



Komparasi Algoritma Machine Learning Regresi untuk Prediksi Durasi Perjalanan Taksi di New York City

Nicholas¹, M.Taufik Wibowo², Ridwan Abdillah³, Raffi Hibban Hafizh⁴, Perani Rosyani⁵

¹⁻⁵Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia

Email: ¹Nicho002104@gmail.com, ²taufikm250@gmail.com, ³readdillah@gmail.com,
⁴Rhbbnhafizh@gmail.com, ⁵dosen00837@unpam.ac.id

Abstrak—Estimasi waktu tempuh atau *Estimated Time of Arrival (ETA)* memiliki dampak langsung terhadap efisiensi operasional armada transportasi dan kepuasan konsumen. Namun, memprediksi durasi perjalanan di kota metropolitan seperti New York City (NYC) merupakan tantangan kompleks karena dipengaruhi oleh faktor dinamis seperti jarak, waktu, dan pola lalu lintas. Penelitian ini bertujuan untuk membangun dan membandingkan kinerja model *machine learning* berbasis regresi untuk memprediksi durasi perjalanan taksi NYC. Algoritma yang diuji meliputi *Linear Regression*, *Random Forest*, *Gradient Boosting*, dan *XGBoost*. Penelitian ini menerapkan tahapan *preprocessing* data spasial-temporal dan *feature engineering* yang komprehensif. Hasil eksperimen menunjukkan bahwa model berbasis *ensemble* dan *boosting* mengungguli model linear sederhana. Secara khusus, algoritma *XGBoost* menghasilkan kinerja terbaik dengan tingkat kesalahan prediksi terendah (RMSE dan MAE minimum) dibandingkan model lainnya serta *baseline* penelitian sebelumnya.

Kata Kunci: *machine learning*, regresi, NYC Taxi, *XGBoost*, *feature engineering*, *travel time prediction*.

Abstract—Travel time estimation or *Estimated Time of Arrival (ETA)* has a direct impact on the operational efficiency of transportation fleets and customer satisfaction. However, predicting trip duration in metropolitan cities like New York City (NYC) is a complex challenge as it is influenced by dynamic factors such as distance, time, and traffic patterns. This study aims to develop and compare the performance of regression-based machine learning models to predict NYC taxi trip duration. The algorithms evaluated include *Linear Regression*, *Random Forest*, *Gradient Boosting*, and *XGBoost*. This research implements comprehensive spatial-temporal data preprocessing and feature engineering stages to enhance data quality. Experimental results demonstrate that ensemble and boosting-based models outperform simple linear models. Specifically, the *XGBoost* algorithm yielded the best performance with the lowest prediction error rates (minimum RMSE and MAE) compared to other models and the baseline from previous studies.

Keywords: *machine learning*, regression, NYC Taxi, *XGBoost*, *feature engineering*, *travel time prediction*.

1. PENDAHULUAN

Perkembangan teknologi informasi telah membawa perubahan signifikan pada sektor transportasi modern. Salah satu tantangan terbesar dalam sistem ini adalah kemampuan memprediksi durasi perjalanan secara tepat, yang krusial untuk penentuan harga dan perencanaan rute. Namun, durasi perjalanan dipengaruhi oleh faktor non-linear seperti cuaca, waktu sibuk, dan kondisi lalu lintas, yang membuat metode perhitungan manual menjadi tidak memadai.

Pendekatan regresi berbasis data menawarkan solusi yang lebih efektif dengan mempelajari hubungan kompleks antar variabel. Penelitian ini bertujuan mengembangkan model prediksi durasi perjalanan menggunakan dataset *NYC Taxi Trip Duration* dari Kaggle. Fokus utama penelitian adalah membandingkan performa algoritma *Linear Regression*, *Random Forest*, *Gradient Boosting*, dan *XGBoost* setelah melalui proses modifikasi *preprocessing* dan *feature engineering* untuk meningkatkan akurasi dibandingkan *baseline* sebelumnya.

2. METODE PENELITIAN

2.1. Dataset

Data yang digunakan adalah dataset publik *NYC Taxi Trip Duration* yang memuat informasi *pickup_datetime*, koordinat lokasi (*latitude/longitude*), dan durasi perjalanan dalam detik.

2.2. Preprocessing Data

Tahapan ini bertujuan membersihkan *noise*. Langkah yang dilakukan meliputi:



JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 9, Februari Tahun 2026
ISSN 3025-0919 (media online)
Hal 2316-2318

- Penghapusan *outlier* durasi ekstrem (≤ 1 detik atau > 7200 detik).
- Validasi koordinat agar tetap berada dalam batas geografis NYC.
- Konversi data waktu menjadi fitur tanggal, jam, dan hari.

2.3. Feature Engineering

Untuk meningkatkan kemampuan model menangkap pola, dilakukan penambahan fitur baru:

- **Jarak Tempuh:** Dihitung menggunakan rumus Haversine dan jarak Manhattan berdasarkan koordinat *pickup* dan *drop-off*.
- **Fitur Temporal:** Ekstraksi fitur jam keberangkatan, hari dalam seminggu, serta indikator *rush hour* (jam sibuk) dan akhir pekan

2.4. Algoritma Model

Empat algoritma regresi dievaluasi dalam studi ini:

1. Linear Regression: Sebagai *baseline* sederhana.
2. Random Forest Regression: *Ensemble* pohon keputusan untuk menangani data non-linear.
3. Gradient Boosting Regression: Model yang memperbaiki kesalahan model sebelumnya secara bertahap.
4. XGBoost Regression: Implementasi *boosting* yang efisien dan memiliki mekanisme regularisasi.

2.5. Evaluasi

Kinerja model diukur menggunakan metrik *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan *Root Mean Squared Error* (RMSE).

3. ANALISA DAN PEMBAHASAN

3.1. Hasil Penelitian

Berdasarkan pengujian pada data validasi, ditemukan perbedaan signifikan antar model:

1. **Linear Regression:** Menunjukkan tingkat kesalahan tertinggi karena keterbatasannya dalam menangkap hubungan non-linear pada data spasial-temporal.
2. **Random Forest:** Memberikan peningkatan performa signifikan dibanding model linear, membuktikan efektivitas metode *ensemble* dalam mengurangi *noise*.
3. **Gradient Boosting:** Menghasilkan prediksi yang lebih stabil dan presisi dibandingkan Random Forest pada beberapa kondisi.
4. **XGBoost:** Mencatatkan performa terbaik dengan error terendah. Keunggulan ini didorong oleh kemampuan regularisasi dan optimasi yang baik terhadap fitur hasil rekayasa.

3.2. Analisis Model dan Fitur

Analisis menunjukkan bahwa *feature engineering*, khususnya penambahan fitur jarak tempuh dan indikator waktu sibuk, berkontribusi besar terhadap akurasi model. Visualisasi hasil memperlihatkan bahwa model XGBoost memiliki distribusi error yang paling kecil dan merata dibandingkan model *baseline*. Modifikasi pada *pipeline* data terbukti menghasilkan akurasi yang lebih tinggi dibandingkan model *baseline* dari Kaggle.

4. KESIMPULAN

Penelitian ini menyimpulkan bahwa pendekatan *machine learning* regresi efektif untuk memprediksi durasi perjalanan taksi. Kualitas data yang ditingkatkan melalui *preprocessing* (penghapusan *outlier*) dan *feature engineering* (fitur jarak dan waktu) memegang peranan vital dalam performa model.

Dari perbandingan algoritma, XGBoost Regression terbukti menjadi model paling superior dibandingkan *Linear Regression*, *Random Forest*, dan *Gradient Boosting*. Untuk penelitian selanjutnya, disarankan menambahkan data eksternal seperti kondisi cuaca dan lalu lintas *real-time* untuk meningkatkan akurasi prediksi lebih lanjut.



JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 9, Februari Tahun 2026
ISSN 3025-0919 (media online)
Hal 2316-2318

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada dosen pengampu mata kuliah Kecerdasan Buatan atas bimbingan dan ilmu yang diberikan selama penyusunan mini project ini. Penulis juga berterima kasih kepada penyedia dataset di platform Kaggle yang telah memfasilitasi data penelitian sehingga eksperimen ini dapat terlaksana dengan baik. Terakhir, terima kasih kepada rekan-rekan mahasiswa yang telah memberikan dukungan serta diskusi yang bermanfaat selama proses pengerjaan laporan ini.

REFERENCES

- Chen, X., & Liu, Y. (2020). *The Impact of Temporal Features in Travel Time Prediction Models*. Journal of Traffic and Transportation Engineering.
- Kaggle. (2017). *NYC Taxi Trip Duration Dataset*. <https://www.kaggle.com/c/nyc-taxi-trip-duration>
- Miller, A., & Zhang, L. (2021). *Predicting NYC Taxi Trip Duration using Gradient Boosting and Spatial Features*. International Conference on Intelligent Transportation Systems.
- Nasrullah, M., et al. (2022). *Comparative Analysis of Machine Learning Algorithms for Travel Time Prediction*. Transportation Research Part C: Emerging Technologies.
- Seleck, T. (2017). *New York City Taxi Trip Duration: Complete EDA*. Kaggle.
- Wang, J. (2024). *Spatial-Temporal Modeling for Estimated Time of Arrival (ETA) in Public Transportation*. Smart City Review.