Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/1 (2026), 121-129. DOI: http://dx.doi.org/10.21609/jiki.v19i1.1755

Predicting Flight Departure Delay Durations Using Ensemble Learning: A Case Study of Soekarno-Hatta International Airport

Zuyina Ayuning Saputri and Denny

Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia

E-mail: zuyina.ayuning@ui.ac.id, denny@cs.ui.ac.id

Abstract

Flight delays at primary hubs like Soekarno-Hatta International Airport (CGK) can disrupt national connectivity and incur substantial operational costs. While existing research often relies on binary classification, tactical airport management requires precise temporal granularity in minutes to optimize resource allocation, such as gate and stand management. This study develops a robust duration prediction model using ensemble learning (XGBoost and Random Forest) integrated with a cost-sensitive learning strategy to address the severe skewness in delay duration distribution. The methodology incorporates advanced preprocessing, including Winsorizing to stabilize gradients and cyclical encoding to capture temporal continuity. Experimental results using 2024 operational data show that the optimized XGBoost model achieves superior performance with a Mean Absolute Error (MAE) of 6.39 minutes and an R^2 score of 0.70. Feature importance analysis identifies scheduled turnaround and ground infrastructure readiness as the primary determinants of delays, highlighting a significant "knock-on effect" where narrow transition windows fail to absorb inbound disruptions. These findings facilitate a transition from reactive reporting to proactive analytics, enabling the Airport Operation Control Center (AOCC) to optimize gate assignments and mitigate delay propagation.

Keywords: *flight departure delay prediction, ensemble learning, XGBoost, Random Forest, cost-sensitive learning*

1. Introduction

Air transport is a critical pillar of national connectivity and economic development. Within the aviation industry, On-Time performance (OTP) serves as a primary benchmark of operational efficiency [1]. High OTP levels reflect effective operational planning and resource management, whereas flight delays indicate disruptions that negatively affect the entire service delivery process.

Flight delays represent a complex operational challenge with substantial global financial implications. Previous studies estimate that delays cost the global economy more than USD 50 billion annually [2], with the United States market accounting for approximately USD 26–33 billion [3]. These costs arise from increased operational expenses, passenger compensation, and productivity losses. Moreover, flight delays exhibit a cascading effect, in which an initial delay propagates across subsequent flight schedules and airport operations [4]. Such propagation reduces the efficiency of gate allocation and ground handling activities, resulting in reactive and suboptimal airport operations [5].

In the Indonesian aviation landscape, InJourney, the national aviation and tourism holding, reported a significant recovery and growth, serving 155.9 million passengers throughout 2024, a 4% increase from the previous year. This volume comprises 118.03 million domestic and 37.90 million international passengers, highlighting the massive scale of human movement within the network [6]. Within this complex ecosystem, Soekarno-Hatta International Airport (CGK) functions not merely as a transit point, but as the primary strategic node and the epicenter of national aircraft rotation.

As the busiest hub in the region, CGK's operational stability is the fundamental determinant of the overall On-Time Performance (OTP) across the Indonesian archipelago. The critical nature of CGK in delay prediction stems from its high nodal centrality, the majority of domestic flight trajectories either originate from, terminate at, or intersect through this hub. Consequently, any operational perturbation at CGK does not remain localized. Due to the tightly coupled nature of aircraft scheduling and high utilization rates, delays at CGK are uniquely positioned to trigger asymmetric delay propagation.

In contrast to secondary airports, which often function as 'spokes' or terminal nodes where delays are typically absorbed or attenuated, CGK acts as a primary source of systemic risk. While a disruption at a secondary airport may impact only a single subsequent leg, a delay at CGK inevitably generates a compounding domino effect, cascading through multiple downstream flight segments and disrupting the synchronization of the entire national fleet. Consequently, the ability to accurately predict flight delays represents a strategic foundation for optimizing airport resource allocation, including gate management and ground personnel readiness [7]. Ultimately, focusing predictive modeling on CGK is imperative for maintaining the operational resilience of the entire Indonesian aviation network.

Despite the critical importance of delay prediction at airports, a review of the existing literature reveals a significant research gap. Most previous studies focus on binary classification approaches that predict only whether a flight will be delayed or not [8][9][10]. However, for operational stakeholders, binary information is often insufficient for tactical requirements. Precise estimation of delay duration (in minutes) is far more crucial for supporting tactical decision-making, enabling more efficient fuel planning, and providing transparent, reliable information to passengers [11][12].

This technological gap in the literature is further compounded by current operational challenges at Soekarno-Hatta International Airport. In practice, the Airport Operation Control Center (AOCC) still operates within a reactive framework, largely dependent on manual delay reporting from airlines. Crucially, the AOC team conducts gate allocation on H-1 or the evening prior to the scheduled departures, establishing the operational baseline for the following day. Currently, delay information is often shared through informal communication channels or reported after the disruption has already begun. This manual and fragmented reporting system creates a significant information lag, preventing the AOCC from optimizing gate allocation and ground resources during this critical planning window.

From a technical perspective, modeling delay duration is inherently challenged by target distribution skewness, where the frequency of short-duration delays significantly outweighs that of substantial delay events. This skewed data distribution poses a risk of model bias toward the majority of low-magnitude samples. To address this challenge, this study employs a cost-sensitive learning strategy. This technique ensures that the model assigns a higher penalty to errors involving high-magnitude delays (extreme values) [13].

Consequently, to bridge the identified literature and operational gaps, this research proposes a robust predictive model for flight delay duration in minutes at Soekarno-Hatta International Airport, specifically designed for a one-day-ahead (H-1) prediction horizon. This timeframe is strategically selected to provide data-driven insights at the exact moment the AOC team performs initial gate allocations. This study employs an ensemble learning approach integrated with cost-sensitive weighting to ensure high predictive performance despite the highly skewed distribution of the target variable. Algorithms such as Random Forest and XGBoost were selected for their proven effectiveness in handling complex aviation datasets and capturing non-linear variable relationships [14].

While popular Deep Learning architectures like Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) are adept at modeling sequential dependencies, they often require larger datasets to avoid overfitting and lack the transparency needed for operational decision-making. In the context of tabular aviation data, tree-based ensemble methods consistently demonstrate superior performance in capturing local patterns and handling feature interactions without the extensive hyperparameter tuning or black-box limitations associated with Deep Learning. By providing accurate delay estimates 24 hours prior to departure, this model serves as a strategic decision-support tool, enabling a transition from reactive reporting to proactive, data-driven gate allocation.

2. Related Works

The evolution of flight delay modeling is characterized by an increasing demand for predictive granularity. Historically, research was heavily weighted toward binary classification, focusing primarily on the occurrence of delays rather than their magnitude. However, recent scholarship has significantly pivoted toward regression-based approaches and multi-phase hybrid models to provide more actionable insights for airport stakeholders [15][16][11].

Studies at major hubs, such as Beijing Capital and John F. Kennedy International Airport, demonstrate that estimating delay duration in minutes rather than simple binary labels is essential for optimizing fuel planning and ground operations [15] [12] [14]. To handle the inherent complexity of aviation data, ensemble learning architectures like Random Forest, XGBoost, and CatBoost have emerged as the superior standard, consistently outperforming traditional statistical methods due to their ability to capture non-linear relationships and high-dimensional features [11] [17] [18] [9].

Despite these advancements, a persistent technical challenge remains in the form of skewed data distribution, where on-time flights represent the overwhelming majority. While previous researchers have employed techniques such as undersampling [11] or synthetic oversampling like SMOTE [9] to mitigate model bias, there is an increasing focus on preserving the original data integrity through more sophisticated weighting and cost-sensitive learning strategies.

While foundational studies at Soekarno-Hatta Airport, have successfully utilized Gradient Boosting with SMOTE for delay classification [9], their binary approach offers limited tactical utility for specific resource planning and relies on synthetic data that may introduce noise.

To address the limitations, this study seeks to advance the current state-of-the-art and bridge identified research gaps by:

- Shifting from classification to duration-based regression, providing the precise temporal granularity (in minutes) required for optimizing the allocation of critical airport resources, such as stands and gates, and ensuring more accurate ground handling readiness.
- Adopting a robust sample weighting strategy instead of synthetic oversampling, which ensures the model prioritizes high-impact, long-duration delay events while strictly preserving the integrity of the original operational data distribution at Soekarno-Hatta International Airport.

3. Methodology

This chapter outlines the research methodology employed in this study. Section 3.1 introduces the case study, providing the necessary context. Section 3.2 details the data preprocessing steps, followed by the feature engineering process in Section 3.3. Section 3.4 describes the techniques used for data encoding and splitting, while Section 3.5 presents the final dataset used in the experiments. The model development phase is discussed in Section 3.6, and finally, Section 3.7 explains the model evaluation metrics and procedures used to assess performance.

3.1 Case study

The case study evaluated in this study addresses the prediction of departure delay durations (in minutes) at Soekarno-Hatta International Airport using data collected between January 1 and December 31, 2024. Unlike previous approaches that were limited to binary classification (delay vs. on-time) [9], this research employs supervised learning algorithms to provide more granular duration estimates.

The dataset is constructed by merging internal operational data from the FARMS Dashboard, weather variables from the Open-Meteo API, and distance data from the Flightradar API. Due to the proprietary nature of the airport operational records, the internal dataset is not publicly available, while the external weather and distance data can be accessed via their respective public APIs. This study seeks to minimize prediction errors while providing explainability for the factors driving delay durations. The research process is divided into several stages, depicted in Figure 1 and elaborated upon in Sections 3.2 to 3.7.

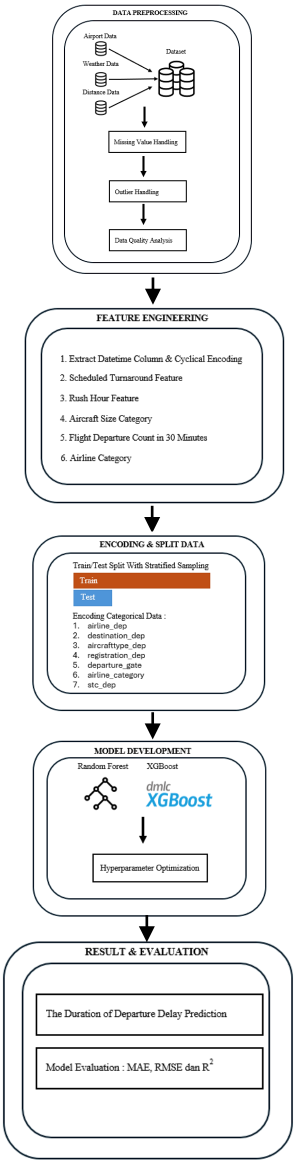

Figure 1. Methodology for predicting departure delay durations.

3.2 Data Preprocessing

The data preparation phase began by merging operational records from the Airport Operational Database (AODB) via the FARMS Dashboard with weather data from the Open-Meteo API and distance data from the Flightradar API. These datasets were joined and filtered to align with the research scope, producing a unified dataset exported in CSV format for subsequent analysis.

To ensure high data quality, the scope of this study is restricted to scheduled commercial passenger flights, specifically excluding ferry flights, cargo operations, and unscheduled services. Cleaning was performed to eliminate duplicate records and operational anomalies. Missing values were addressed through targeted strategies:

- The "*eldt (estimated landing date time)*" attribute was discarded due to significant data sparsity. Instead, this study utilizes estimated in block time (eibt) as the primary arrival metric, which is more relevant given that the focus of this research is on departure delay durations.
- Categorical imputation was applied to the "*Flighttype*" variable to distinguish between domestic and international.
- A cross-reference method was utilized to maintain the consistency of "*aircraft registration*" numbers.

Furthermore, to mitigate model bias and ensure gradient stability within the skewed dataset, outliers were managed using the Interquartile Range (IQR) method. Rather than employing traditional trimming techniques that reduce sample size, this study implemented Winsorizing (Capping) at the 99.9th percentile, as introduced by Tukey (1962, 1977). This approach strategically caps extreme values to maintain the total sample volume while preventing extreme delay fluctuations from distorting the model's learning process [19].

Finally, the consolidated dataset is detailed in Table 1. This refined data serves as the essential foundation for the subsequent feature engineering phase, where raw attributes are transformed to better capture the underlying patterns of flight delay durations.

Table 1. Summary of primary variables in the consolidated dataset.

Column	Description
eobt	estimated off block time: the estimated time a flight is expected to depart from the gate.
aobt	actual off block time: the exact moment the aircraft pushes back or leaves its parking position.
eibt	estimated in block time: the estimated time of arrival at the destination parking stand.
terminal_dep	airport terminal where the aircraft operates (e.g., terminal 1, terminal 2, terminal 3)
airline_dep	the unique identifier for the airline operator (e.g., GA, ID, JT, etc).
flighttype_dep	type of flight: domestic or international.
aircrafttype_dep	the specific designator for the aircraft frame and engine variant (e.g., B738, A320, etc).
stc_dep	standardized code for flight service type (e.g., passenger, cargo, ferry flight, etc)
distance_dep	actual distance between origin and destination airport
temperature_2m	air temperature measured at 2 meters above the ground.
dew_point_2m	the atmospheric temperature at 2 meters at which water vapor condenses into liquid.
relative_humidity_2m	the percentage of water vapor present in the air relative to the saturation point at 2 meters.
pressure_msl	mean sea level pressure: atmospheric pressure adjusted to the average sea level.
surface_pressure	atmospheric pressure measured at the ground surface level.
precipitation	total volume of water falling from the atmosphere (including rain, drizzle, etc.).
rain	specific intensity or volume of liquid rainfall.
cloud_cover	the total percentage of the sky obscured by cloud layers.
cloud_cover_low	the percentage of cloud coverage within the low-level atmospheric layer.
cloud_cover_mid	the percentage of cloud coverage within the mid-level atmospheric layer.
cloud_cover_high	the percentage of cloud coverage within the high-level atmospheric layer.
wind_speed_10m	sustained horizontal air velocity at 10 meters height
wind_direction_10m	the direction from which the wind originates.
wind_gust_10m	maximum instantaneous wind speed recorded at 10 meters height.

3.3 Feature Engineering

Following the data preprocessing phase, a rigorous feature engineering process was undertaken to transform raw operational variables into a high-dimensional feature set. This process was designed to capture the complex and non-linear dynamics of departure delays through the following six developments:

- **Extract Datetime Column & Cyclical Encoding:** Temporal attributes, including the month, day of the week (coded 0–6), and hour, were extracted from the estimated off-block time (eobt). To address the limitations of linear encoding in representing temporal continuity, a trigonometric transformation was applied to map the hour variable into a cyclical coordinate system using sine and cosine functions, ensuring the model preserves the proximity between late-night and early-morning operations.

- **Scheduled Turnaround Feature:** The scheduled turnaround time was calculated as the delta between the estimated off-block time (eobt) and the estimated in-block time (eibt). This derived variable serves as a critical proxy for the available time buffer required to absorb reactionary delays from preceding flight legs.
- **Rush Hour Feature:** To incorporate domain-specific operational constraints, several categorical and binary indicators were engineered to represent peak demand. This includes a specific *rush_hour* feature for daily peak periods, as well as *is_weekend* and *is_peak_month* (targeting June, July, and December) to account for intensified traffic during holiday seasons.
- **Aircraft Size Category**: Aircraft were categorized based on ground handling complexity into wide-body or narrow-body vessels using a threshold of 250 seats. This classification quantifies the operational demand and time required for ramp services and passenger boarding processes.
- **Flight Departure Count in 30 Minutes:** To quantify instantaneous airport throughput and congestion, a departure density metric was derived. This was achieved by aggregating the total number of scheduled departures within 30-minute sliding intervals, representing the pressure on airport infrastructure at the time of departure.
- **Airline Category:** Airlines were stratified into Low-Cost Carriers (LCC) and Full-Service Carriers (FSC). This distinction captures the influence of different business models, turnaround strategies, and fleet utilization patterns on flight delay characteristics.

3.4 Encoding and Split Data

Upon completing the feature engineering phase, the dataset was subjected to a systematic transformation and partitioning process. This stage is critical for converting qualitative operational data into a machine-readable format and ensuring that the subsequent model evaluation is conducted on a statistically representative and unbiased subset of data.

To facilitate numerical computation, label encoding was applied to the primary categorical features: *airline_dep, stc_dep, destination_dep, aircrafttype_dep, registration_dep, departure_gate,* and *airline_category.*

The operational dataset exhibits a significant distribution skewness, where the frequency of short-duration delays far outweighs those experiencing extreme delays. To mitigate this target skewness, a cost-sensitive learning strategy was employed [13]. The specific weight assignments applied to different delay magnitudes to balance the learning process are detailed in Table 2. These weights (1, 2, 4, 7, 8, 10) were derived by applying a natural logarithmic transformation to the inverse frequencies of the samples within each duration range. This approach scales the disproportionate distribution of the target variable into a computationally stable range while maintaining the necessary penalty for errors in minority target ranges. By utilizing this logarithmic inverse frequency in a cost-sensitive strategy, the model's objective function is adjusted to assign a higher cost to prediction errors in severe delay intervals. This ensures the regression remains sensitive to less frequent but critical long-duration events, preventing the model from being overly optimized toward the majority of short-duration samples.

Table 2. Cost-sensitive weighting scheme for regression.

Category	Delay Duration	Weight
Cat-5	>240 minutes	10
Cat-4	181-240 minutes	8
Cat-3	121-180 minutes	7
Cat-2	61-120 minutes	4
Cat-1	30-60 minutes	2
<30 minutes	-	1

The processed dataset was bifurcated into a training set (80%) and a testing set (20%). This partitioning was executed using Stratified Random Sampling to ensure that the distribution of the target labels, specifically the proportions of different delay severities remain consistent across both subsets. Unlike simple random splitting, the stratified approach guarantees that the testing set is a true representation of the original population, thereby providing a more robust and reliable measure of the model's generalization performance [20].

3.5 Dataset

The finalized dataset utilized in this study comprises a total of 163.087 records. Following the implementation of the stratified partitioning strategy, the data were divided into a training set of 130.469 samples (80%) and a testing set of 32.628 samples (20%). Following the categorical encoding and feature engineering phases, the final feature space comprises 55 input variables.

This balanced and stratified configuration, coupled with the sample weighting scheme detailed in Table 2, establishes the final model-ready dataset. This rigorous framework is specifically designed to ensure that the ensemble learning algorithms are trained and validated on data that accurately represents the complexities of flight delay durations across various severity categories.

3.6 Model Development

Model development is a critical phase where the prepared dataset is utilized to train supervised learning algorithms for the regression task. To address the variance and complexity of flight delays at Soekarno-Hatta International Airport, this study employs advanced ensemble techniques, specifically Random Forest and XGBoost. The development process is structured into three distinct experimental scenarios to systematically evaluate the impact of data preprocessing and optimization strategies:

- Scenario 1: Comparison of XGBoost and Random Forest regression algorithms using the dataset without outlier treatment. This scenario utilizes default library hyperparameters to establish an initial performance benchmark.
- Scenario 2: Comparison of XGBoost and Random Forest regression algorithms using default library hyperparameters following the implementation of Winsorizing (Capping) at the 99.9th percentile. This stage aims to mitigate the influence of extreme values while maintaining standard model settings.
- Scenario 3: Comparative analysis of XGBoost and Random Forest built upon the processed dataset from Scenario 2 (with outlier handling), integrated with systematic hyperparameter tuning via GridSearchCV. This scenario focuses on optimizing configurations (e.g., *max_depth*, *learning_rate*, and *n_estimators*) to achieve the highest possible precision in flight delay prediction.

During the development stage, a 5-fold Cross-Validation framework is integrated into the training process, applied to 80% of the dataset (training set), to rigorously assess model stability and prevent overfitting. This iterative approach involves partitioning the training data into five distinct subsets, where each subset serves as a validation set while the remaining four subsets are used for training. This methodology ensures that the predictive models maintain high generalizability across various data segments before final evaluation on the 20% hold-out test set.

For the most advanced stage (Scenario 3), optimal predictive performance is sought through systematic hyperparameter optimization conducted within the validation cycle using Grid Search Cross-Validation (GridSearchCV). This exhaustive search strategy is applied to the two primary ensemble algorithms: the XGBoost Regressor and the Random Forest Regressor. For the XGBoost model, the tuning process investigates critical parameters such as the number of boosting rounds (n_estimators), the learning rate, the maximum depth of the trees (max_depth), and the minimum sum of instance weight required in a child (min_child_weight). Simultaneously, the Random Forest model is refined by optimizing the number of decision trees, the maximum tree depth, the minimum samples required to split an internal node (min_samples_split), the minimum samples required at a leaf node (min_samples_leaf), and the number of features considered for the best split (max_features).

3.7 Model Evaluation

The predictive efficacy of the ensemble models is rigorously quantified through a nested validation framework. In this study, the dataset is partitioned into an 80% training set and a 20% independent hold-out test set. Within the training phase, the evaluation is executed through a tiered approach where the outcomes of Scenarios 1 and 2 serve as the foundational baseline to inform and justify the optimization strategies implemented in Scenario 3. Specifically, a 5-fold GridSearchCV (inner loop) is employed to identify the optimal hyperparameters for the XGBoost and Random Forest, utilizing Mean Absolute Error (MAE) as the primary scoring metric.

Once the best parameters are determined, a final model is retrained on the entire 80% training set. This model is then evaluated against the independent 20% test set to ensure that the reported performance metrics are derived from data that remained completely unseen during the tuning process. This comprehensive evaluation strategy ensures that the models are assessed not only for their average accuracy but also for their robustness against outliers and their ability to explain the underlying variance in flight departure delay durations at Soekarno-Hatta International Airport.

The evaluation is conducted through a multi-metric framework comprising MAE, Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2 Score). While all three metrics provide a holistic performance profile, MAE is designated as the primary benchmark for identifying the superior algorithm. The prioritization of MAE is predicated on its high interpretability in a temporal context, offering the most practical and reliable metric for airport stakeholders to gauge the accuracy of predicted delay durations in real-world operations.

4. Results and Discussion

This section presents a comparative analysis of the experimental results obtained from the three predefined scenarios. From a computational perspective, the algorithms exhibited significant variance in efficiency, which is a critical factor for real-time airport operational deployment. For a single cross-validation cycle (5 iterations), Random Forest required 450 seconds, whereas XGBoost demonstrated superior efficiency with a runtime of only 6 seconds. The consistency of the error metrics across these 5 iterations, along with the narrow margin between training and validation errors, confirms that the models were robust and effectively avoided overfitting. Furthermore, the hyperparameter tuning process using Grid Search involved a substantial computational overhead, taking approximately 3 hours to identify the optimal parameter configurations for both the Random Forest and XGBoost algorithms. The comprehensive performance metrics, including MAE, RMSE, and R^2 scores across all scenarios, are detailed in Table 3.

Table 3. Performance comparison of different scenarios.

Scenario	Algorithm	MAE	RMSE	R^2
Scenario 1	XGBoost	6.95	9.97	0.657
	Random Forest	7.07	11.13	0.573
Scenario 2	XGBoost	6.84	9.45	0.661
	Random Forest	6.95	10.22	0.604
Scenario 3	XGBoost	**6.39**	**8.89**	**0.70**
	Random Forest	6.65	9.42	0.65

Specifically, Scenario 3 showcases the models' performance after GridSearchCV optimization using the Winsorized dataset. The results indicate that the optimized XGBoost model in Scenario 3 is the best-performing model with an R^2 of 0.70 and the lowest MAE of 6.39 and RMSE of 8.89, outperforming Random Forest. Technically, XGBoost's superiority stems from its sequential boosting mechanism that iteratively minimizes residuals to reduce bias, whereas Random Forest's bagging approach only reduces variance through averaging. By utilizing L1 and L2 regularization, XGBoost more effectively manages the bias-variance trade-off, allowing it to capture the complex, non-linear patterns of flight delays without overfitting.

Moreover, the experiments show that scenarios with outlier handling (Scenarios 2 and 3) are superior to Scenario 1. The reduction in RMSE across all models in Scenario 2 most notably in Random Forest, where RMSE dropped from 11.13 to 10.22 provides empirical evidence that mitigating extreme values led to higher predictive stability. For XGBoost, this outlier mitigation was particularly vital in preventing gradient instability, as extreme residuals can cause disproportionate weight updates, thereby ensuring a more stable and precise convergence during training.

Feature importance was extracted from the best-performing model (XGBoost, Scenario 3) to interpret the predictors of flight delay duration. The top 15 features contributing to the model's accuracy are visualized in Figure 2.

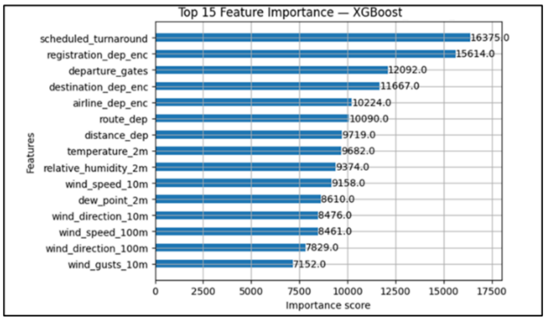

Figure 2. Feature importance.

Feature importance analysis reveals that operational ground stability is the primary driver of model accuracy. The dominance of *scheduled_turnaround* confirms a 'knock-on effect,' where narrow transition windows fail to absorb inbound delays. This vulnerability is most significant during the identified peak periods: the morning peak (04:00–07:00) and the evening peak (15:00–20:00). During these windows, high traffic density leaves minimal room for error; a short turnaround duration in these specific hours means that any delay from an arriving flight will almost certainly propagate to the next departure. Meanwhile, *departure_gates* highlights the impact of infrastructure logistics on boarding and pushback efficiency. Furthermore, the high ranking of aircraft and airline IDs (*registration_dep_enc, airline_dep_enc*) captures technical reliability variations and distinct Standard Operating Procedures. Notably, while meteorological factors are included, they serve only as supplementary constraints. Ultimately, this hierarchy proves that managerial efficiency and schedule integrity are far more deterministic of delay durations than stochastic environmental events.

Several limitations of this study should be noted. First, the use of Winsorizing in preprocessing may introduce bias by capping extreme delay values to ensure model stability. Second, while major force majeure events typically result in flight cancellations (which are excluded from this duration prediction study), the model remains sensitive to smaller-scale atypical disruptions that are inherently difficult to predict. Lastly, as the study is calibrated specifically to the operational dynamics of Soekarno-Hatta International Airport (CGK), its generalizability to airports with different infrastructure constraints requires further validation.

5. Conclusions

This study concludes that the XGBoost, optimized through hyperparameter tuning, is the superior model for predicting departure delay durations at Soekarno-Hatta International Airport. Throughout the research, the investigator was closely supervised by the airport operations team to ensure operational relevance. By leveraging a Gradient Boosting-based ensemble learning framework, the model consistently outperformed the Random Forest Regressor, achieving peak performance with a Mean Absolute Error (MAE) of 6.39 minutes and a coefficient of determination (R^2) of 0.70. These results were confirmed by the operations team, who validated that the model's accuracy is acceptable and highly applicable for delay management within a complex aviation environment.

A critical finding of this research is revealed through feature importance analysis, which identifies operational variables, specifically *scheduled_turnaround, registration_dep_enc*, and *departure_gate* as the primary determinants of delay duration. This evidence suggests that delays at Soekarno-Hatta International Airport are not merely a product of stochastic environmental factors, such as weather, but are deeply rooted in schedule integrity and the readiness of ground infrastructure. The significance of turnaround durations reinforces the presence of a "knock-on effect," where operational bottlenecks in one cycle inevitably propagate into subsequent flights.

Ultimately, these results advocate for a paradigm shift in airport management: transitioning from reactive manual reporting toward proactive predictive analytics. Implementing an accurate Predictive Duration Delay system provides the Airport Operation Control Center (AOCC) with the foresight required for strategic resource allocation. By accurately estimating delay durations, the AOCC can optimize the assignment of gates and stands, thereby mitigating delay propagation and enhancing the overall operational efficiency of the airport ecosystem.

References

[1] W. Cheevachaipimol, B. Teinwan, and P. Chutima, "Flight delay prediction using a hybrid deep learning method," *Engineering Journal*, vol. 25, no. 8, pp. 99–112, Aug. 2021, doi: 10.4186/ej.2021.25.8.99.

[2] K. Cai, Y. Li, Y. P. Fang, and Y. Zhu, "A Deep Learning Approach for Flight Delay Prediction Through Time-Evolving Graphs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11397–11407, Aug. 2022, doi: 10.1109/TITS.2021.3103502.

[3] C. Lonzius and A. Lange, "Aircraft routing clusters and their impact on airline delays," *J. Air Transp. Manag.*, vol. 114, Jan. 2024, doi: 10.1016/j.jairtraman.2023.102493.

[4] J. S. Patil, S. Kanase, G. M. Lonare, S. Khetree, P. Kiran Sawant, and F. Jahan Khan, "An Efficient Method to Detection Flight Delay Propagation based on Supplement GRU," in *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, Sep. 2023, pp. 1490–1495. doi: 10.1109/ICOSEC58147.2023.10275823.

[5] S. Wandelt, X. Chen, and X. Sun, "Flight Delay Prediction: A Dissecting Review of Recent Studies Using Machine Learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 4, pp. 4283–4297, 2025, doi: 10.1109/TITS.2025.3528536.

[6] PT Aviasi Pariwisata Indonesia (Persero), "Annual Report 2024: The Legacy of Indonesian Heritage," 2025. [Online]. Available: https://www.injourney.id

[7] PT Angkasa Pura II (Persero), *Airport Collaborative Decision-Making (A-CDM) Implementation Guidelines*, Version 1.0. Tangerang, Indonesia: PT Angkasa Pura II (Persero), 2021.

[8] Y. Tang, "Airline Flight Delay Prediction Using Machine Learning Models," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Oct. 2021, pp. 151–154. doi: 10.1145/3497701.3497725.

[9] A. Kusumasetya and E. Miranda, "Predicting Flight Delays at Soekarno-Hatta Airport Using Machine Learning: a Comparative Analysis of Random Forest and Gradient Boosting with SMOTE," in *Proceedings of the International Conference on Computer Engineering, Network and Intelligent Multimedia, CENIM 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/CENIM64038.2024.10882734.

[10] Y. Tijil, N. Dwivedi, S. K. Srivastava, and A. Ranjan, "Flight Delay Prediction Using Machine Learning Techniques," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Feb. 2024, pp. 1909–1913. doi: 10.1109/IC2PCT60090.2024.10486482.

[11] M. Beltman, M. Ribeiro, J. de Wilde, and J. Sun, "Dynamically forecasting airline departure delay probability distributions for individual flights using supervised learning," *J. Air Transp. Manag.*, vol. 126, Jun. 2025, doi: 10.1016/j.jairtraman.2025.102788.

[12] X. Wang, Z. Wang, L. Wan, and Y. Tian, "Prediction of Flight Delays at Beijing Capital International Airport Based on Ensemble Methods," *Applied Sciences (Switzerland)*, vol. 12, no. 20, Oct. 2022, doi: 10.3390/app122010621.

[13] C. Elkan, "The Foundations of Cost-Sensitive Learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, vol. 1, 2001, pp. 973–978.

[14] R. T. Reddy, P. Basa Pati, K. Deepa, and S. T. Sangeetha, "Flight Delay Prediction Using Machine Learning," in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, Apr. 2023, pp. 1–5. doi: 10.1109/I2CT57861.2023.10126220.

[15] Y. Yuan, Y. Wang, and C. S. Lai, "Multi-Attribute Data-Driven Flight Departure Delay Prediction for Airport System Using Deep Learning Method," *Aerospace*, vol. 12, no. 3, Mar. 2025, doi: 10.3390/aerospace12030246.

[16] J. Ningthoukhongjam, G. Mahesh, M. S. Alam, P. Kumar, and T. Kiran Kumar, "Feature Engineering and Hybrid Machine Learning Approach for Flight Delay Prediction," in *2nd IEEE International Conference on Data Science and Network Security, ICDSNS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICDSNS62112.2024.10690998.

[17] N. Sharma and S. Vijayalakshmi, "Flight Arrival Delay Prediction Using Deep Learning," in *TQCEBT 2024 - 2nd IEEE International Conference on Trends in Quantum Computing and Emerging Business Technologies 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/TQCEBT59414.2024.10545034.

[18] R. Khan, S. Akbar, and T. A. Zahed, "Flight Delay Prediction Based on Gradient Boosting Ensemble Techniques," in *2022 16th International Conference on Open Source Systems and Technologies (ICOSST)*, Dec. 2022, pp. 1–5. doi: 10.1109/ICOSST57195.2022.10016828.

[19] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley Publishing Company, 1977.

[20] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA: Morgan Kaufmann, 2012.