Minería de datos en un Data Mart

ASIGNATURA: ALMACENES Y MINERÍA DE DATOS

PRÁCTICA 5

CURSO: 2025/26

Pascual Albericio, Irene (NIP 871627)

Vázquez Martín, Lucía (NIP 871886)

Lucía Vázquez Martín, 871886
Irene Pascual Albericio, 871627

Índice

1. Introducción

En esta práctica se llevan a cabo diversos procesos de minería de datos sobre el data mart construido en sesiones anteriores, con el objetivo de analizar en profundidad los factores que influyen en los retrasos de los vuelos comerciales en Estados Unidos. Para ello, se emplea el entorno estadístico R, a partir de un fichero exportado, y se aplican técnicas como clasificación, análisis por franjas horarias y extracción de reglas de asociación. El trabajo se centra tanto en descubrir patrones relevantes como en interpretar adecuadamente los resultados obtenidos, extrayendo conocimiento útil a partir de los datos.

2. Problemas a desarrollar

En el siguiente apartado se desarrollan en detalle los distintos problemas planteados para el análisis del retraso de los vuelos, aplicando diversas técnicas de minería de datos. Cada cuestión se aborda de forma individual para identificar patrones, factores influyentes y conclusiones significativas a partir del data mart.

2.1. Problema 1: Factores relevantes en el retraso de los vuelos

Con el objetivo de detectar qué factores son los que influyen más en el retraso de los vuelos comerciales en Estados Unidos, se ha implementado el algoritmo de selección del mejor subconjunto. Para ello, se han generado iterativamente todos los modelos con las combinaciones desde 1 predictor hasta 7 predictores, utilizando bucles anidados.

Para la selección del modelo óptimo, se han comparado tres métricas: **R^2 ajustado**, que determina la capacidad explicativa del modelo, y **AIC** y **BIC**, que son criterios de información que penalizan la complejidad del modelo. Se busca que estos dos últimos valores sean lo menor posible.

A continuación, se muestra el mejor modelo encontrado para cada subconjunto probado:

```
[1] "--- COMPARATIVA DE LOS MEJORES MODELOS (p=1 hasta 7) ---"
> print(mejores_modelos[, c("Num_Variables", "R2_Ajustado", "AIC", "BIC")])
    Num_Variables R2_Ajustado      AIC      BIC
2               1 0.002693623 12539.42 12555.67
18              2 0.003091061 12539.76 12561.42
47              3 0.003478060 12540.11 12567.19
89              4 0.003604749 12540.90 12573.39
104             5 0.003700071 12541.74 12579.65
123             6 0.003324553 12543.36 12586.68
127             7 0.002809238 12545.21 12593.95
```

Figura 1. Valores R^2 ajustado, AIC y BIC para los modelos probados

En la tabla de resultados anterior, se puede observar que para todas las combinaciones el valor de R^2 ajustado es muy bajo, en torno a 0.002-0.003. Este valor no es algo extraño debido a que en el contexto a analizar influyen muchos otros factores externos, como el clima, las huelgas de empleados o el tráfico aéreo, las cuales no se están considerando en el análisis.

Por otro lado, el valor AIC y BIC es mínimo para el modelo formado por una sola variable, con unos valores de 12539.42 y 12555.67 respectivamente. Se puede apreciar que dichos valores aumentan ligeramente cada vez que se añade una nueva variable, reflejando que la complejidad que se añade al llevar a cabo esta acción no compensa la ganancia predictiva.

Por ello, y siguiendo el principio de la navaja de Occam, se selecciona como modelo óptimo el formado por una única variable: *id_operadora*. Esto indica que la compañía que gestiona el vuelo es el predictor estadístico más relevante del retraso dentro del conjunto de datos analizado, constituyendo el factor más robusto a la hora de explicar la variabilidad observada.

```
[1] "--- EL MEJOR MODELO DE TODOS (Menor AIC) ---"
> mejor_global <- resultados[which.min(resultados$AIC), ]
> print(mejor_global)
  Num_Variables    Variables R2_Ajustado      AIC      BIC
2             1 id_operadora 0.002693623 12539.42 12555.67
```

Figura 2. Modelo más influyente en el retraso de los vuelos

2.2. Problema 2: Clasificación de vuelos en función de variables

Para clasificar los vuelos en tres niveles de retraso (bajo, medio y alto), se aplicó un proceso de agrupamiento mediante K-Means sobre el retraso medio de cada entidad, lo que permitió asignar de forma automática cada aeropuerto, aerolínea o avión a uno de los tres grupos de severidad. Con el fin de mejorar la interpretación de las gráficas, se sustituyeron los identificadores internos por nombres más descriptivos y se incorporó una lógica de seguimiento para destacar entidades representativas (las que mejor, peor y moderadamente rinden) en todas las comparativas. Además, se ajustó la disposición visual de los elementos y las etiquetas para asegurar una presentación clara y legible de los resultados.

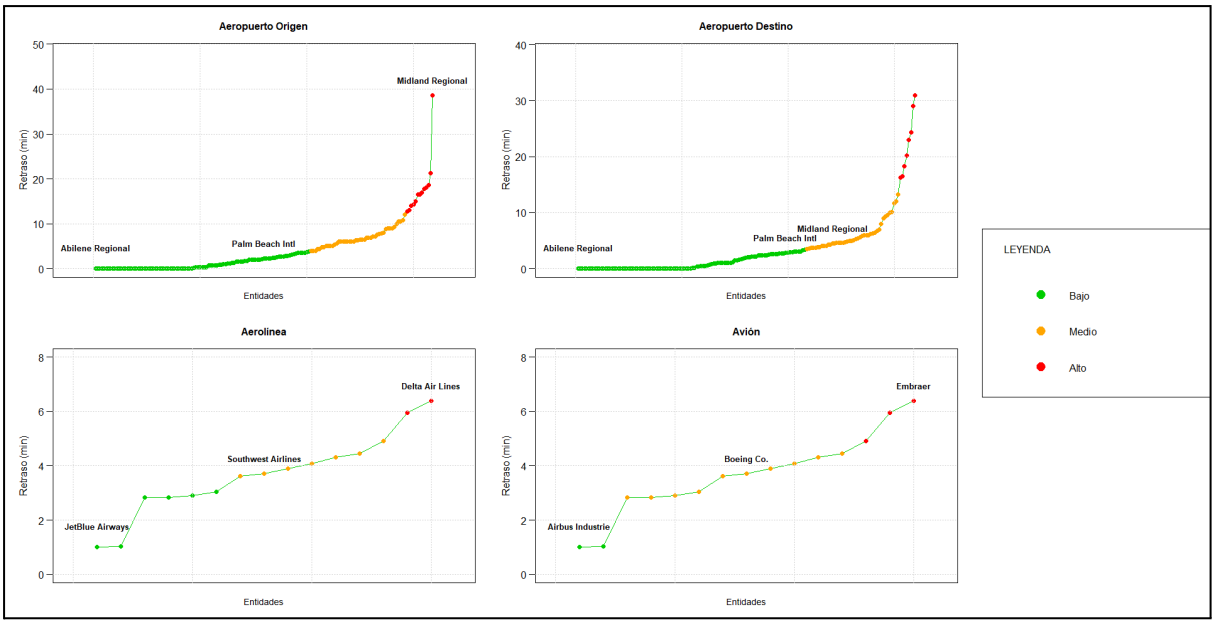

Figura 3. Gráficas de clasificación de vuelos en tres niveles según variables

Aeropuerto Origen

La distribución muestra un claro predominio del nivel de retraso bajo (verde), con la mayoría de aeropuertos situándose por debajo de los 5 minutos de demora media. A partir de este punto comienzan a aparecer entidades clasificadas en nivel medio (naranja) y únicamente un pequeño conjunto alcanza el nivel alto (rojo). El caso más destacado es *Midland Regional*, que se comporta como un outlier al superar los 35 minutos de retraso medio en las salidas.

Aeropuerto Destino

El patrón es muy similar al observado en los aeropuertos de origen: la mayor parte de los destinos presentan retrasos bajos, con una transición progresiva hacia el nivel medio en torno a los 5 minutos. Solo un grupo reducido alcanza el nivel alto. En este caso, *Midland Regional* destaca, por tener apenas 10 minutos de retraso medio en llegadas en contraste con sus casi 40 minutos en salidas. Por otro lado, *Abilene Regional* mantiene un rendimiento muy estable, situándose como el aeropuerto con menor retraso tanto en origen como en destino.

Aerolínea

En este caso la distribución es más equilibrada entre los tres niveles. Aproximadamente la mitad de las aerolíneas se encuentran en la franja de retraso bajo (hasta 3 minutos), mientras que un número significativo se concentra en el rango medio (3–5 minutos). Solo unas pocas superan estos valores y se clasifican en el nivel alto. *JetBlue Airways* destaca como la aerolínea más eficiente, con un retraso medio cercano al minuto, mientras que *Delta Air Lines* aparece como la menos favorable, superando los 6 minutos.

Avión

Los resultados por modelo de avión presentan un comportamiento muy similar al de las aerolíneas. La zona de retrasos bajos alcanza el minuto, seguida de una amplia franja correspondiente al nivel medio, que va de 3 a 5 minutos. Solo un reducido número de modelos supera ese umbral y se clasifica en el nivel alto. *Airbus Industrie* muestra el mejor desempeño con un retraso próximo a un minuto, mientras que *Embraer* se sitúa como el modelo con mayor demora, superando los 6 minutos de media.

2.3. Problema 3: Análisis de retrasos por franja horaria

Con el propósito de determinar si el momento del día influye en la puntualidad de los vuelos, se llevó a cabo un análisis de la varianza (ANOVA). Previamente, se realizó una transformación de los datos originales cruzando la tabla de hechos con la dimensión temporal para obtener la hora real y discretizarla en: Mañana, Tarde y Noche.

Se aplicó el modelo ANOVA utilizando la función aov de R para comparar las medias de retraso entre los grupos. A continuación, se presentan los resultados estadísticos obtenidos:

```
> modelo_anova <- aov(retardo_minutos ~ franja, data = datos_completos)
> print(summary(modelo_anova))
              Df Sum Sq Mean Sq F value Pr(>F)
franja         2    254   127.2   1.149  0.317
Residuals   1659 183686   110.7
```

Figura 4. Resultados del test ANOVA para evaluar la influencia de la franja horaria en el retraso.

Como se observa en la columna Pr(>F), el p-valor obtenido es de 0.317. Al ser este valor superior al nivel de significancia estándar (0.05), no existe evidencia estadística suficiente para rechazar la hipótesis nula. Esto confirma matemáticamente que la franja horaria, por sí sola, no explica la varianza en los retrasos.

Posteriormente, se ejecutó el test de Tukey para comparaciones por pares, obteniendo p-valores ajustados de 0.99, 0.49 y 0.33, lo que corrobora la inexistencia de diferencias significativas entre volar por la mañana, tarde o noche.

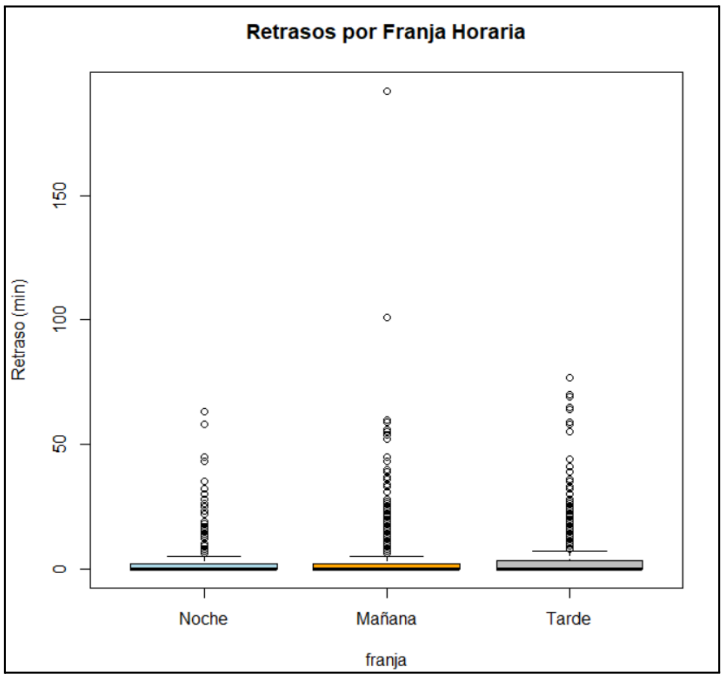

Figura 5. Boxplot comparativo de la distribución del retraso en función de las franjas horarias

2.4. Problema 4: Detección de patrones con reglas de asociación

Para profundizar en el análisis, se aplicó el algoritmo *Apriori* (librería arules) con el objetivo de descubrir combinaciones de variables que resulten en retrasos graves (>15 minutos).

Se configuró el algoritmo buscando reglas que tuvieran como consecuente (RHS) el "Retraso Grave", filtrando aquellas con un *Lift* elevado para detectar correlaciones positivas fuertes.

```
> reglas <- apriori(datos_reglas,
+                   parameter = list(supp = 0.001, conf = 0.1, minlen = 2),
+                   appearance = list(rhs = "tipo_retraso=Grave", default = "lhs"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
        0.1    0.1    1 none FALSE            TRUE       5   0.001      2     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 1

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[180 item(s), 1662 transaction(s)] done [0.00s].
sorting and recoding items ... [146 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [98 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
> reglas_ordenadas <- sort(reglas, by = "lift")
>
> print("--- TOP 10 REGLAS DE ASOCIACIÓN ---")
[1] "--- TOP 10 REGLAS DE ASOCIACIÓN ---"
> inspect(head(reglas_ordenadas, 10))
     lhs                                                                                rhs                      support     confidence coverage    lift      count
[1]  {nombre_origen=Will Rogers World, franja=Tarde}                                 => {tipo_retraso=Grave} 0.001203369 1.0000000  0.001203369 11.622378 2
[2]  {nombre_origen=Midland Regional Air Trml}                                       => {tipo_retraso=Grave} 0.001203369 0.6666667  0.001805054  7.748252 2
[3]  {nombre_aerolinea=Envoy Air, nombre_origen=Miami International, franja=Noche}   => {tipo_retraso=Grave} 0.001203369 0.6666667  0.001805054  7.748252 2
[4]  {nombre_origen=Jackson Hole}                                                    => {tipo_retraso=Grave} 0.001203369 0.5000000  0.002406739  5.811189 2
[5]  {nombre_origen=Jackson Hole, franja=Mañana}                                     => {tipo_retraso=Grave} 0.001203369 0.5000000  0.002406739  5.811189 2
[6]  {nombre_origen=Tulsa International, franja=Mañana}                              => {tipo_retraso=Grave} 0.001203369 0.5000000  0.002406739  5.811189 2
[7]  {nombre_aerolinea=Mesa Airlines Inc., nombre_origen=Reno/Tahoe International}   => {tipo_retraso=Grave} 0.001203369 0.5000000  0.002406739  5.811189 2
[8]  {nombre_aerolinea=Envoy Air, nombre_origen=Cincinnati/Northern Kentucky International} => {tipo_retraso=Grave} 0.001203369 0.5000000  0.002406739  5.811189 2
[9]  {nombre_aerolinea=Mesa Airlines Inc., nombre_origen=Reno/Tahoe International, franja=Tarde} => {tipo_retraso=Grave} 0.001203369 0.5000000  0.002406739  5.811189 2
[10] {nombre_origen=Tulsa International}                                             => {tipo_retraso=Grave} 0.001203369 0.4000000  0.003008424  4.648951 2
```

Figura 6. Extracción de reglas de asociación mediante algoritmo Apriori y patrones con mayor Lift

La regla más destacada identificó que los vuelos con origen en el aeropuerto “Will Rogers World”, durante la franja de Tarde presentaron una probabilidad del 100% de sufrir retrasos graves en los registros observados, con un Lift de 11.62. Esto implica que es 11 veces más probable sufrir un retraso grave en esta situación específica que en un vuelo promedio.

También se detectaron otros patrones relevantes como la combinación de la aerolínea “Envoy Air” operando desde “Miami International” por la noche, la cual mostró un riesgo de retraso grave 7.7 veces superior a la media (Lift = 7.7).

Estos resultados demuestran que aunque las variables como la hora no sean significativas a nivel global, como se vió en la ANOVA, sí existen combinaciones específicas de aeropuerto, aerolínea y franja horaria que garantizan un alto riesgo de impuntualidad.

3. Tabla de esfuerzos invertidos

A continuación, se detallan las tareas realizadas para el desarrollo de la práctica. De acuerdo con la metodología de trabajo establecida, todas las actividades se llevaron a cabo de forma conjunta, mediante programación en pareja y revisión simultánea de los resultados.

Apartados del trabajo realizados	Lucía	Irene
ETL y preprocesamiento	2	2
Selección de variables (ejercicio 1)	1.5	1.5
Clustering y visualización (ejercicio 2)	2	2
Análisis ANOVA (ejercicio 3)	2	2
Reglas de Asociación (ejercicio 4)	2.5	2.5
Memoria	3.5	3.5
	12.5	**12.5**

4. Conclusiones

El análisis nos ha permitido ver que el retraso aéreo no es un problema uniforme.

Los modelos lineales mostraron que añadir complejidad no mejora la predicción, siendo la **operadora** el factor individual más decisivo. Además, la clasificación con **K-Means** confirmó que, aunque la mayoría de entidades funcionan correctamente, existe un pequeño grupo de "alto riesgo" claramente diferenciado.

Lo más interesante surgió al cruzar los resultados. Aunque el **ANOVA** indicó que la hora del día no es significativa por sí misma, las **reglas de asociación** matizaron esto encontrando excepciones graves. Se detectaron patrones muy concretos (como ciertos orígenes por la tarde) donde el fallo es casi seguro. Por tanto, la solución no parece estar en medidas generales, sino en controlar esos puntos críticos específicos que se han localizado.