AAGE - Práctica 1 (MLlib):

Andrés Lires Saborido[1] and Ángel Vilariño García[2]

[1] Universidade da Coruña, andres.lires@udc.es
[2] Universidade da Coruña, angel.vilarino.garcia@udc.es

Abstract. En este proyecto se presenta un problema de clasificación y predicción de retrasos de vuelos durante el año 2019, haciendo uso de un conjunto de datos extraído de Kaggle, *2019 Airline Delays w/Weather and Airport Detail* [1]. El objetivo es desarrollar un modelo de clasificación binaria capaz de estimar si un vuelo se retrasará a partir de una serie de variables operativas y meteorológicas conocidas antes de la salida. Dado el gran volumen y la heterogeneidad de los datos, se emplea Apache Spark MLlib [2] para realizar todas las fases del trabajo: preprocesamiento de datos y el entrenamiento de diferentes modelos de aprendizaje automático a gran escala. Se comparan distintos algoritmos, así como diferentes modelos gracias a combinaciones de características o hiperparámetros para tratar de obtener el mejor modelo predictivo.

Keywords: Big Data · Machine Learning · Apache Spark · MLlib · Binary Classification · Prediction · Flight Delays · Airports · Weather · Data Analysis · Data Science

1 Introducción

1.1 Dataset

El conjunto de datos elegido es *2019 Airline Delays w/Weather and Airport Detail* [1], disponible en Kaggle, plataforma gratuita. Contiene más de 6 millones de vuelos realizados en Estados Unidos durante el año 2019, con información sobre aerolíneas, el vuelo, condiciones del aeropuerto, de la aeronave y meteorológicas.

El conjunto presenta 26 variables, explicadas con mayor detalle en la Tabla 1. Las diferentes características, tanto categóricas como numéricas, pueden ser utilizadas de manera conjunta para analizar las condiciones operativas o meteorológicas que llevan al retraso de vuelos.

El número de filas del dataset justifica el uso de Apache Spark, motor de procesamiento distribuido, para el entrenamiento de modelos de aprendizaje automático a gran escala.

1.2 Definición del problema

El objetivo del proyecto es predecir si un vuelo será retrasado o no antes de su salida, utilizando la información disponible en el dataset.

Table 1. Resumen de las variables del dataset

Variable	Tipo	Descripción
MONTH	Numérica	Mes del vuelo
DAY_OF_WEEK	Numérica	Día de la semana (1 = Lunes)
DEP_DEL15	Binaria	Salida retrasada 15 min o más (1 = Retrasado)
DEP_TIME_BLK	Categórica	Bloque horario de salida
DISTANCE_GROUP	Numérica	Grupo de distancia del vuelo
SEGMENT_NUMBER	Numérica	Vuelos previos del avión hoy
CONCURRENT_FLIGHTS	Numérica	Vuelos simultáneos en el mismo bloque
NUMBER_OF_SEATS	Numérica	Número de asientos
CARRIER_NAME	Categórica	Aerolínea
AIRPORT_FLIGHTS_MONTH	Numérica	Promedio de vuelos aeropuerto/mes
AIRLINE_FLIGHTS_MONTH	Numérica	Promedio de vuelos aerolínea/mes
AIRLINE_AIRPORT_FLIGHTS_MONTH	Numérica	Promedio de vuelos aerolínea+ aeropuerto/mes
AVG_MONTHLY_PASS_AIRPORT	Numérica	Promedio pasajeros aeropuerto/mes
AVG_MONTHLY_PASS_AIRLINE	Numérica	Promedio pasajeros aerolínea/mes
FLT_ATTENDANTS_PER_PASS	Numérica	Tripulantes por pasajero
GROUND_SERV_PER_PASS	Numérica	Personal tierra por pasajero
PLANE_AGE	Numérica	Edad del avión
DEPARTING_AIRPORT	Categórica	Aeropuerto de salida
LATITUDE	Numérica	Latitud aeropuerto
LONGITUDE	Numérica	Longitud aeropuerto
PREVIOUS_AIRPORT	Categórica	Aeropuerto previo
PRCP	Numérica	Precipitación (pulgadas)
SNOW	Numérica	Nieve caída (pulgadas)
SNWD	Numérica	Profundidad de nieve (pulgadas)
TMAX	Numérica	Temp. máxima (°F)
AWND	Numérica	Velocidad máxima viento (m/s)

Se trata de un problema de clasificación binaria supervisada, donde la variable objetivo es DEP_DEL15 (valor 1 si el vuelo salió con más de 15 minutos de retraso, 0 en otro caso). Entre las variables que nos ayudarán a lograr la predicción se encuentran características del vuelo (aerolínea, origen, destino, mes, día de la semana) y condiciones meteorológicas en el aeropuerto de origen.

Para resolver el problema se emplearán distintos modelos de Spark MLlib, haciendo uso de diferentes algoritmos, características o hiperparámetros de manera justificada. Estos modelos se evaluarán de acuerdo a métricas adecuadas al problema.

2 Análisis exploratorio de datos (EDA)

En esta sección se llevará a cabo un análisis exploratorio de los datos para comprender mejor las características del dataset y la relación entre las variables. Se utilizarán técnicas de visualización y estadísticas descriptivas para identificar patrones, tendencias y posibles problemas en los datos.

Table 2. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. **??**).

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [3], an LNCS chapter [4], a book [5], proceedings without editors [6], and a homepage [7]. Multiple citations are grouped [3–5], [3, 5–7].

Acknowledgments. A bold run-in heading in small font size at the end of the paper is used for general acknowledgments, for example: This study was funded by X (grant number Y).

Disclosure of Interests. It is now necessary to declare any competing interests or to specifically state that the authors have no competing interests. Please place the statement with a bold run-in heading in small font size beneath the (optional) acknowledgments[3], for example: The authors have no competing interests to declare that are relevant to the content of this article. Or: Author A has received research grants from Company W. Author B has received a speaker honorarium from Company X and owns stock in Company Y. Author C is a member of committee Z.

References

1. Página de Kaggle del dataset, https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations/data, descargado a 08/10/2025

[3] If EquinOCS, our proceedings submission system, is used, then the disclaimer can be provided directly in the system.

2. Página sobre MLlib en la web de Spark, https://spark.apache.org/mllib/
3. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
4. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
5. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
6. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
7. LNCS Homepage, http://www.springer.com/lncs, last accessed 2023/10/25