Predicción del desfase de salida de vuelos con aprendizaje automático usando datos BTS TranStats (2024)

Sebastian Pelaez Acevedo, Simón Sánchez Sepulveda
Facultad de Ingeniería, Universidad de Antioquia
Medellin, Colombia
sebastian.pelaez@udea.edu.co, simon.sanchezs@udea.edu.co

I. Introducción

Las aerolíneas y aeropuertos necesitan anticipar cuántos minutos se adelantará o retrasará cada vuelo para asignar puertas, rotar aeronaves y tripulaciones, proteger conexiones y comunicar ETAs realistas. Hoy se usan reglas generales o avisos tardíos, lo que genera ineficiencias y costos. Un modelo de *Machine Learning* (ML) regresivo que prediga el desfase en minutos integra señales operativas y externas (histórico por ruta y aerolínea, rotaciones previas, congestión, franja horaria, clima, restricciones ATC) y capta relaciones no lineales que las reglas no ven. Con pronósticos más precisos y anticipados, operaciones puede actuar antes: reprogramar recursos, evitar demoras en cascada y mejorar la experiencia del pasajero.

Este artículo se organiza así: en la Sección II se describe el problema y la base de datos; en la Sección III se justifica el paradigma de aprendizaje; la Sección IV expone el estado del arte.

II. Descripción del problema y de los datos

II-A. Fuente y construcción del conjunto de datos

La base de datos se construyó a partir de la *On-Time Performance* de *Bureau of Transportation Statistics* (BTS TranStats), un repositorio público con información de vuelos domésticos en EE. UU. actualizado mensualmente. Un autor en Kaggle consolidó todos los meses de 2024, realizó limpieza de faltantes y estandarizó nombres de características. El conjunto unificado contiene más de 7 millones de vuelos y 35 características por muestra.

Para reducir la complejidad computacional del entrenamiento se aplicó un *submuestreo* preservando la proporción de vuelos por mes, obteniendo una muestra de 10 000 registros.

II-B. Cobertura y campos

Esta base describe origen y destino, rendimiento operativo, horas programadas y reales, minutos de retraso (negativos si salió antes), cancelaciones y desvíos, entre otros. En la Tabla I se listan las variables principales con su descripción.

Cuadro I
Características del dataset y su descripción.

Nombre de la columna	Descripción
year	Año del vuelo
month	Mes del vuelo (1–12)
day_of_month	Día del mes
day_of_week	Día de la semana (1=Lunes … 7=Domingo)
fl_date	Fecha del vuelo (AAAA-MM-DD)
op_unique_carrier	Código único de la aerolínea operadora
op_carrier_fl_num	Número de vuelo de la aerolínea reportante
origin	Código del aeropuerto de origen
origin_city_name	Nombre de la ciudad de origen
origin_state_nm	Nombre del estado de origen
dest	Código del aeropuerto de destino
dest_city_name	Nombre de la ciudad de destino
dest_state_nm	Nombre del estado de destino
crs_dep_time	Hora programada de salida (local, hhmm)
dep_time	Hora real de salida (local, hhmm)
dep_delay	**Retraso en la salida en minutos (negativo si salió antes)**
taxi_out	Tiempo de rodaje de salida (min)
wheels_off	Hora en que el avión despega (local, hhmm)
wheels_on	Hora en que el avión aterriza (local, hhmm)
taxi_in	Tiempo de rodaje de llegada (min)
crs_arr_time	Hora programada de llegada (local, hhmm)
arr_time	Hora real de llegada (local, hhmm)
arr_delay	Retraso en la llegada (min; negativo si llegó antes)
cancelled	Indicador de vuelo cancelado (0=No, 1=Sí)
cancellation_code	Motivo de la cancelación (si aplica)
diverted	Indicador de vuelo desviado (0=No, 1=Sí)
crs_elapsed_time	Tiempo programado de vuelo (min)
actual_elapsed_time	Tiempo real de vuelo (min)
air_time	Tiempo en el aire (min)
distance	Distancia entre origen y destino (millas)
carrier_delay	Retraso atribuible a la aerolínea (min)
weather_delay	Retraso por condiciones meteorológicas (min)
nas_delay	Retraso por el Sistema Nacional de Aviación (NAS/ATC) (min)
security_delay	Retraso por motivos de seguridad (min)
late_aircraft_delay	Retraso por llegada tardía de la aeronave (min)

II-C. Faltantes, codificación e imputación

Las variables temporales (`dep_time`, `arr_time`, `wheels_off`, `wheels_on`) se tratan como enteros `hhmm` con extracción de atributos (hora, minuto, franja). Las categóricas (`origin`, `dest`, `op_unique_carrier`) se codifican vía *one-hot* o *target/impact encoding* según el modelo. Para faltantes en demoras por causa, se aplica imputación conservadora (cero si la causa no aplica) o mediana condicional por aeropuerto/mes; para incoherencias horarias, se descartan registros. La variable objetivo es `dep_delay` (min).

III. Paradigma de aprendizaje

El paradigma seleccionado es **regresión supervisada** para predecir el desfase continuo (minutos) respecto a la salida programada. Inicialmente se consideró una clasificación binaria (retraso vs. no retraso), pero al tratarse de una variable objetivo numérica con signo (positivos = retraso; negativos = adelanto), la regresión ofrece mayor valor operativo: permite estimar magnitudes y umbrales adaptativos, priorizar recursos y evaluar impactos de *what-if*. La evaluación se plantea con MAE y RMSE; si el coste de subestimar es mayor, puede emplearse una función de pérdida asimétrica.

IV. Estado del arte

La literatura sobre puntualidad y retrasos de vuelo se estructura en tres formulaciones: (i) **clasificación binaria** (retraso vs. no retraso) para decisiones rápidas; (ii) **regresión** del desfase en minutos, más útil para asignación fina de recursos; y (iii) **esquemas de dos etapas** que primero detectan si habrá retraso y luego estiman su magnitud. Los modelos más eficaces reportados combinan **árboles potenciados** (Gradient Boosting, XGBoost/LightGBM), **Random Forest** y **MLP/DNN**; varias propuestas incorporan **segmentación previa** (p. ej., K-Means) para entrenar clasificadores/regresores especializados por perfil operativo. La validación típica separa entrenamiento/validación/prueba y, cuando procede, se añade K-Fold o particiones temporales; se trata el desbalance con *class weights* u *oversampling*. En clasificación se priorizan precisión, *recall* y F1 (macro/ponderado), además de *accuracy*; en regresión se emplean MAE y RMSE/MSE, y a veces R^2 o regresión cuantil para intervalos.

Thiagarajan et al. presentan un *pipeline* supervisado en dos etapas con datos históricos BTS y meteorología: (1) **clasificación** (retraso sí/no) y (2) **regresión** (minutos). Comparan Random Forest, Extra-Trees, AdaBoost, **Gradient Boosting** y MLP/DNN; usan partición *train/test* (3:1) y validaciones K-Fold (K=5) y *Leave-One-Year-Out*, con *grid search* para hiperparámetros. En clasificación reportan *accuracy*, precisión y *recall*; en regresión, MSE y R^2. Sus mejores resultados son: **94,35 % de *accuracy*** en llegada con Gradient Boosting; en regresión, **Extra-Trees** obtiene **MSE=68,31** (arribo) y, tras *entrenamiento selectivo* por ruta (p. ej., ATL–CLT), **MSE=26,36** [4].

Ortega (Trabajo Final Master) estudia BTS (aeropuerto Internacional Hartsfield–Jackson de Atlanta) y llegadas a Madrid-Barajas (MAD) enriquecidas con METAR (informe meteorológico rutinario de aeródromo usado en aviación). Formula principalmente **clasificación binaria** y explora un enfoque híbrido **K-Means y Random Forest** (segmentación y clasificador especializado), además de **XGBoost** y MLP. La validación separa *train/valid/test*, emplea *early stopping* y búsquedas de hiperparámetros (grid/aleatoria); trata el desbalance con *class weights*/SMOTE. Para el **aeropuerto Internacional Hartsfield–Jackson de Atlanta (BTS)**, reporta: Naive Bayes (Acc. 85, W-F1 0,85), Reg. Logística (84, 0,84), Random Forest (86, 0,84), **XGBoost (87, 0,87)**, **MLP (87, 0,87)** y **K-Means y Random Forest (90, W-F1 0,90)**; las variables más influyentes en XGBoost son *matrícula*, *día de la semana* y *número de vuelo*. Para **MAD**, la inclusión directa de meteorología reduce el tamaño útil del set y no mejora de forma consistente: **XGBoost** (Acc. 72, W-F1 0,72) y **MLP** (Acc. 78, W-F1 0,70); destacan *callsign*, *día de la semana* y *aeronave* como atributos relevantes [3]. En conjunto, confirma que los **árboles potenciados** y la **segmentación previa** ofrecen mejor equilibrio rendimiento–interpretabilidad, y que la **calidad/volumen de datos** y las **variables operativas** aportan más señal que el clima sin mayor ingeniería.

Referencias

[1] Bureau of Transportation Statistics (BTS), "On-Time Performance (TranStats)." [En línea]. Disponible en: https://www.transtats.bts.gov/ontime/. Accedido: oct. 2025.

[2] Autor del dataset en Kaggle, "US Flights On-Time Performance 2024 (compilación BTS)." 2024. [En línea].

[3] Universitat Politècnica de València, "Predicción de retrasos de vuelos: estudio comparativo y modelos híbridos." 2020. Disponible en: https://riunet.upv.es/server/api/core/bitstreams/9fbb9fff-b9d8-45bd-bdb0-c4225b988984/content. Accedido: oct. 2025.

[4] B. Thiagarajan, L. Srinivasan, A. V. Sharma, D. Sreekanthan y V. Vijayaraghavan, "A Machine Learning Approach for Prediction of On-Time Performance of Flights," en *Proc. IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, 2017, pp. 1–10. doi: 10.1109/DASC.2017.8102138.