

PREDIKSI DURASI KETERLAMBATAN PENERBANGAN MENGUNAKAN ALGORITMA RANDOM FOREST

Fajar Raharjo, Amat Wiratno, Kristiawan Nugroho, Eka Ardianto

Program Studi Magister Teknologi Informasi, Fakultas Teknologi Informasi dan Industri, Universitas Stikubank
Jalan Tri Lomba Juang No 1, Indonesia
fajarraharjo0017@mhs.unisbank.ac.id

ABSTRAK

Keterlambatan penerbangan merupakan masalah krusial dalam industri transportasi udara yang berdampak pada kerugian finansial maskapai dan ketidaknyamanan penumpang. Amerika Serikat, dengan volume lalu lintas udara tertinggi di dunia, menghadapi tantangan besar dalam mengelola efisiensi jadwal. Penelitian ini bertujuan untuk membangun model prediktif machine learning untuk mengestimasi durasi keterlambatan penerbangan (dalam menit) dengan memanfaatkan algoritma Random Forest. Dataset yang digunakan bersumber dari dataset kaggle, yang mencakup fitur-fitur seperti (year, month, kode maskapai, nama maskapai, kode airport, nama airport, total jumlah penerbangan, jumlah penerbangan yang terlambat lebih dari 15 menit, penyebab, waktu delay karena keamanan, waktu delay karena keterlambatan sebelumnya). Metodologi penelitian meliputi tahap preprocessing data, penanganan missing values, serta optimasi hyperparameter pada model Random Forest. Algoritma Random Forest dipilih karena kemampuannya dalam menangani data berdimensi tinggi dan hubungan non-linear antar variabel melalui mekanisme ensemble learning. Performa model dievaluasi menggunakan metrik Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan R-squared (R^2). Penelitian ini memberikan kontribusi bagi pengelola bandara dan maskapai dalam melakukan mitigasi risiko keterlambatan secara lebih proaktif.

Kata kunci : *Random forest, machine learning, mean absolute error, root mean square error, R-squared.*

1. PENDAHULUAN

Industri transportasi udara merupakan komponen vital dalam mendukung mobilitas global dan pertumbuhan ekonomi nasional, khususnya di Amerika Serikat yang memiliki salah satu jaringan lalu lintas udara tersibuk di dunia. Namun, efisiensi operasional industri ini sering terhambat oleh masalah keterlambatan penerbangan (*flight delays*). Keterlambatan penerbangan didefinisikan sebagai perbedaan waktu antara jadwal keberangkatan atau kedatangan dengan waktu aktual keberangkatan atau kedatangan [1].

Keterlambatan tidak hanya menyebabkan kerugian finansial yang signifikan bagi maskapai penerbangan akibat pembengkakan biaya operasional dan manajemen awak pesawat, tetapi juga berdampak negatif pada tingkat kepuasan serta produktivitas penumpang. Menurut laporan *Bureau of Transportation Statistics* (BTS), ribuan penerbangan mengalami keterlambatan setiap tahunnya yang disebabkan oleh berbagai faktor kompleks, mulai dari cuaca ekstrem, kepadatan lalu lintas udara, hingga kendala teknis dan manajerial [2].

Kemampuan untuk memprediksi durasi keterlambatan secara akurat menjadi sangat krusial bagi para pemangku kepentingan. Bagi maskapai, prediksi yang presisi memungkinkan mitigasi proaktif dan alokasi sumber daya yang lebih efisien. Bagi penumpang, informasi estimasi keterlambatan membantu dalam pengambilan keputusan perjalanan yang lebih baik. Meskipun sistem manajemen lalu lintas udara tradisional telah memiliki mekanisme penjadwalan, data penerbangan yang sangat dinamis,

non-linear, dan memiliki dimensi yang besar membuat metode statistik konvensional seringkali kurang optimal dalam memberikan hasil prediksi yang akurat [3].

2. TINJAUAN PUSTAKA

Penerapan algoritma machine learning dapat berguna karena data penerbangan yang dihasilkan cukup besar serta mencakup berbagai fitur dan kombinasi yang berpotensi memengaruhi keterlambatan [4]. Dalam beberapa tahun terakhir, implementasi *Machine Learning* telah menunjukkan hasil yang menjanjikan dalam menyelesaikan permasalahan kompleks di sektor transportasi. Berbagai algoritma seperti *Support Vector Machine* (SVM), *Artificial Neural Networks* (ANN), dan *Gradient Boosting* telah diuji dalam konteks klasifikasi keterlambatan. Namun, banyak penelitian terdahulu hanya berfokus pada klasifikasi biner, yakni memprediksi apakah sebuah penerbangan akan terlambat atau tidak. Terdapat celah penelitian (*research gap*) dalam memprediksi durasi spesifik keterlambatan (regresi), yang secara teknis lebih menantang karena melibatkan variansi data yang sangat tinggi [5].

Algoritma *Random Forest* dipilih karena keunggulannya sebagai metode *ensemble learning* yang mampu menangani dataset besar dengan banyak fitur kategorikal dan numerik. *Random Forest* memiliki ketahanan yang baik terhadap *outliers* dan mampu menangkap hubungan non-linear antar variabel tanpa memerlukan asumsi distribusi data yang ketat. Selain itu, kemampuan algoritma ini dalam

memberikan skor kepentingan fitur (*feature importance*) memungkinkan identifikasi variabel mana yang paling dominan dalam memicu durasi keterlambatan [6].

Penelitian ini bertujuan untuk mengembangkan model prediksi durasi keterlambatan penerbangan di Amerika Serikat dengan memanfaatkan algoritma *Random Forest*. Dataset yang digunakan mencakup data historis penerbangan domestik yang diintegrasikan dengan variabel-variabel kunci seperti waktu keberangkatan, informasi maskapai, serta penyebab keterlambatan. Diharapkan penelitian ini dapat memberikan kontribusi pada pengembangan sistem pendukung keputusan yang lebih cerdas dalam ekosistem penerbangan modern [7].

3. METODE PENELITIAN

Penelitian ini terdiri dari beberapa tahap yang terdiri dari pengumpulan data, analisis data, data preprocessing, pembuatan model, pengujian model dan evaluasi.

3.1. Pengumpulan Data

Adapun sumber data yang digunakan dalam penelitian ini berasal dari Kaggle dengan judul *Airline Delay Cause*, dengan total data sebanyak 318.087 baris dan 21 kolom data.

3.2. Analisis Data

Tahap analisis data ini dilakukan untuk menganalisis dataset yang akan digunakan. Dengan analisis data ini dapat mengetahui apakah data memiliki noise atau tidak. Jika terdapat noise maka data harus dilakukan preprocessing agar data yang digunakan merupakan data yang valid dan berguna [8].

3.3. Data Preprocessing

Pra-pemrosesan merupakan tahapan krusial untuk meningkatkan kualitas data sebelum masuk ke tahap pemodelan. Langkah-langkah yang dilakukan meliputi :

- Pembersihan Data: Menghapus entri yang memiliki nilai kosong (*missing values*).
- Transformasi Fitur: Melakukan Encoding pada variabel kategorikal (seperti maskapai dan bandara) menggunakan teknik *One-Hot Encoding* atau Label Encoding agar dapat diproses oleh algoritma numerik.
- Normalisasi: Menggunakan *Standard Scaler* untuk menyeragamkan skala pada fitur numerik agar proses konvergensi model menjadi lebih cepat.
- Data Splitting: Data dibagi menjadi *Training Set* (80%) untuk melatih model dan *Testing Set* (20%) untuk mengevaluasi performa model pada data yang belum pernah dilihat sebelumnya [8].

3.4. Pembuatan Model

Algoritma *Random Forest* digunakan sebagai model regresi utama. *Random Forest* bekerja dengan

membangun sekumpulan pohon keputusan (*decision trees*) secara paralel melalui mekanisme *Bootstrap Aggregating* (Bagging). Dalam penelitian ini, optimasi model dilakukan melalui *hyperparameter tuning* pada parameter berikut :

- n_estimators*: Jumlah pohon dalam hutan (100).
- max_depth*: Kedalaman maksimum setiap pohon untuk mencegah *overfitting*.
- min_samples_split*: Jumlah minimum sampel yang diperlukan untuk membagi simpul internal.

Secara matematis, prediksi akhir dari Random Forest Regressor ($\hat{y}$) adalah rata-rata dari prediksi seluruh pohon individu ($hk(x)$) [9].

$$\hat{y} = \frac{1}{K} \sum_{k=1}^K hk(x) \quad (1)$$

$\hat{y}$: Nilai prediksi akhir (estimasi durasi keterlambatan dalam menit).

K : Jumlah total pohon (*trees*) yang dibangun (misalnya 100 atau 500 pohon).

$hk(x)$: Hasil prediksi dari satu pohon individu ke- k berdasarkan input data x .

$\sum_{k=1}^K hk(x)$: Simbol sigma yang berarti kita menjumlahkan semua prediksi dari pohon pertama hingga pohon terakhir.

3.5. Mean Absolute Error (MAE)

MAE digunakan untuk mengetahui rata-rata kesalahan absolut antara prediksi model dengan data aktual [10].

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

n : Jumlah total sampel data yang diuji.

y_i : Nilai aktual durasi keterlambatan.

$\hat{y}_i$: Nilai prediksi durasi keterlambatan yang dihasilkan model.

$|y_i - \hat{y}_i|$: Nilai absolut dari selisih (error).

3.6. Root Mean Squared Error (RMSE)

Metrik evaluasi model regresi yang mengukur rata-rata besar kesalahan antara nilai prediksi model dan nilai aktual, dihitung dengan mengambil akar kuadrat dari rata-rata kuadrat selisihnya, nilai RMSE yang lebih rendah menunjukkan model yang lebih akurat karena prediksinya lebih mendekati nilai sebenarnya, dan nilainya memiliki satuan yang sama dengan data yang diprediksi [10].

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

n : Jumlah total sampel data yang diuji.

y_i : Nilai aktual durasi keterlambatan.

$\hat{y}_i$: Nilai prediksi durasi keterlambatan yang dihasilkan model.

$(y_i - \hat{y}_i)^2$: Selisih antara nilai aktual dan prediksi dikuadratkan.

3.7. R-squared (R^2 Score)

Metrik ini digunakan untuk mengukur seberapa baik model mampu menjelaskan variasi pada data target.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (4)$$

4. HASIL DAN PEMBAHASAN

Setelah melakukan pelatihan model menggunakan algoritma *Random Forest* dengan dataset penyebab keterlambatan penerbangan Amerika Serikat, performa model dievaluasi menggunakan data pengujian (*testing set*). Hasil evaluasi menunjukkan bahwa model mampu memprediksi durasi keterlambatan dengan tingkat akurasi yang cukup tinggi untuk model regresi. Dan menghasilkan metrik evaluasi :

Tabel 1. Metrik Evaluasi

	Metrik	Value
0	Mean Absolute Error (MAE)	1231.132892
1	Root Mean Squared Error (RMSE)	3898.031260
2	R-Squared (R^2)	0.875782

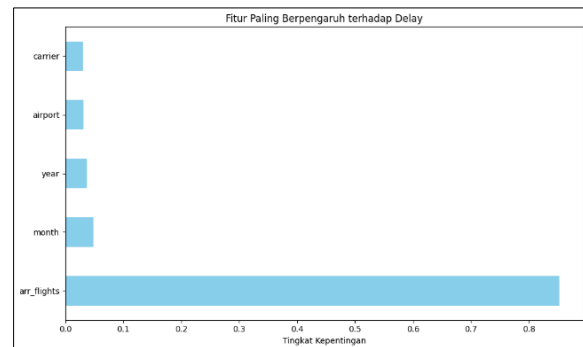
Nilai R^2 sebesar 0.8757 menunjukkan bahwa model mampu menjelaskan sekitar 87,57% variansi pada variabel dependen (target) berdasarkan variabel independen (fitur) yang digunakan. Umumnya dikategorikan sebagai model dengan tingkat kecocokan (*goodness of fit*) yang kuat. Model ini sudah cukup representatif, namun masih ada sekitar 12,43% variansi yang dipengaruhi oleh faktor lain di luar model atau adanya *noise* dalam data [11].

Nilai Mean Absolute Error (MAE) sebesar 1231.13 menunjukkan rata-rata selisih absolut antara nilai aktual dengan nilai prediksi. Secara rata-rata, prediksi model meleset sekitar 1231.13 menit dari nilai aslinya. Karena MAE bersifat linear, setiap kesalahan (error) diberi bobot yang sama. Nilai ini memberikan gambaran tentang performa model pada kondisi normal atau mayoritas data.

Nilai Root Mean Squared Error (RMSE) sebesar 3898.03 menunjukkan akar dari rata-rata kuadrat error.

Terdapat perbedaan yang cukup signifikan antara nilai MAE (1231.13) dan RMSE (3898.03). Dalam konteks statistik, jika nilai RMSE jauh lebih besar dari MAE (sekitar 3 kali lipat), ini menunjukkan bahwa model menghasilkan beberapa kesalahan prediksi dengan magnitudo yang sangat besar. Implikasi teknis, karena RMSE mengkuadratkan error sebelum dirata-ratakan, metrik ini sangat sensitif terhadap nilai ekstrem. Selisih yang lebar ini mengindikasikan bahwa meskipun model bekerja cukup baik pada sebagian besar data, ada beberapa observasi di mana model gagal memprediksi dengan akurat secara signifikan [10].

Salah satu keunggulan *Random Forest* adalah kemampuannya memberikan informasi mengenai kontribusi masing-masing fitur terhadap hasil prediksi. Berdasarkan hasil ekstraksi fitur, diperoleh urutan kepentingan variabel sebagai berikut :

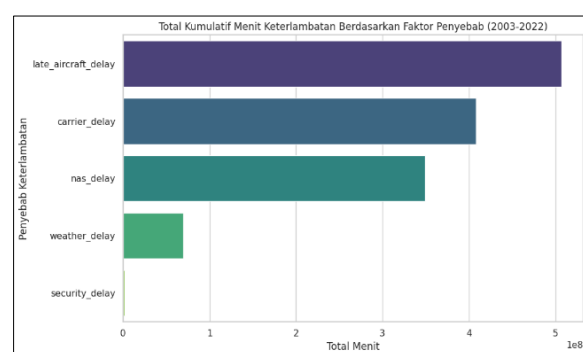


Gambar 1. Fitur paling berpengaruh terhadap delay

Fitur *arr_flights* memiliki skor kepentingan yang sangat kontras dibandingkan fitur lainnya, yaitu di kisaran 0.85 (85%). Model sangat bergantung pada jumlah penerbangan yang tiba (*arr_flights*) untuk menentukan apakah terjadi *delay* atau tidak. Secara statistik, ini berarti variabel ini memiliki korelasi linear atau non-linear yang sangat kuat dengan target. Dalam konteks manajemen bandara, kepadatan lalu lintas udara (*arr_flights*) memang faktor utama keterlambatan. Semakin banyak pesawat yang dijadwalkan tiba di waktu bersamaan, semakin besar beban infrastruktur dan kemungkinan terjadi *delay*.

Empat fitur lainnya *month*, *year*, *airport*, dan *carrier* memiliki tingkat kepentingan yang sangat rendah (masing-masing di bawah 0.05 atau 5%).

- *Month*, *year* : Faktor musiman (seperti libur panjang atau cuaca tahunan) ternyata tidak lebih berpengaruh dibandingkan volume trafik harian.
- *Airport*, *carrier*: maskapainya atau di mana bandarannya memiliki pengaruh yang kecil dalam model. Ini menunjukkan bahwa masalah *delay* dalam dataset bersifat sistemik (tergantung volume trafik) daripada spesifik pada maskapai atau lokasi tertentu.

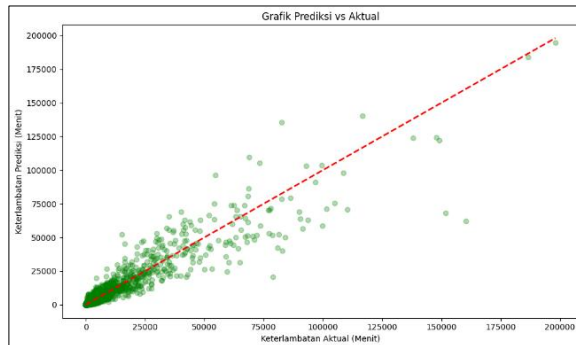


Gambar 2. Total waktu kumulatif keterlambatan berdasarkan faktor penyebab

Penyebab terbesar keterlambatan (dalam jumlah menit) adalah :

- Late Aircraft Delay (keterlambatan pesawat sebelumnya yang berdampak pada penerbangan berikutnya).
- Carrier Delay (masalah internal maskapai).

- c. Nas Delay (keterlambatan disebabkan oleh sistem penerbangan nasional, seperti kepadatan lalu lintas udara, masalah navigasi, atau keterbatasan infrastruktur bandara).
- d. Weather Delay (keterlambatan akibat faktor cuaca).
- e. Security Delay (keterlambatan akibat gangguan keamanan).



Gambar 3. Grafik prediksi Vs actual

Menunjukkan scatter plot yang membandingkan hasil prediksi model dengan data keterlambatan aktual (dalam menit).

- a. Korelasi positif yang kuat, mayoritas titik data (hijau) berkumpul mengikuti garis putus-putus merah ($y = x$). Ini menandakan bahwa model memiliki akurasi yang cukup baik dalam memprediksi keterlambatan.
- b. Akurasi pada skala kecil, model sangat akurat pada keterlambatan durasi pendek (0–25.000 menit), di mana titik-titik data sangat rapat pada garis referensi.
- c. Variansi pada nilai ekstrim: semakin besar nilai keterlambatan aktualnya (di atas 75.000 menit), persebaran titik menjadi lebih lebar. Artinya, model mulai kesulitan memprediksi secara tepat ketika terjadi gangguan besar (outlier), cenderung memberikan hasil yang sedikit lebih rendah atau lebih tinggi dari kenyataan (*underprediction* atau *overprediction*).
- d. Indikasi model: sebaran data ini umum ditemukan pada model regresi linear atau *tree-based models* yang mencoba menangkap pola dari data historis yang memiliki banyak noise [12].

5. KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan, berikut adalah kesimpulan utama dari analisis prediksi durasi keterlambatan penerbangan menggunakan algoritma *Random Forest*.

Performa dan Keandalan Model, Model *Random Forest* menunjukkan performa yang sangat kuat dengan nilai R^2 sebesar 0,8757, yang berarti model mampu menjelaskan 87,57% variansi data keterlambatan penerbangan.

Meskipun memiliki tingkat kecocokan yang tinggi, terdapat selisih signifikan antara nilai MAE

(1231,13) dan RMSE (3898,03). Hal ini mengindikasikan bahwa model sangat akurat dalam memprediksi keterlambatan pada kondisi normal, namun masih rentan terhadap kesalahan prediksi besar pada nilai-nilai ekstrem (*outliers*).

Faktor Penentu Keterlambatan, Fitur *arr_flights* (jumlah penerbangan yang tiba) merupakan variabel paling dominan dengan skor kepentingan mencapai 85%. Hal ini membuktikan bahwa kepadatan lalu lintas udara adalah faktor utama penyebab *delay*. Variabel lain seperti bulan, tahun, jenis maskapai, dan lokasi bandara memiliki pengaruh yang sangat kecil (di bawah 5%). Ini menunjukkan bahwa keterlambatan lebih bersifat sistemik akibat volume trafik, daripada spesifik pada identitas maskapai atau lokasi bandara tertentu.

Dinamika Penyebab Keterlambatan, Secara kumulatif, *Late Aircraft Delay* (efek dari keterlambatan pesawat sebelumnya) dan *Carrier Delay* (masalah internal maskapai) menjadi penyumbang menit keterlambatan terbesar. Faktor cuaca (*Weather Delay*) dan keamanan (*Security Delay*) memberikan kontribusi yang jauh lebih kecil dibandingkan masalah manajemen operasional dan infrastruktur.

Karakteristik Prediksi, Visualisasi grafik Prediksi vs Aktual menunjukkan akurasi yang sangat tinggi pada rentang keterlambatan 0–25.000 menit. Penyimpangan mulai terlihat pada keterlambatan di atas 75.000 menit, di mana model cenderung mengalami *underprediction* atau *overprediction* akibat sifat data yang memiliki banyak *noise* pada skala ekstrem.

DAFTAR PUSTAKA

- [1] E. A. Savela *et al.*, “Pengaruh Keterlambatan Penerbangan Terhadap Kepuasan Penumpang Maskapai Garuda Indonesia di Bandar Udara Jenderal Ahmad Yani Semarang,” vol. 2, no. 1, pp. 28–36, 2025.
- [2] R. Sistem, M. Genta, A. Shandi, R. Adhitama, and A. B. Arifa, “Penerapan Long Short Term Memory untuk Memprediksi Flight Delay,” vol. 1, no. 10, pp. 1–3, 2021.
- [3] A. M. Sany, C. Natalie, C. Samuel, L. Haga, M. Abriel, and G. Riser, “Begawan Abioso Evaluasi Implementasi Regulasi Hak Penumpang,” vol. 15, 2024.
- [4] S. Lolitha, S. Lukas, and F. Panduwina, “Prediksi Keterlambatan Penerbangan Menggunakan Metode Decision Tree Untuk Penentuan Premi Asuransi Perjalanan,” vol. 5, no. 2502, pp. 50–58, 2020, doi: 10.22236/teknoka.v5i.326.
- [5] M. Mamdouh, M. Ezzat, and H. A. Hefny, “A novel intelligent approach for flight delay prediction,” *J. Big Data*, 2023, doi: 10.1186/s40537-023-00854-w.
- [6] P. Keterlambatan, P. Pt, and K. Api, “PREDICTIVE ANALYTICS MENGGUNAKAN MACHINE LEARNING

- UNTUK MEMPREDIKSI WAKTU KETERLAMBATAN BERDASARKAN,” vol. 03, no. 02, pp. 165–180, 2020.
- [7] M. Y. Syam, E. F. Ramadani, V. Yoga, P. Ardhana, and E. A. M. Sampetoding, “Prediksi Flight Delay Berbasis Algoritma Neural Network,” vol. 2, no. 1, pp. 26–30, 2022.
- [8] J. Kuswanto and L. Hakim, “Penerapan Algoritma Random Forest untuk memprediksi Performa Akademik Mahasiswa,” vol. 5, no. 1, pp. 262–270, 2025.
- [9] D. Puspitasari, N. Q. Nada, A. Tri, J. Harjanta, and P. S. Informatika, “Penerapan Algoritma Random Forest Untuk Prediksi Biaya Kontruksi Berbasis Web,” vol. 8, no. 2, pp. 314–326, 2025.
- [10] M. S. Efendi and A. K. Zyen, “Penerapan Algoritma Random Forest Untuk Prediksi Penjualan Dan Sistem Persediaan Produk,” vol. 5, no. 1, pp. 12–20, 2024, doi: 10.30865/resolusi.v5i1.2149.
- [11] A. H. Arayadiba, E. K. Nurnawati, and N. Rahmawati, “Prediksi Tingkat Risiko Keterlambatan Pengiriman Barang Menggunakan Algoritma Decision Tree,” vol. 13, no. 01, pp. 51–57, 2025.
- [12] U. M. Kudus, “PENERAPAN ALGORITMA RANDOM FOREST DALAM,” vol. 14, no. September, pp. 3986–3995, 2025, doi: 10.34127/jrlab.v14i3.1770.