Ejercicio 2 Puntuable

1.- En este ejercicio utilizareis el mismo CSV que utilizasteis en el primer ejercicio y lo pasareis por algoritmos no supervisados, si no es posible utilizar el mismo CSV, utilizad otro, pero debéis comentar por qué no habéis podido utilizarlo y por qué habéis utilizado el segundo.

He utilizado el mismo CSV para esta práctica, ya que permite realizar los análisis pedidos sin problema.

2.- ¿Qué vamos a predecir esta vez?

Predecir si un vuelo determinado se retrasará, basándonos en información sobre la salida programada, la hora de salida, el aeropuerto.. etc.

3.- Hacer una batch prediction con el algoritmo que creáis más adecuado para esos datos.

Tras realizar unas pruebas con los 4 algoritmos no supervisados:

- **Clusters**: A partir de unos resultados se van haciendo agrupaciones. A partir de esas agrupaciones da resultados (posiblemente más específicos). El algoritmo intenta descubrir estructuras o patrones en los datos por sí mismo.

  Para el ejercicio se ordena y verifica las columnas que tengan mayor importancia (orden de importancia, en este caso delay, delay_1 y time), vemos si los clusters coinciden. Este es el algoritmo que más me ha convencido con mi CSV.

- **Association**: Asocié los campos delay, delay_01 y time. Explicación de los campos generados:
  - **Support**: Es la probabilidad de que dos o más ítems ocurran juntos.
  - **Lift**: Mide la importancia de la regla, comparando la probabilidad de que X e Y ocurran juntos con la probabilidad de que ocurran por separado.
  - **Confidence**: Mide la probabilidad de que Y ocurra dado que X ha ocurrido. Es una medida de cuán fuerte es la relación entre los ítems X e Y.
  - **Leverage**: Mide la diferencia entre la probabilidad observada de que dos ítems ocurran juntos y la probabilidad de que ocurran juntos de manera aleatoria. Es una forma de evaluar qué tan fuerte es la relación entre los ítems.

  Básicamente con este algoritmo, que, si pasa algo, es probable que ocurra otra cosa adicionalmente a partir de ese algo, habiendo así relación entre los datos.

- **PCA**: Con este algoritmo, lo que se hace es que se reduce el número de variables dentro de un conjunto de datos, mientras se intenta tener la mayor cantidad de información posible.
- **Anomaly**: Utilizada para identificar patrones inusuales o raros dentro de un conjunto de datos. En este caso tendremos una columna score con la probabilidad del retraso (Ordenado por score).