Hasil

Hasil penelitian konsisten dengan temuan Johnson & Lee (2021) bahwa XGBoost unggul dalam prediksi transportasi urban. Keunggulan XGBoost terletak pada:

- Kemampuan regularisasi yang mencegah *overfitting*
- Penanganan *missing values* secara otomatis
- Efisiensi komputasi dengan optimasi paralel

4. Fleksibilitas dalam menangani berbagai tipe data

Faktor geografis dan temporal terbukti paling berpengaruh, sesuai dengan karakteristik lalu lintas NYC yang sangat bergantung pada lokasi dan waktu (Singh & Kaur, 2022). Sistem yang dikembangkan memberikan akurasi lebih tinggi dibandingkan penelitian Smith et al. (2020) yang menggunakan LSTM dengan R^2 0,82.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa:

1. XGBoost merupakan model terbaik untuk prediksi durasi perjalanan taxi NYC dengan MAE 189,23 detik, RMSE 287,91 detik, dan R^2 0,8657.
2. Faktor paling berpengaruh adalah jarak perjalanan, waktu, dan kondisi cuaca.
3. Sistem prediksi yang dibangun mampu memberikan estimasi *real-time* dengan akurasi tinggi dan antarmuka intuitif.
4. Pendekatan *end-to-end* terbukti efektif dalam konteks prediksi transportasi urban. Penelitian selanjutnya dapat mengintegrasikan data *real-time* dari API cuaca dan lalu lintas, mengembangkan platform *mobile*, serta mengeksplorasi model *deep learning* untuk peningkatan akurasi lebih lanjut.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan dan kontribusi dalam pelaksanaan penelitian ini, khususnya kepada dosen pembimbing atas arahan dan bimbingan yang diberikan, serta kepada pihak-pihak yang telah menyediakan dataset dan referensi yang digunakan dalam penelitian ini. Penulis juga menyampaikan apresiasi kepada institusi dan rekan-rekan yang turut membantu hingga penelitian ini dapat diselesaikan dengan baik. Terima kasih juga disampaikan kepada NYC Taxi & Limousine Commission yang menyediakan dataset terbuka untuk penelitian ini.

REFERENCES

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer.
- Johnson, M., & Lee, S. (2021). Comparative analysis of machine learning models for urban transportation prediction. *Journal of Urban Informatics*, 8(2), 112–125.
- Kumar, A., Kaur, R., & Kumar, M. (2019). A comparative study of machine learning algorithms for travel time prediction in smart transportation systems. *Journal of Big Data*, 6(1), 1-21.
- NYC Taxi & Limousine Commission. (2023). *TLC Trip Record Data*.
- Rahman, M., & Abdullah, S. (2023). Performance comparison of regression algorithms for real-time traffic prediction. *Journal of Big Data*, 10(1), 24.
- Singh, A., & Kaur, R. (2022). Predictive analysis of taxi trip duration using machine learning models. *International Journal of Transportation Science and Technology*, 11(3), 245–256.
- Smith, R., Brown, K., & Davis, L. (2020). Neural network-based taxi trip duration prediction using GPS trajectory features. *Transportation Research Part C: Emerging Technologies*, 121, 102830.
- Utsalina, R. (2017). Analisis performa algoritma machine learning untuk prediksi waktu tempuh kendaraan. *Jurnal Teknologi Informasi*, 15(2), 89-102.