Mineria_LAB_2

Joel Jaquez, Luis Gonzalez, Fabian Morales

2026-02-23

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(cluster)
library(fpc)
library(NbClust)
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(hopkins)
library(GGally)
library(pheatmap)
library(hopkins)
```

```
# Cargamos el dataset desde la carpeta Datos
movies <- read.csv("./Datos/movies_2026.csv")
```

Clustering

Ejercicio 1.1 Haga el preprocesamiento del dataset, explique qué variables no aportan información a la generación de grupos y por qué. Describa con qué variables calculará los grupos.

```
# Seleccionar solo las variables numéricas que aportan información de éxito y magnitud
datos <- movies[, c("popularity", "budget", "revenue", "runtime", "voteCount", "voteAvg")]

# Quitamos valores nulos
set.seed(123)
datos <- datos[complete.cases(datos),]
```

```r
# Escalado los datos
datos_scaled <- scale(datos)
```

Para la generación de los grupos, se descartaron las variables categóricas, fechas y de texto libre (como id, title, director, genres y releaseDate) , ya que los algoritmos de clustering se basan en el cálculo de distancias matemáticas (como la euclidiana) y su inclusión requeriría transformaciones que aumentarían drásticamente la dimensionalidad, introduciendo ruido y dificultando la interpretación de los clústeres. En su lugar, el agrupamiento se calculará utilizando exclusivamente variables numéricas que capturan la magnitud financiera y la recepción del público: popularity, budget, revenue, runtime, voteCount y voteAvg . Finalmente, estas variables seleccionadas fueron previamente escaladas para asegurar que las diferencias extremas de magnitud (por ejemplo, presupuestos en millones de dólares frente a calificaciones en una escala de 0 a 10) no dominen ni sesguen el cálculo de las distancias en el algoritmo.

Ejercicio 1.2 Analice la tendencia al agrupamiento usando el estadístico de Hopkings y la VAT (Visual Assessment of cluster Tendency). Esta última hágala si es posible, teniendo en cuenta las dimensiones del conjunto de datos. Discuta sus resultados e impresiones.

```r
# Calcular el estadístico de Hopkins
set.seed(123)
valor_hopkins <- hopkins(datos_scaled)
print(paste("Valor de Hopkins:", valor_hopkins))
```

```
## [1] "Valor de Hopkins: 0.999999535140567"
```

Para analizar la tendencia al agrupamiento, se calculó el estadístico de Hopkins utilizando la totalidad de los datos numéricos escalados, obteniendo un valor de 0.9999. Al estar este resultado sumamente alejado de 0.5 (y prácticamente en 1), se rechaza de forma contundente la hipótesis de aleatoriedad espacial, lo que confirma que las películas poseen una altísima tendencia natural a formar agrupaciones reales y estructuradas. Respecto a la Evaluación Visual de Tendencia (VAT), se tomó la decisión técnica de omitir su generación gráfica, esto se justifica por las altas dimensiones del conjunto de datos (19,883 registros), ya que procesar una matriz de distancias de tal magnitud saturaría la memoria computacional. Por lo tanto, la pertinencia de aplicar algoritmos de clustering queda plenamente respaldada y demostrada por la prueba de Hopkins.