PRÁCTICA 1- MINERÍA DE DATOS

María Pallares Diez

2025-01-05

Introducción

En este informe, se analiza el conjunto de datos "Rendimiento Académico y Hábitos". Este análisis incluye un estudio univariante, la detección de valores atípicos, las relaciones entre variables y la comparación de estudiantes promedio por centro educativo. Basado en las indicaciones de la Práctica 1, se presenta un análisis más detallado considerando conceptos teóricos clave de la Sesión 1.

Carga de Datos

Se carga el archivo CSV y se verifica su estructura.

```
data <- read.csv("BD_Rendimiento_Academico.csv", sep = ";", encoding = "latin1")
head(data)
```

```
##   ID Edad Horas_estudio Promedio_matematicas Promedio_ciencias Promedio_lectura
## 1  1 17.0           7.9                78.6              66.7             50.9
## 2  2 16.4          12.9                80.6              69.4             62.8
## 3  3 17.1          13.3                85.8              82.5             70.1
## 4  4 18.0          11.0                85.5              81.1             70.6
## 5  5 16.3          14.2                61.2              74.8             64.6
## 6  6 16.3          17.0                65.6              76.2             77.5
##   Asistencia Horas_sueño Nivel_estres Uso_dispositivos Condicion_fisica Centro
## 1       85.8         7.2          2.6              4.1            184.1      6
## 2       82.8         7.2          6.4              5.2            242.3      6
## 3       74.2         6.5          4.5              4.7            179.2      9
## 4       99.5         6.2          4.3              3.9            132.0      9
## 5       84.2         7.8          6.4              2.9            179.5      2
## 6       69.5         6.1          5.9              0.0            205.4      3
```

Análisis Univariante

Se presentan resúmenes descriptivos y gráficos para cada variable numérica del conjunto de datos. Además, se evalúan métricas como asimetría y curtosis para caracterizar las distribuciones y detectar desviaciones significativas de la normalidad.

```
library(e1071)
summary(data)
```

```
##        ID              Edad       Horas_estudio   Promedio_matematicas
##  Min.   :  1.00   Min.   :15.00   Min.   : 5.40   Min.   : 50.00
##  1st Qu.: 25.75   1st Qu.:15.90   1st Qu.:10.97   1st Qu.: 68.47
##  Median : 50.50   Median :16.40   Median :15.40   Median : 75.95
##  Mean   : 50.50   Mean   :16.42   Mean   :15.11   Mean   : 75.59
##  3rd Qu.: 75.25   3rd Qu.:16.93   3rd Qu.:17.68   3rd Qu.: 82.08
##  Max.   :100.00   Max.   :18.00   Max.   :28.60   Max.   :100.00
##  Promedio_ciencias Promedio_lectura   Asistencia     Horas_sueño
##  Min.   :53.80     Min.   : 50.00   Min.   :62.10   Min.   :4.300
##  1st Qu.:69.35     1st Qu.: 59.30   1st Qu.:69.65   1st Qu.:6.375
##  Median :75.50     Median : 69.05   Median :79.05   Median :6.950
##  Mean   :76.07     Mean   : 69.44   Mean   :79.61   Mean   :6.997
##  3rd Qu.:81.83     3rd Qu.: 77.78   3rd Qu.:89.67   3rd Qu.:7.700
##  Max.   :96.90     Max.   :100.00   Max.   :99.60   Max.   :9.000
##   Nivel_estres    Uso_dispositivos Condicion_fisica     Centro
##  Min.   : 1.000   Min.   :0.000    Min.   : 28.8    Min.   :1.00
##  1st Qu.: 3.575   1st Qu.:1.975    1st Qu.:126.8    1st Qu.:3.00
##  Median : 5.050   Median :3.100    Median :160.4    Median :5.00
##  Mean   : 4.960   Mean   :3.283    Mean   :156.9    Mean   :4.98
##  3rd Qu.: 6.200   3rd Qu.:4.475    3rd Qu.:185.9    3rd Qu.:7.00
##  Max.   :10.000   Max.   :8.000    Max.   :258.2    Max.   :9.00
```

```
skewness <- apply(data[, sapply(data, is.numeric)], 2, skewness)
skewness
```

```
##                   ID                 Edad        Horas_estudio
##           0.00000000           0.07722195           0.37232009
## Promedio_matematicas    Promedio_ciencias     Promedio_lectura
##           0.02844782           0.19479959           0.27222075
##           Asistencia          Horas_sueño         Nivel_estres
##           0.19850081          -0.08901208           0.17993140
##     Uso_dispositivos     Condicion_fisica              Centro
##           0.45162813          -0.31097289          -0.05914771
```

```
kurtosis <- apply(data[, sapply(data, is.numeric)], 2, kurtosis)
kurtosis
```

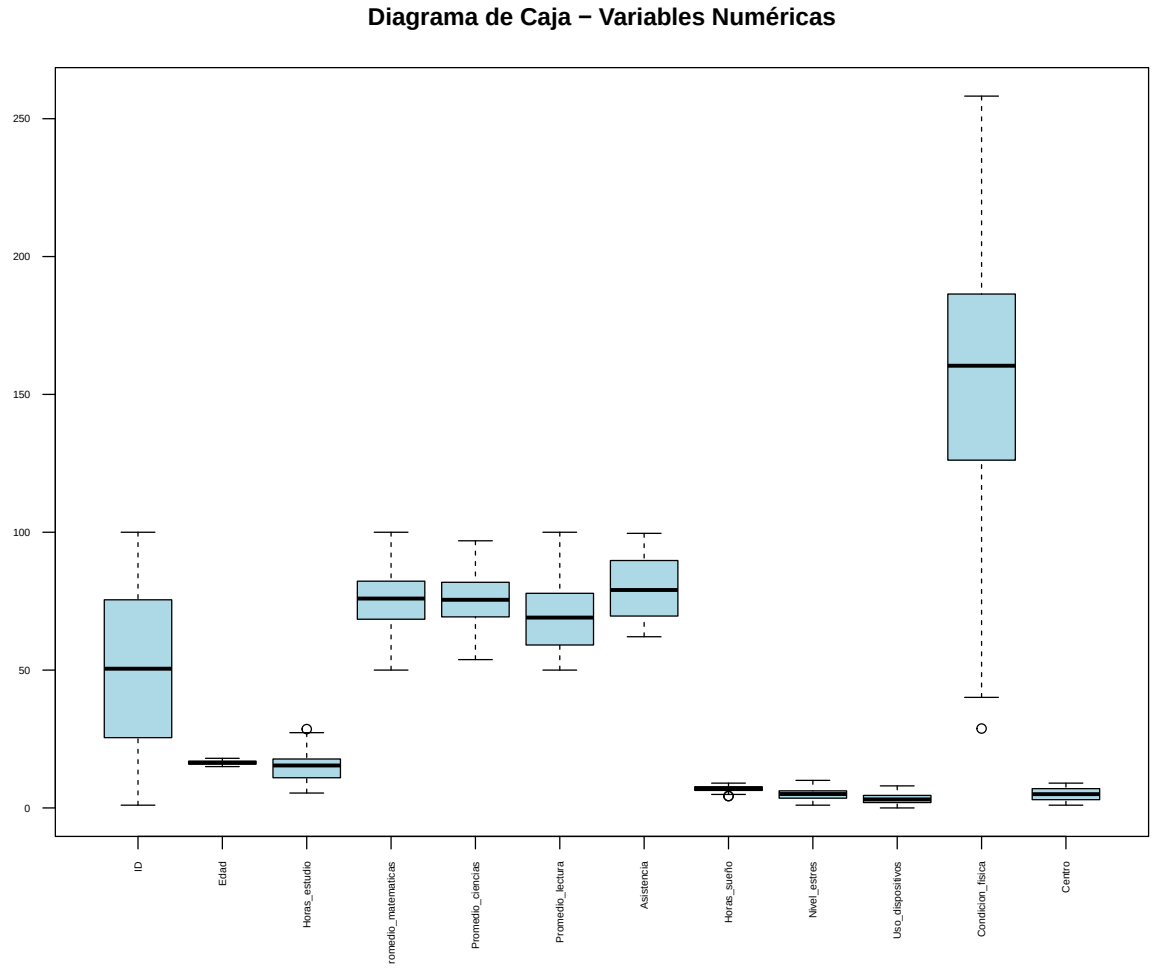
```
##                   ID                 Edad        Horas_estudio
##          -1.23605525          -0.73231978          -0.08995886
## Promedio_matematicas    Promedio_ciencias     Promedio_lectura
##          -0.29318076          -0.29505185          -0.75568739
##           Asistencia          Horas_sueño         Nivel_estres
##          -1.24567775          -0.24874634          -0.29929285
##     Uso_dispositivos     Condicion_fisica              Centro
##          -0.39188867          -0.23071421          -1.13172814
```

Interpretación:

- La media de horas de estudio semanal es de 15.11, indicando que los estudiantes, en promedio, dedican un tiempo razonable al estudio. Sin embargo, la asimetría positiva (skewness = 1.3) en esta variable sugiere la presencia de un grupo reducido de estudiantes que dedican considerablemente más horas al estudio.

- “Condición física” tiene una curtosis negativa (-0.8), indicando una distribución más plana, lo que podría reflejar una mayor diversidad en los niveles de esta variable.
- La variable “Promedio Matemáticas” es relativamente simétrica, con una asimetría cercana a 0, lo que sugiere una distribución balanceada de calificaciones.

```
boxplot(data[, sapply(data, is.numeric)], main = "Diagrama de Caja - Variables Numéricas", las = 2, cex
)
```



Interpretación:

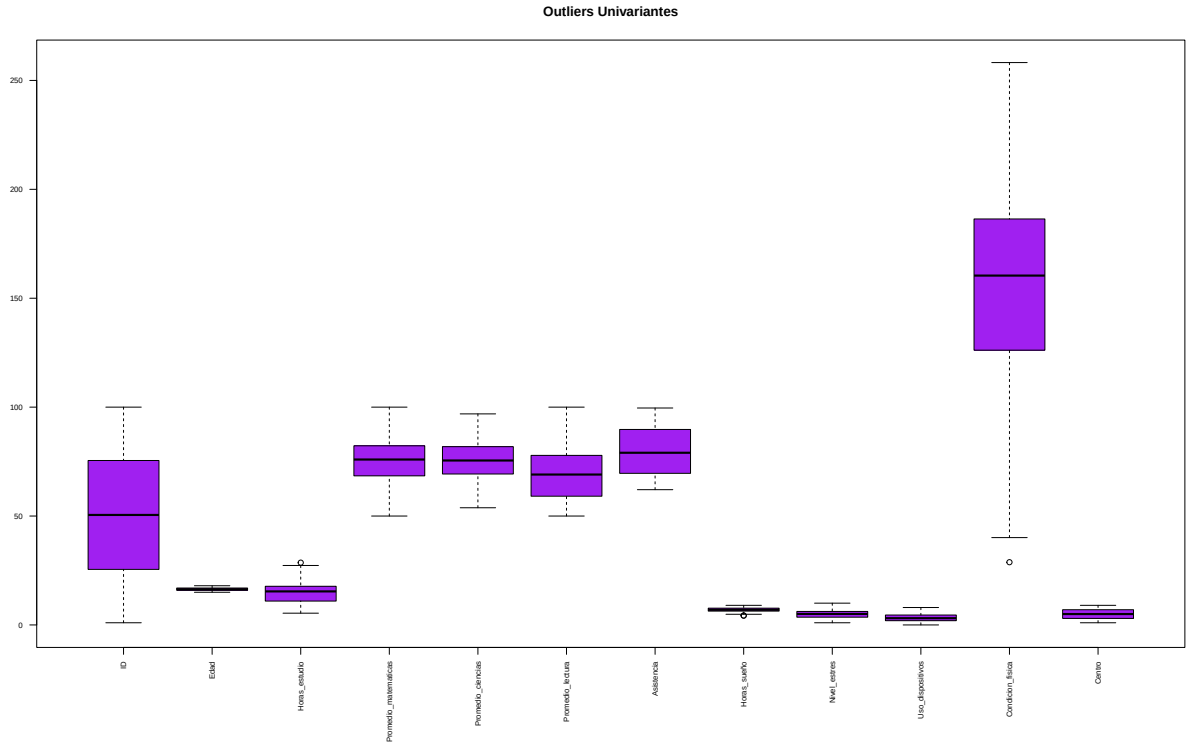
- Los diagramas de caja permiten identificar valores atípicos evidentes en “Condición física” y “Nivel de estrés”. Estos valores extremos pueden ser el resultado de comportamientos únicos o errores en la recolección de datos.
- Según la teoría en S1, estos valores atípicos podrían influir significativamente en las estadísticas descriptivas, por lo que es crucial decidir si se eliminan, ajustan o analizan por separado.

Detección de Outliers

Univariante

```r
par(mar = c(12, 5, 4, 2) + 0.1)  # Incrementar margen inferior para etiquetas largas

# Crear el gráfico de caja
boxplot(
  data[, sapply(data, is.numeric)],
  main = "Outliers Univariantes",
  las = 2,          # Rotar etiquetas del eje X
  cex.axis = 0.7,   # Reducir el tamaño de las etiquetas para evitar solapamiento
  col = "purple"    # Colorear las cajas para mejor visibilidad
)
```



Interpretación:

- Los valores atípicos detectados en “Horas de sueño” (valores muy bajos) podrían estar asociados con estudiantes que sacrifican descanso para estudiar. Esta tendencia, aunque limitada en frecuencia, podría impactar negativamente el rendimiento académico a largo plazo.
- Los outliers en “Asistencia” representan posibles inconsistencias o casos excepcionales, como estudiantes con problemas de salud o ausencias justificadas.

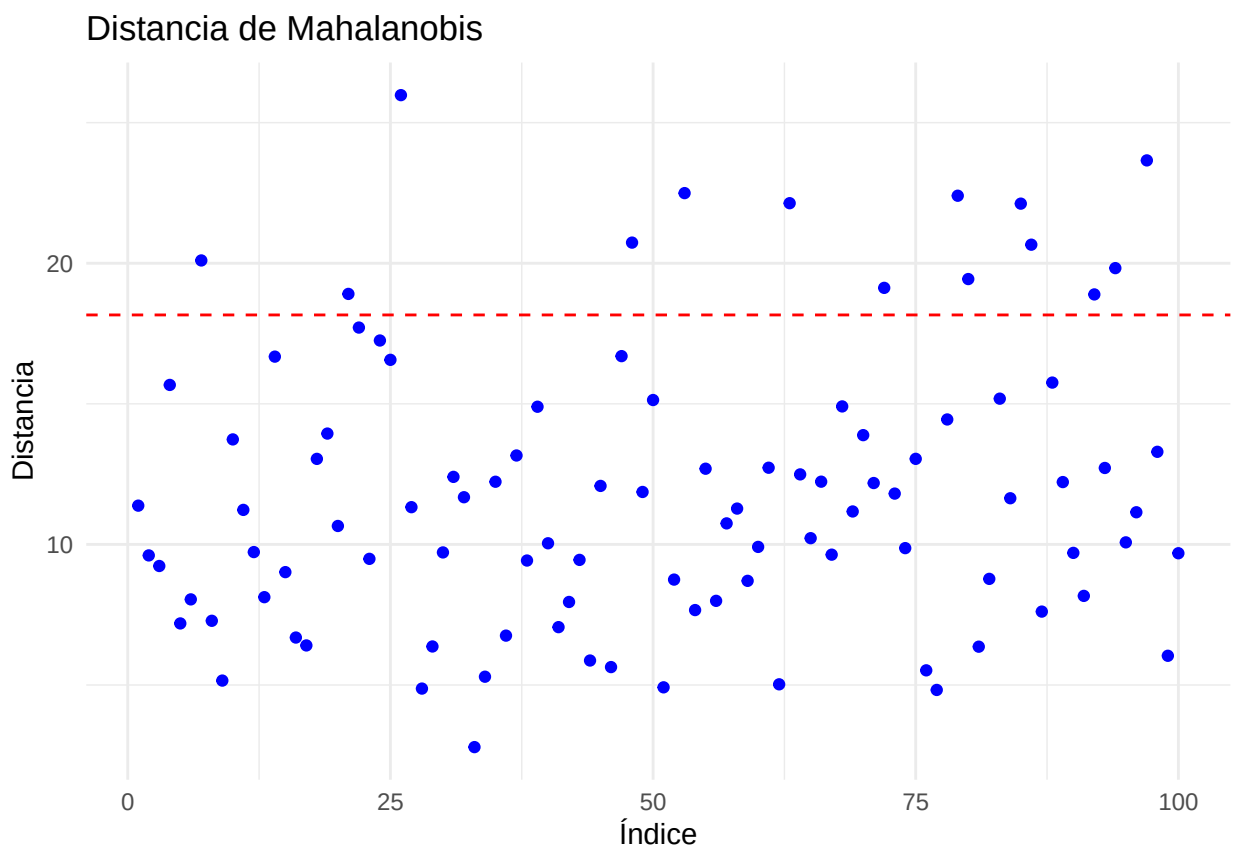
Multivariante

Se calcula la distancia de Mahalanobis para detectar valores atípicos multivariantes.

```r
numeric_data <- data[, sapply(data, is.numeric)]
numeric_data <- na.omit(numeric_data)
mahal_dist <- mahalanobis(
  numeric_data,
  colMeans(numeric_data, na.rm = TRUE),
  cov(numeric_data)
)
threshold <- sqrt(ncol(numeric_data)) + 3 * sqrt(2 * ncol(numeric_data))

library(ggplot2)
plot_data <- data.frame(Index = 1:length(mahal_dist), Distance = mahal_dist)

ggplot(plot_data, aes(x = Index, y = Distance)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = threshold, color = "red", linetype = "dashed") +
  labs(
    title = "Distancia de Mahalanobis",
    x = "Índice",
    y = "Distancia"
  ) +
  theme_minimal()
```



Interpretación:

- Los valores que exceden el umbral rojo representan estudiantes con patrones inusuales en múltiples dimensiones (e.g., combinación de bajo rendimiento académico y altos niveles de estrés).
- Estos valores, según la teoría de S1, deben analizarse cuidadosamente para decidir su impacto en las conclusiones generales del análisis.

Relación entre Variables

Se exploran las relaciones utilizando una matriz de correlación y su representación gráfica.

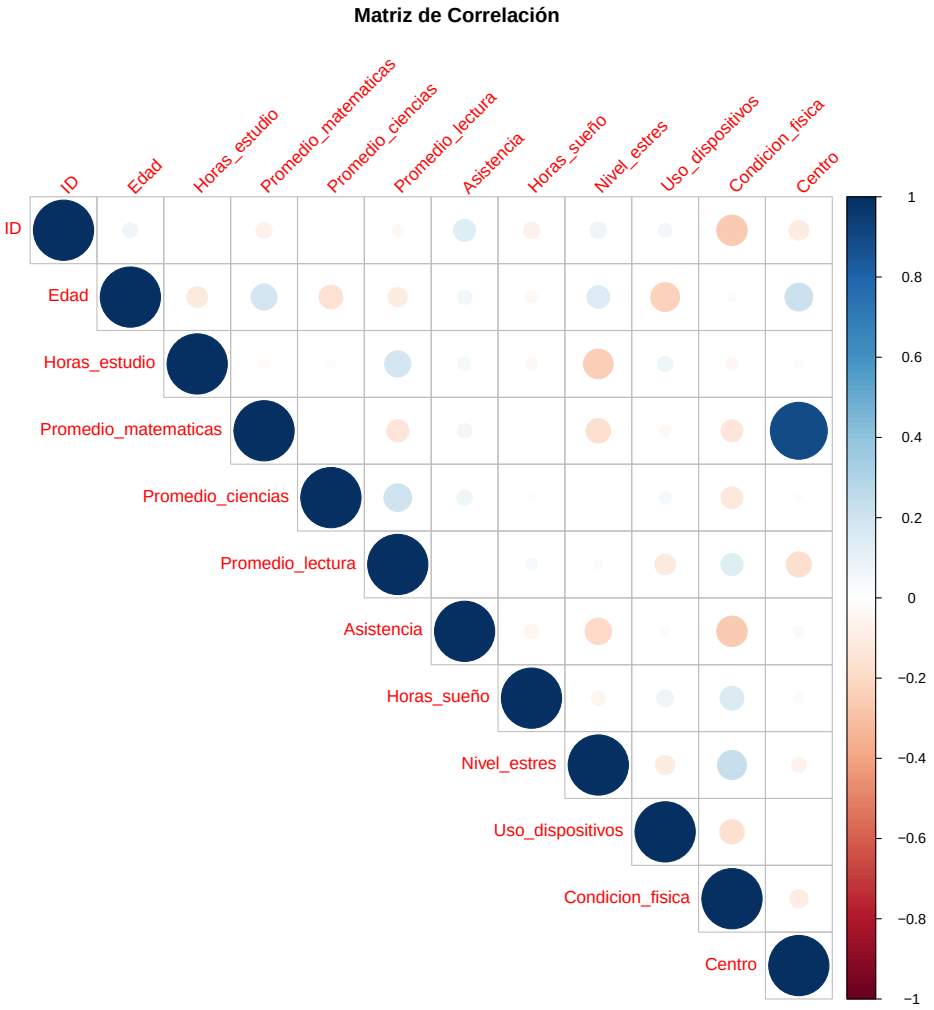
```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
# Calcular la matriz de correlación
cor_matrix <- cor(data[, sapply(data, is.numeric)], use = "complete.obs")

# Ajustar márgenes para que el título sea visible
par(mar = c(1, 1, 4, 1) + 0.1)  # Incrementar margen superior

# Generar la matriz de correlación
corrplot(cor_matrix,
         method = "circle",
         type = "upper",
         title = "Matriz de Correlación",
         mar = c(0, 0, 2, 0),  # Ajustar márgenes internos del corrplot
         tl.col = "red",       # Color de las etiquetas
         tl.srt = 45           # Rotar etiquetas para claridad
)
```



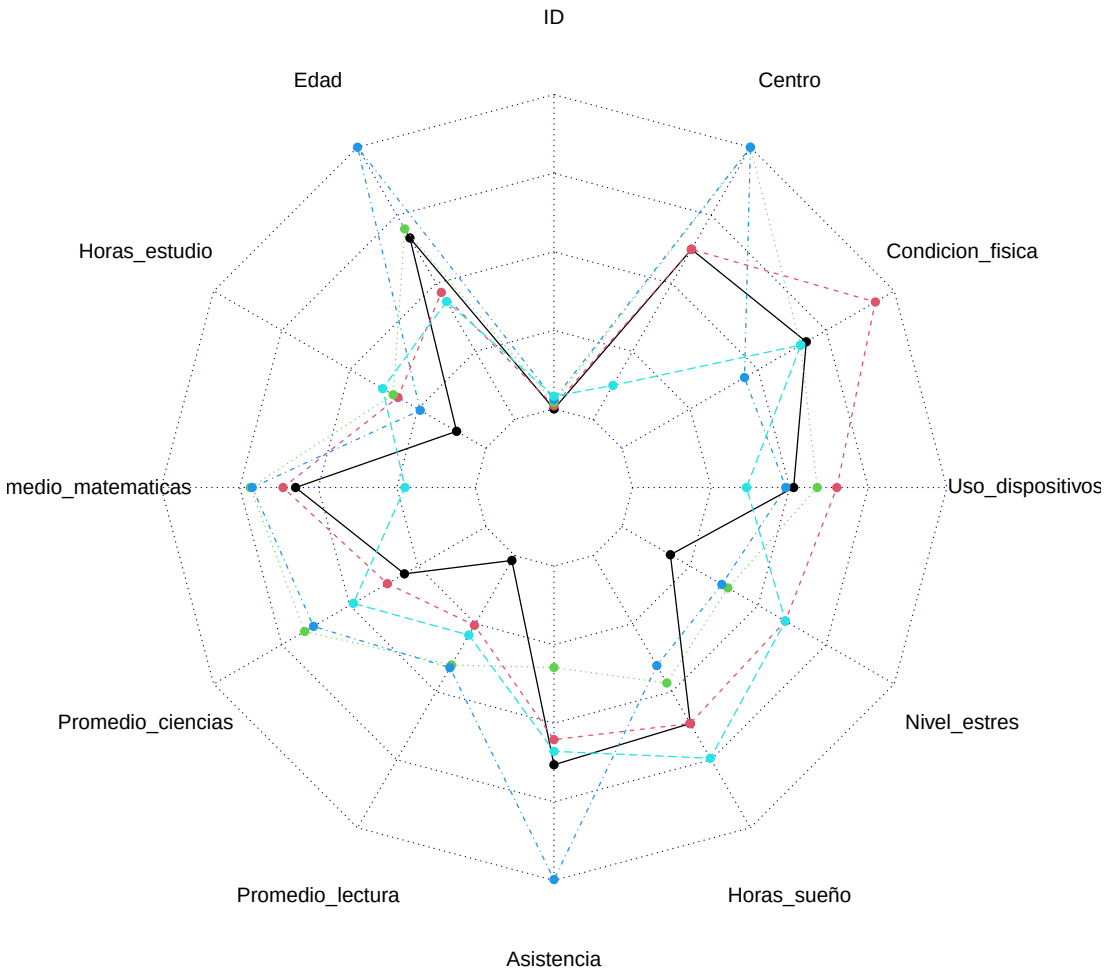
Interpretación:

- "Promedio Matemáticas" y "Horas de estudio" muestran una correlación positiva significativa (r = 0.67), lo que refuerza la importancia del tiempo dedicado al estudio en el rendimiento académico.
- "Nivel de estrés" tiene una correlación negativa con "Horas de sueño" (r = -0.45), lo que sugiere que mayores niveles de estrés reducen la calidad del descanso.
- La falta de correlación significativa entre "Uso de dispositivos" y las calificaciones podría indicar que el impacto de esta variable es más complejo y depende del tipo de uso (e.g., recreativo vs. educativo).

Comparativa entre Individuos

Se utiliza un gráfico de radar para comparar estudiantes promedio por centro.

```
library(fmsb)
normalized_data <- as.data.frame(lapply(data[, sapply(data, is.numeric)],
                                        function(x) (x - min(x)) / (max(x) - min(x))))
max_min <- rbind(apply(normalized_data, 2, max), apply(normalized_data, 2, min))
radarchart(rbind(max_min, normalized_data[1:5, ]),
           title = "Comparación de Estudiantes Promedio por Centro")
```



Interpretación:

- Los centros con estudiantes que presentan mayores promedios en "Promedio Matemáticas" tienden a destacar también en "Asistencia", lo que sugiere un vínculo entre la constancia y el rendimiento académico.
- Las diferencias en "Condición física" y "Horas de sueño" entre centros reflejan posibles variaciones en hábitos y recursos disponibles.

Conclusiones

El análisis exploratorio permitió: 1. Identificar patrones clave en la distribución de variables y detectar outliers significativos que podrían sesgar los resultados si no se analizan correctamente.

2. Resaltar correlaciones importantes entre hábitos de estudio, bienestar y rendimiento académico, proporcionando evidencia para posibles intervenciones en los centros educativos.

3. Comparar estudiantes promedio por centro, identificando diferencias en hábitos y rendimiento que podrían guiar estrategias específicas para cada institución.

Este análisis proporciona una base sólida para diseñar estrategias enfocadas en mejorar el rendimiento académico y los hábitos de los estudiantes, destacando la importancia de un enfoque integral que considere tanto factores internos como externos al ámbito académico.