# Activitat 3: Modelització predictiva

Semestre 2020.2

# Índex



---

# Introducció

En aquesta activitat usarem un conjunt de dades sobre l'aeroport internacional de San Francisco (SFO). Ha estat guardonat dues vegades, com el millor aeroport a Amèrica del Nord. En aquest estudi s'analitzaran les dades de vols recollits durant l'any 2015.

L'arxiu **dat_SFO** conté aproximadament 145000 registres i 28 variables.

Les principals variables són:

- Month: Dia del mes de sortida del vol.
- Day of week: Dia de la setmana de sortida del vol.
- Airline: Nom en sigles de la companyia aèria.
- Destination Airport: Aeroport de destinació.
- Scheduled Departure: Hora de sortida del vol estimada per la companyia.
- Departure Time: Hora de sortida real del vol.
- Departure Delay: Diferència entre l'hora de sortida estimada i la real.
- Air Time: Temps real de vol en aire.
- Distance: Distància entre els aeroports origen i arribada.
- Scheduled Arrival: Hora d'arribada del vol estimada per la companyia.
- Arrival Time: Hora d'arribada del vol.
- Arrival Delay: Diferència entre l'hora d'arribada estimada i la real.

- Late Aircraft Delay: Retard per arribada tard de l'avió.
- Diverted: Indicador de vol desviat, sent zero si el vol s'ha efectuat amb normalitat i un si ha estat desviat.
- Cancelled: Indicador de vol cancel · lat, sent zero si el vol s'ha efectuat i un si no.

Cada any una quantitat considerable de vols de diferents aerolínies es retarda o cancel · la, costant al sistema de transport aeri milers de milions d'euros en pèrdues de temps i diners. En aquesta activitat es pretén realitzar un estudi dels retards dels vols, tant en sortides com arribades. Per a això, s'estudiaran les relacions entre els mateixos i diverses variables. Primer s'estudiaran les relacions lineals i posteriorment s'avaluaran els possibles factors de risc d'aquests retards.

A continuació, s'especifiquen els passos a seguir. En el lliurament, s'ha de respectar la mateixa numeració dels apartats de l'índex.

# 1 Model de regressió lineal

## 1.1 Model de regressió lineal(regresores quantitatius)

a) Estimar per mínims quadrats ordinaris un model lineal que expliqui la variable DEPARTURE_DELAY en funció de la variable ARRIVAL_DELAY. S'avaluarà la bondat de l'ajust, a partir del coeficient de determinació. Calcular el coeficient de correlació i explicar la seva relació amb el coeficient de determinació.

NOTA: En la base de dades els noms de les variables estan en majúscules.

b) S'afegirà al model anterior la variable independent DISTÀNCIA. Existeix una millora de l'ajust?. Raonar.

c) Posteriorment, es procedirà a dividir la mostra en dues, segons els vols siguin o no més llargs. Es prendrà per llarga distància aquells amb un recorregut superior a 600 milles. Ara cal repetir el model de regressió de l'apartat anterior per cada mostra. Raonar els resultats.

## 1.2 Model de regressió lineal múltiple (regresors quantitatius i qualitatius)

En aquest apartat s'estudiarà la relació de DEPARTURE_DELAY, amb les variables explicatives ARRIVAL_DELAY i LATE_AIRCRAFT_DELAY. Per això es procedirà a la recodificació de la variable LATE_AIRCRAFT_DELAY, en major i menor o igual a 15 minuts.

## 1.3 Diagnosi del model

Per a la diagnosi es tria el model construído en l'apartat b) i es pintaran dos gràfics: un amb els valors ajustats enfront dels residus (que ens permetrà veure si la variància és constant) i el gràfic quantil-quantil que compara els residus del model amb els valors d'una variable que es distribueix normalment(QQ plot). Interpretar els resultats.

## 1.4 Predicció del model

Segons el model de l'apartat b), calcular el retard en la sortida d'un avió, que després de recórrer 2500 milles ha arribat al seu destí amb 30 minuts més tard.

# 2 Model de regressió logística

## 2.1 Estudi de relacions entre variables

Es vol estudiar la probabilitat que té un avió de sofrir un retard.

Per a això, primer es crearà una nova variable dicotòmica anomenada **delay_SFO**. Aquesta nova variable està relacionada amb els valors de la variable DEPARTURE_DELAY. Es codificarà de la següent manera: Si el valor d'aquesta variable és menor a 15 minuts, es pot assumir que el vol no va amb retard i es codificarà amb el valor 0, en cas contrari, es codificarà amb el valor 1.

a) Visualitzar la relació entre delay_SFO i les variables independents:DAY_OF_WEEK i AIRLINE. Calcular les freqüències relatives per fila i columna. Interpretar el significat. Visualitzar amb barplot.

b) Per a comprovar si existeix associació entre les variable dependent i cadascuna de les variables explicatives, s'aplicarà el test Chi-quadrat de Pearson. Un resultat significatiu ens dirà que existeix associació. Interpretar.

### 2.2 Model de regressió logística

a) Estimar el model de regressió logística prenent com a variable dependent delay_SFO i variable explicativa DAY_OF_WEEK. Es prendrà com a dia de referència el dilluns. ¿Es pot considerar que el dia de la setmana és un factor de risc?. Justifica la teva resposta.

b) Idem a l'anterior prenent com a variable explicativa AIRLINE. Es prendrà com a aerolínia de referència AA. ¿Es pot considerar que l'aerolínia és un factor de risc?. Justifica la teva resposta.

c) Es crearà un model amb la variable dependent i les variable explicatives DAY_OF_WEEK (l'obtinguda en l'apartat a) i DISTANCE. S'observa una millora amb referència als anteriors?. Explicar.

d) Es crearà un nou model amb la variable dependent i prenent com a variables explicatives, aquelles que han estat significatives en els apartats anteriors, i a més s'afegirà la variable ARRIVAL_DELAY.S'observa una millora amb referència als anteriors?. Explicar. Realitzeu el càlcul de les OR.

### 2.3 Predicció

Segons el model de l'apartat c), calcula la probabilitat de retard en el vol, si el nostre destí està a 1500 milles i viatgem en dijous.

### 2.4 Bondat de l'ajust

Usa el test de Hosman-Lemeshow per a veure la bondat d'ajust, prenent el model de l'apartat c). En la llibreria ResourceSelection hi ha una funció que ajusta el test de Hosmer-Lemeshow.

### 2.5 Corba ROC

Dibuixar la corba ROC, i calcular l'àrea sota la corba amb els models dels apartats c) i d). Discutir el resultat.

## 3 Conclusions de l'anàlisi

En aquest apartat s'hauran d'exposar les conclusions sobre la base dels resultats obtinguts en tot l'estudi. Regressió lineal i logística.

---

**Puntuació dels apartats**

- Apartat 1.1 (10%)
- Apartat 1.2 (10%)
- Apartat 1.3 (10%)
- Apartat 1.4 (5%)

- Apartat 2.1 (10%)
- Apartat 2.2 (15%)
- Apartat 2.3 (10%)
- Apartat 2.4 (10%)
- Apartat 2.5 (10%)
- Apartat 3 i Qualitat de l'informe dinàmic (10%)