# RESULTADOS

## ANÁLISIS EXPLORATORIO DE DATOS

Este estudio parte de un conjunto de 209 respuestas, conseguidas a través del cuestionario online.
Para empezar, se realizó un análisis exploratorio de los datos. Se eliminaron las columnas que no iban a ser necesarias para dicho estudio ("fecha", "consentimiento informado", "email", "comentarios" y "tipo de modificaciones"). Esta última fue eliminada porque se decidió que la mejor forma de estudiar dicha variable era sumar el número de notificaciones que se habían marcado (creando una nueva columna), en lugar de estudiar notificación por notificación. Este razonamiento parte de la base de que, a mayor número de notificaciones, mayor es el estímulo recibido por el móvil, aumentando las ganas de consultar las redes sociales.
Acto seguido, se modificaron las columnas para que estas fueran variables categóricas ordinales o dicotómicas, compuestas por números. Este paso facilitaría el análisis de datos posterior. En la **Tabla 1** se muestra un ejemplo de cómo quedó la base de datos:

**Tabla 1:** Muestra del formato de los datos una vez efectuados los cambios anteriormente comentados

| | Age | Gender | Min_attention | Stress | Happiness | Social_relationship | Sleep_hours | Meditation_time | Reading_time | Mind_wandering | Sport | Alcohol_amount | Social_media_time | Mobile_reachable | num_notifications |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2 | 0 | 3 | 6 | 8 | 5 | 4 | 1 | 1 | 2 | 5 | 3 | 8 | 0 | NaN |
| **1** | 2 | 0 | 4 | 7 | 10 | 5 | 3 | 1 | 1 | 2 | 6 | 2 | 6 | 1 | 1 |
| **2** | 2 | 1 | 8 | 8 | 8 | 3 | 4 | 1 | 1 | 4 | 3 | 2 | 6 | 1 | 1 |
| **3** | 2 | 0 | 2 | 5 | 8 | 3 | 5 | 1 | 8 | 5 | 6 | 1 | 5 | 1 | 1 |
| **4** | 2 | 0 | 5 | 8 | 5 | 4 | 3 | 1 | 5 | 1 | 3 | 2 | 6 | 1 | 3 |

A continuación, se quiso investigar los valores nulos que tenía el conjunto de datos. Para ello, se creó un mapa de calor (**Figura 1**) que mostraba aquellas instancias con valores nulos. En este punto, llama mucho la atención la columna de "número de notificaciones". Volviendo a repasar el cuestionario, salta a la vista que entre las opciones de la pregunta "¿Qué tipo de notificaciones tiene activadas…?" no se contempló la opción "No tengo notificaciones". Por tanto, para esta pregunta, se asumió que el valor nulo correspondía a la opción de tener cero notificaciones. Así pues, en dicha columna se cambiaron los valores nulos por ceros. En los otros casos, se decidió que la mediana de la variable era la mejor opción para rellenar los valores nulos.

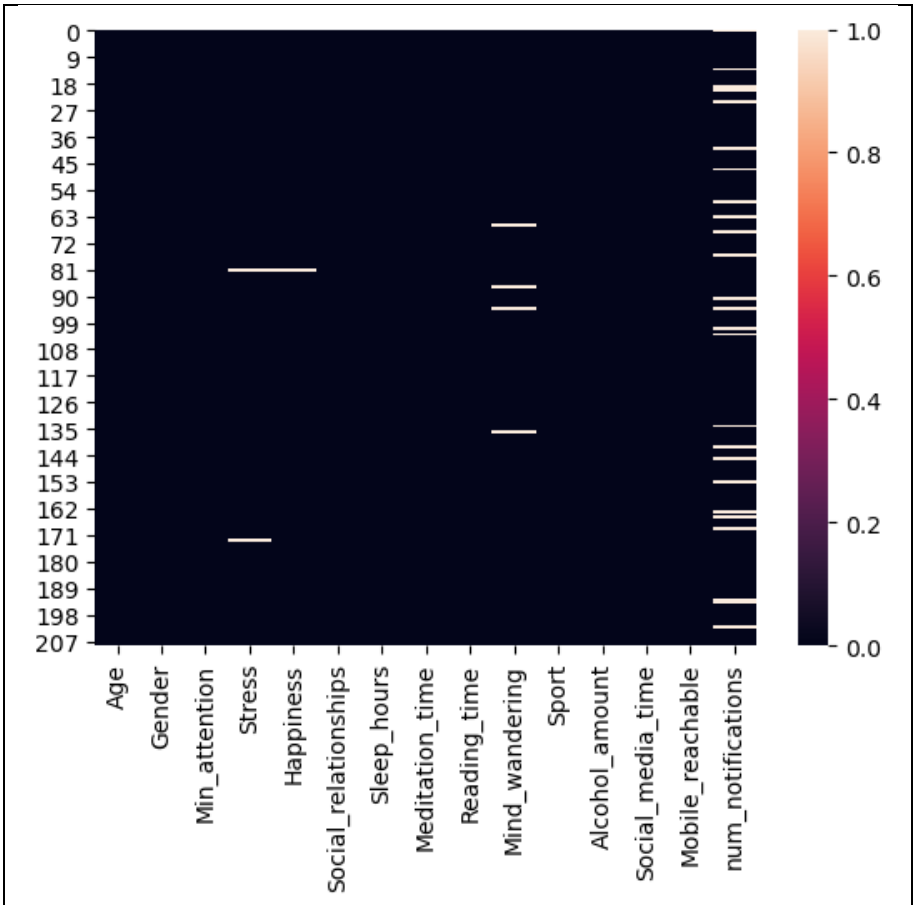



**Figura 1:** Mapa de calor de los valores nulos o faltantes de la base de datos

El siguiente paso fue empezar con el análisis estadístico. Se crearon dos tipos de marco de datos. Uno contenía las variables categóricas ordinales y el otro las binarias ("género" y "alcance móvil"). Se resumieron los datos, tal y como se puede ver en la **Tabla 2**. Con este resumen, se puede comprobar que no hay valores atípicos en las respuestas, dado que el mínimo y el máximo de todas las variables no supera el número de grupos que contiene cada una. Los resultados de este resumen podrán visualizarse mejor en el apartado de gráficas.

**Tabla 2:** Resumen del conjunto de datos numéricos

| | Age | Min_attention | Stress | Happiness | Social_relationships | Sleep_hours | Meditation_time | Reading_time | Mind_wandering | Sport | Alcohol_amount | Social_media_time | num_notifications |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 209 | 209 | 209 | 209 | 209 | 209 | 209 | 209 | 209 | 209 | 209 | 209 | 209 |
| **mean** | 3,081 | 5,062 | 6,258 | 7,541 | 3,608 | 4,426 | 1,483 | 3,201 | 2,928 | 2,794 | 1,880 | 4,775 | 1,464 |
| **std** | 1,509 | 2,542 | 1,951 | 1,506 | 1,566 | 1,007 | 1,057 | 2,276 | 1,438 | 1,445 | 1,010 | 2,630 | 0,971 |
| **min** | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| **25%** | 2 | 3 | 5 | 7 | 2 | 4 | 1 | 1 | 2 | 2 | 1 | 3 | 1 |
| **50%** | 2 | 5 | 7 | 8 | 3 | 4 | 1 | 3 | 3 | 3 | 2 | 4 | 1 |
| **75%** | 5 | 8 | 8 | 9 | 5 | 5 | 1 | 5 | 4 | 4 | 2 | 7 | 2 |
| **max** | 6 | 9 | 10 | 10 | 6 | 7 | 5 | 8 | 5 | 6 | 5 | 9 | 3 |
| **median** | 2 | 5 | 7 | 8 | 3 | 4 | 1 | 3 | 3 | 3 | 2 | 4 | 1 |
| **mode** | 2 | 9 | 7 | 8 | 2 | 5 | 1 | 1 | 3 | 3 | 1 | 3 | 1 |

Seguidamente, se utilizó el test de Shapiro para comprobar si alguna de las variables seguía una distribución normal. Sin embargo, en todas ellas se puede rechazar la hipótesis nula, lo que indica que ninguna variable sigue una distribución normal (**Tabla 3**).

**Tabla 3:** Resultados del Saphiro test para el estudio de la normalidad de las distribuciones.

| **Variable** | Shapiro_Test_Statistic | Shapiro_Test_P_Value | Shapiro_Test_Result |
|---|---|---|---|
| **Age** | 0,86760783 | 1,61e-12 | No sigue una distribución normal |
| **Min_attention** | 0,90282112 | 1,99e-10 | No sigue una distribución normal |
| **Stress** | 0,95402044 | 2,96e-6 | No sigue una distribución normal |
| **Happiness** | 0,92187113 | 4,42e-9 | No sigue una distribución normal |
| **Social_relationships** | 0,90825951 | 4,63e-10 | No sigue una distribución normal |
| **Sleep_hours** | 0,91126341 | 7,47e-10 | No sigue una distribución normal |
| **Meditation_time** | 0,50552833 | 7,15e-24 | No sigue una distribución normal |

| | | | |
|---|---|---|---|
| **Reading_time** | 0,84541905 | 1,20e-13 | No sigue una distribución normal |
| **Mind_wandering** | 0,87854695 | 6,49e-12 | No sigue una distribución normal |
| **Sport** | 0,89903748 | 1,13e-10 | No sigue una distribución normal |
| **Alcohol_amount** | 0,78423178 | 2,89e-16 | No sigue una distribución normal |
| **Social_media_time** | 0,92117071 | 3,92e-9 | No sigue una distribución normal |
| **num_notifications** | 0,81261736 | 3,97e-15 | No sigue una distribución normal |

Debido a los resultados del test anterior, la correlación de Spearman fue la utilizada entre las variables categóricas ordinales para construir la matriz de correlación. Para estudiar la correlación entre variables dicotómicas y ordinales, se utilizó de igual forma la correlación de Spearman. Finalmente, se utilizó un test chi cuadrado con corrección de Yates para estudiar la correlación entre variables binarias. Los resultados se muestran en la **Figura 2**.

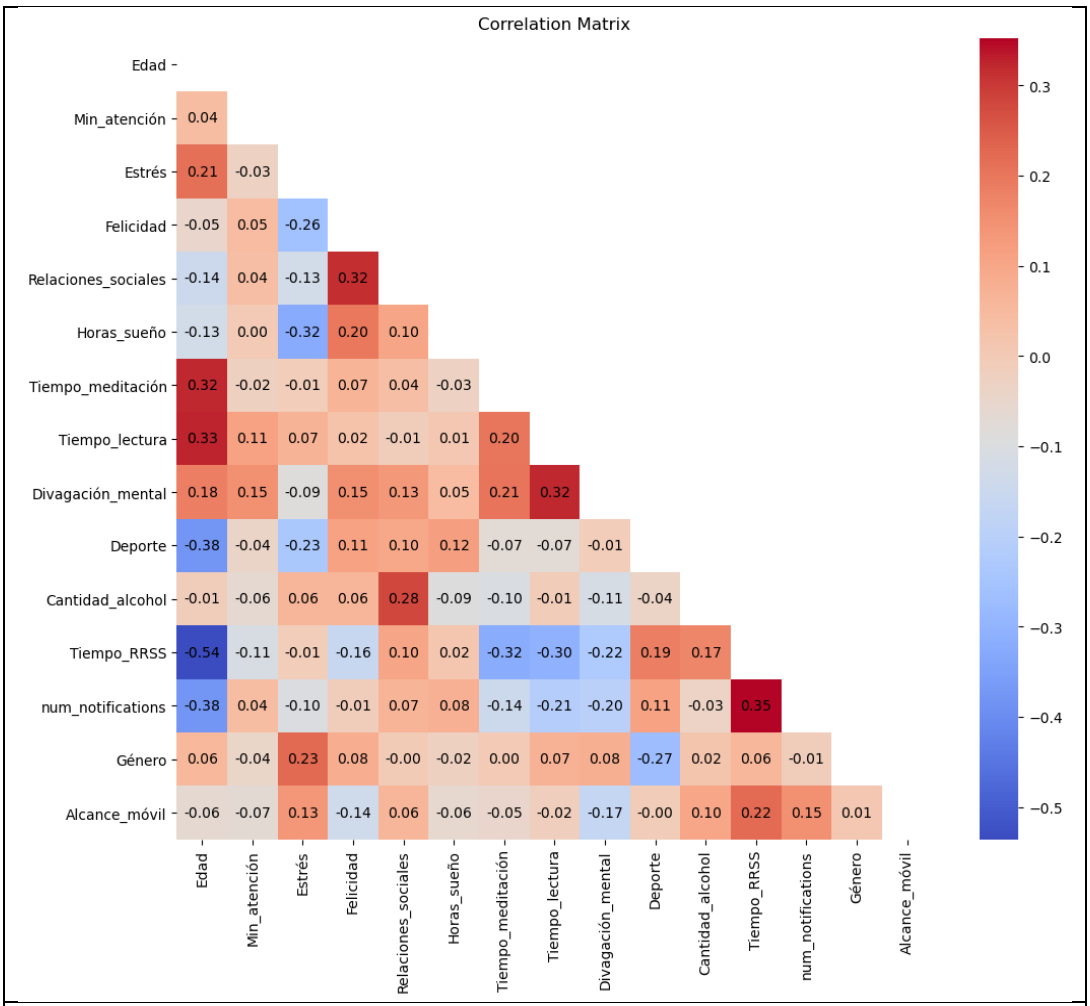


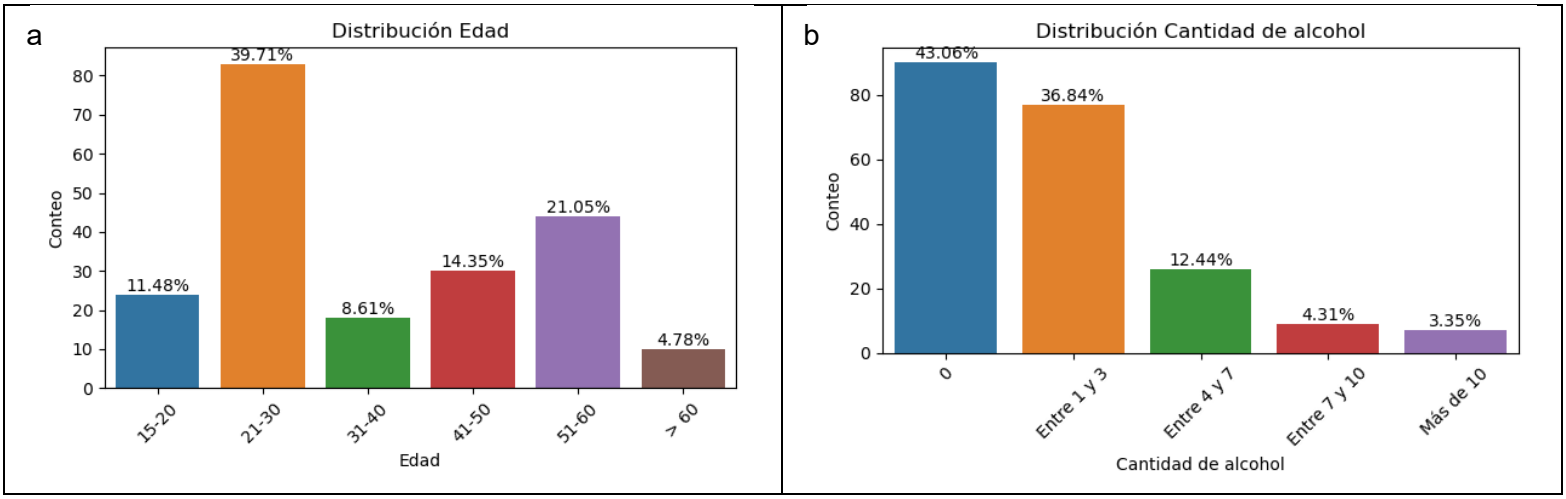

**Figura 2**: Mapa de calor de las correlaciones entre las diferentes variables. Se utilizó la correlación de Spearman entre variables ordinales, así como entre ordinales y binarias. Finalmente, la prueba de chi-cuadrado con corrección de Yates se utilizó para estudiar la correlación entre variables binarias.

En esta última figura, se puede observar cómo no hay ninguna correlación lo suficientemente fuerte (>0.8) para poder eliminar alguna de las 2 variables correlacionadas. Destaca la correlación negativa entre edad y tiempo en redes sociales (-0.54), así como edad y número de notificaciones (-0.38), es decir, a mayor edad, menor uso de las redes sociales y número de notificaciones en el móvil. Otra que también destaca es la asociación negativa entre edad y horas de deporte (-0.38), indicando que hay una tendencia a reducir las horas de actividad a medida que nos hacemos mayores.

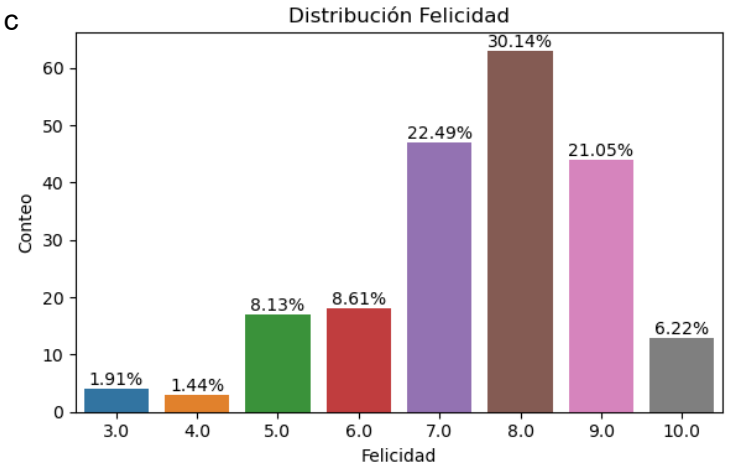
En cuanto a correlaciones con la variable de atención (objetivo de este trabajo), se puede comprobar que no hay ninguna mayor de |0.11|. Esto indica que los datos parecen no mostrar asociaciones claras con la atención. Este resultado no está en concordancia con muchos de los estudios que sí encuentran relaciones más fuertes entre las variables presentadas en este proyecto y la atención (como las presentadas en el documento adjunto de este proyecto). Este hecho puede deberse al gran sesgo que tienen los datos de este proyecto, en especial los de la atención.
Esto se debe a que los datos de atención parten de una visión subjetiva de las personas que rellenaron el cuestionario, y no de un test de atención. Por tanto, cada persona podría haber asumido dicha pregunta de forma diferente, dando lugar a datos erróneos y sin ningún tipo de reproducibilidad. Esta limitación era conocida antes de empezar con el cuestionario, pero, de haberlo hecho de manera correcta (con un test de atención ANT, por ejemplo), no se hubieran conseguido tantas respuestas debido al aumento del tiempo y la dificultad para rellenar el cuestionario. Y dado que este trabajo no es más que un proyecto final de un curso de análisis de datos, se prefirió un mayor número de respuestas a respuestas con mayor reproducibilidad.

Sin embargo, otras asociaciones, como las destacadas, sí que tienen sentido y están en concordancia con otros estudios, por lo que los datos, a pesar de tener sesgos, dejan entrever asociaciones que sí que se observan en investigaciones científicas de mayor nivel (como el estrés y sueño, con una correlación negativa de -0.32 o relaciones sociales y felicidad, con una correlación positiva de 0.32).

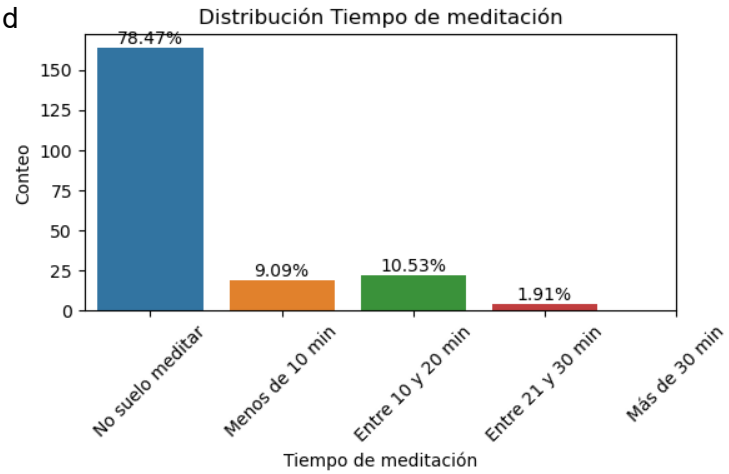
A continuación, se crearon los gráficos para poder ver las distribuciones de las variables y mostrar algunas de las correlaciones destacadas en la matriz de correlación.
Para visualizar las distribuciones (**Figura 3**), se utilizaron las gráficas de “countplot” de la librería “Seaborn”.



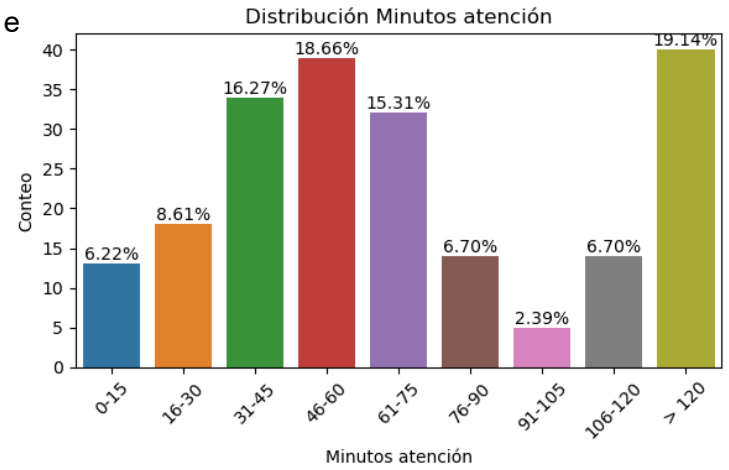
c


Distribución Felicidad
Conteo
Felicidad
1.91%
1.44%
8.13%
8.61%
22.49%
30.14%
21.05%
6.22%
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0

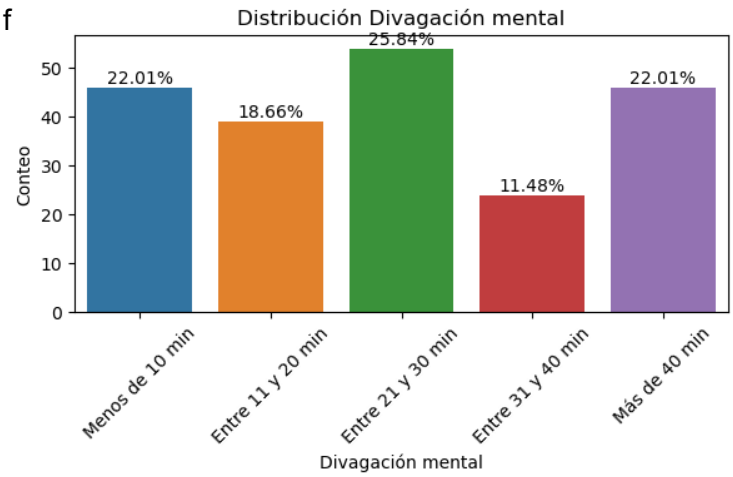
d


Distribución Tiempo de meditación
Conteo
Tiempo de meditación
78.47%
9.09%
10.53%
1.91%
No suelo meditar
Menos de 10 min
Entre 10 y 20 min
Entre 21 y 30 min
Más de 30 min

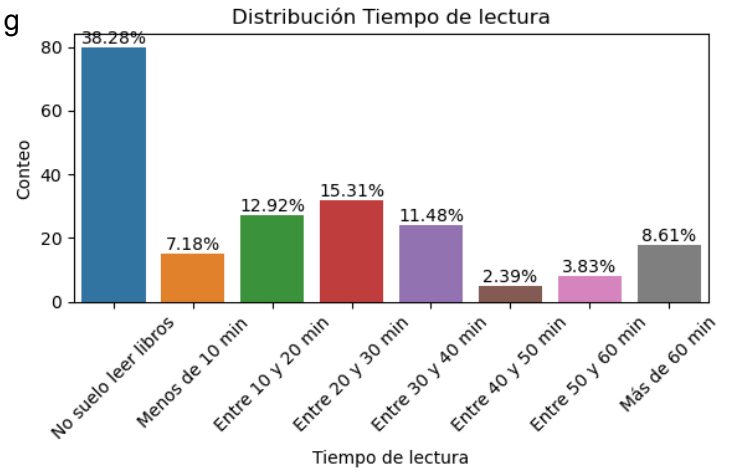
e


Distribución Minutos atención
Conteo
Minutos atención
6.22%
8.61%
16.27%
18.66%
15.31%
6.70%
2.39%
6.70%
19.14%
0-15
16-30
31-45
46-60
61-75
76-90
91-105
106-120
> 120

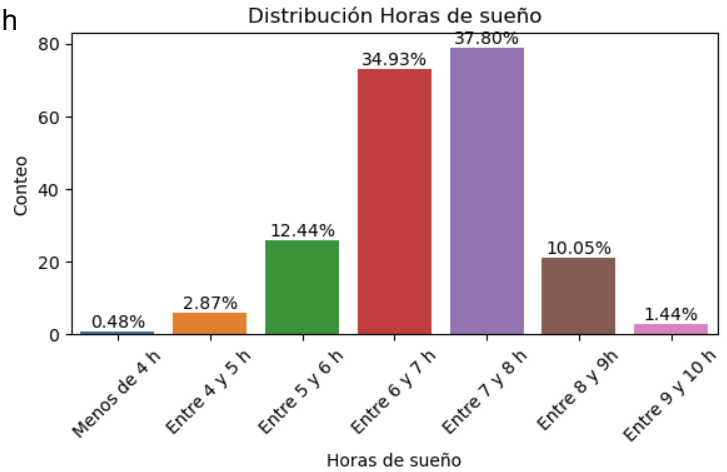
f


Distribución Divagación mental
Conteo
Divagación mental
22.01%
18.66%
25.84%
11.48%
22.01%
Menos de 10 min
Entre 11 y 20 min
Entre 21 y 30 min
Entre 31 y 40 min
Más de 40 min

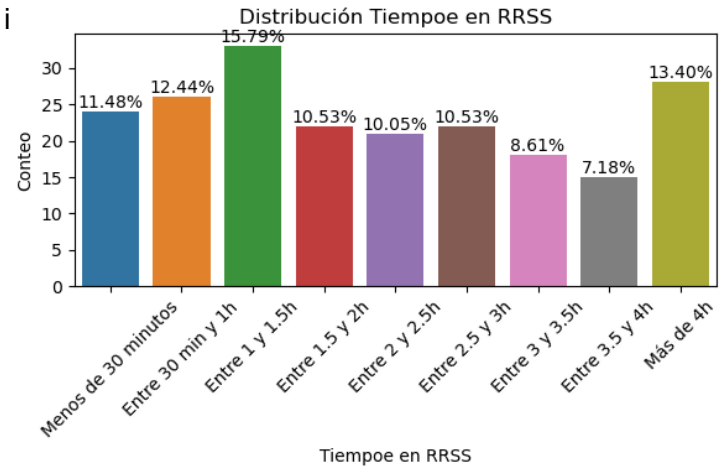
g


Distribución Tiempo de lectura
Conteo
Tiempo de lectura
38.28%
7.18%
12.92%
15.31%
11.48%
2.39%
3.83%
8.61%
No suelo leer libros
Menos de 10 min
Entre 10 y 20 min
Entre 20 y 30 min
Entre 30 y 40 min
Entre 40 y 50 min
Entre 50 y 60 min
Más de 60 min

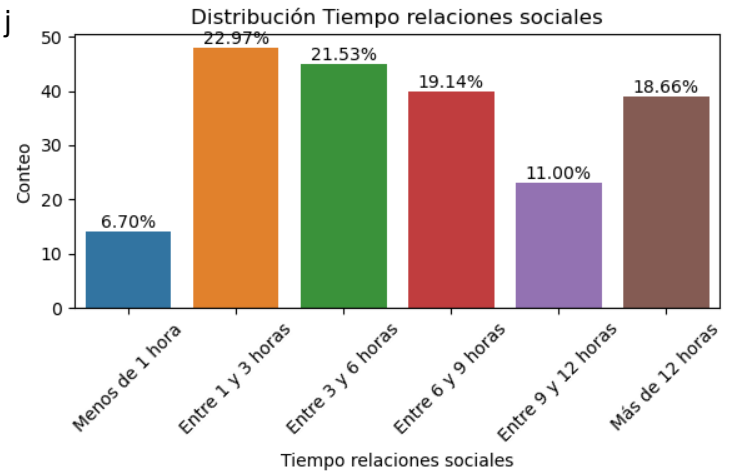
h


Distribución Horas de sueño
Conteo
Horas de sueño
0.48%
2.87%
12.44%
34.93%
37.80%
10.05%
1.44%
Menos de 4 h
Entre 4 y 5 h
Entre 5 y 6 h
Entre 6 y 7 h
Entre 7 y 8 h
Entre 8 y 9h
Entre 9 y 10 h

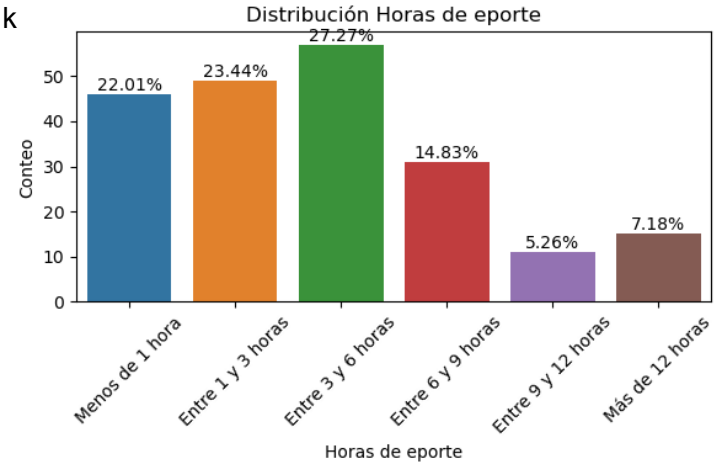
i


Distribución Tiempoe en RRSS
Conteo
11.48%
12.44%
15.79%
10.53%
10.05%
10.53%
8.61%
7.18%
13.40%
Menos de 30 minutos
Entre 30 min y 1h
Entre 1 y 1.5h
Entre 1.5 y 2h
Entre 2 y 2.5h
Entre 2.5 y 3h
Entre 3 y 3.5h
Entre 3.5 y 4h
Más de 4h
Tiempoe en RRSS

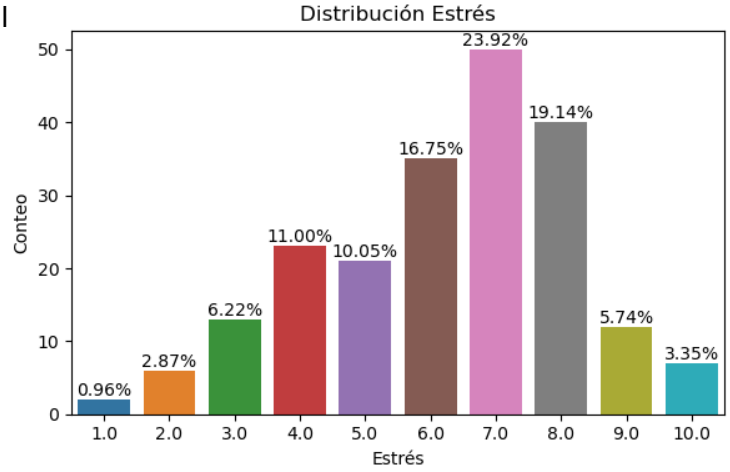
j


Distribución Tiempo relaciones sociales
Conteo
6.70%
22.97%
21.53%
19.14%
11.00%
18.66%
Menos de 1 hora
Entre 1 y 3 horas
Entre 3 y 6 horas
Entre 6 y 9 horas
Entre 9 y 12 horas
Más de 12 horas
Tiempo relaciones sociales

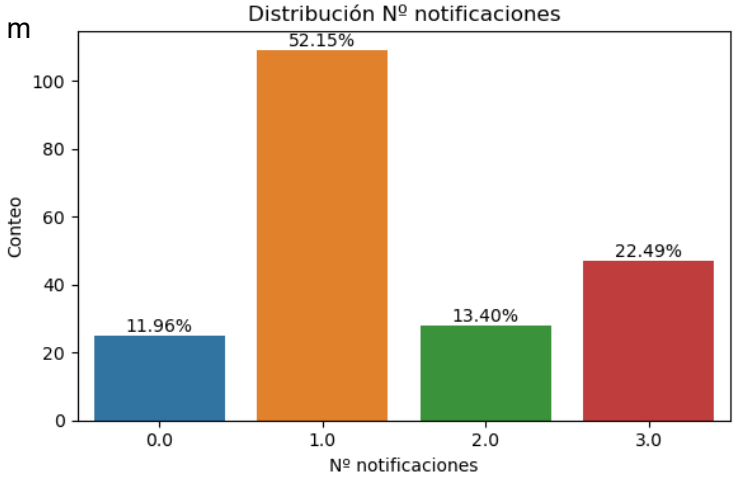
k


Distribución Horas de eporte
Conteo
22.01%
23.44%
27.27%
14.83%
5.26%
7.18%
Menos de 1 hora
Entre 1 y 3 horas
Entre 3 y 6 horas
Entre 6 y 9 horas
Entre 9 y 12 horas
Más de 12 horas
Horas de eporte

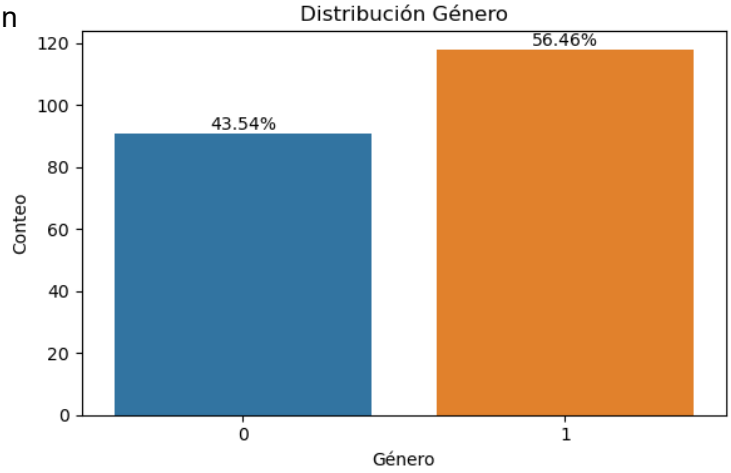
l


Distribución Estrés
Conteo
0.96%
2.87%
6.22%
11.00%
10.05%
16.75%
23.92%
19.14%
5.74%
3.35%
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
Estrés

m


Distribución Nº notificaciones
Conteo
11.96%
52.15%
13.40%
22.49%
0.0
1.0
2.0
3.0
Nº notificaciones

n


Distribución Género
Conteo
43.54%
56.46%
0
1
Género

o

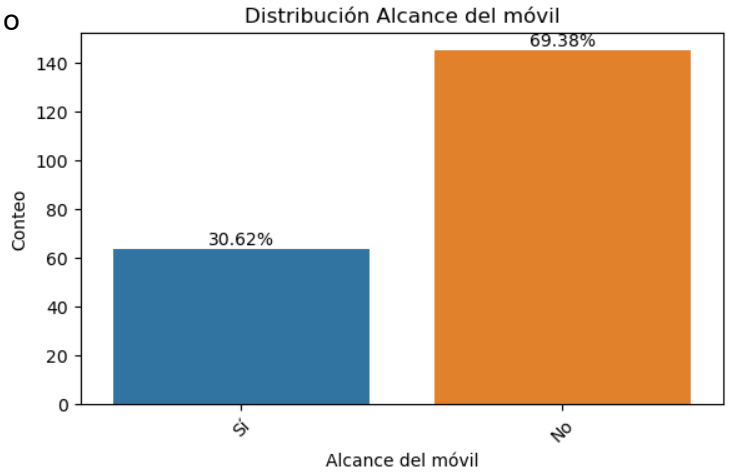



**Figura 3:** Distribución de las variables del proyecto: a) Edad, b) Cantidad de alcohol a la semana, c) Felicidad, d) Meditación, e) Minutos de concentración, f) Divagación mental, g) Lectura, h) Horas de sueño, i) Tiempo en redes sociales, j) Tiempo relaciones sociales, k) Horas de deporte, l) Estrés, m) número de notificaciones, n) género, o) Alcance del móvil

De esta gráfica, se podría destacar la rara distribución de la variable atención, siendo la última opción (>120 minutos de atención) la moda de la misma. Esto puede tener dos explicaciones: mala interpretación de la pregunta (debido a que esta no era lo suficientemente clara) o a un sesgo de respuesta social (como se explica en el apartado de Limitaciones).
Otra que también está comentado en dicho apartado es la distribución de la variable 'Edad', mostrando una falta de homogeneidad de las diferentes edades. Esto seguramente se deba al tipo de distribución del cuestionario (redes sociales) y la autoselección de los participantes.

Para las gráficas a pares (**Figura 4**), se escogieron las gráficas de violines de Seaborn, dado que en ellas también se representa la forma de la distribución, la simetría de las misma y contiene mayor cantidad de información que las cajas de bigotes. La **Figura 4** muestra dos de las asociaciones más significativas encontradas en la matriz de correlación.

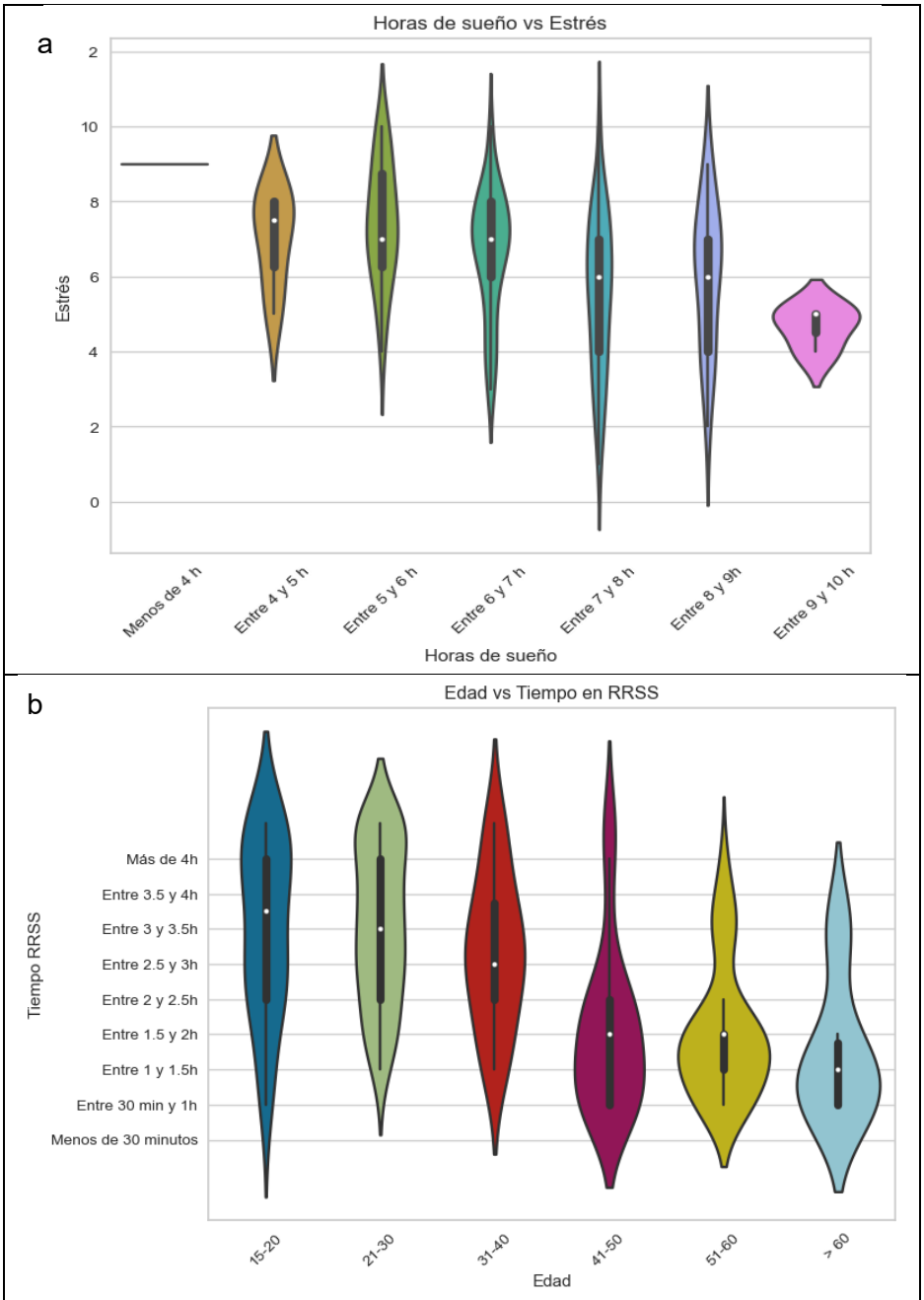



**Figura 4:** Visualización de las la distribución de a) horas de sueño vs estrés y b) edad vs tiempo en RRSS

# ANÁLISIS DE COMPONENTES PRINCIPALES

El PCA es una técnica matemática utilizada en el análisis de datos para simplificar y comprender conjuntos de datos complejos. Para este proyecto, la finalidad de realizar un PCA sobre los datos era ayudar a la visualización y comprensión de la variabilidad que tiene la base de datos. En la **Figura 5**, donde se muestran los PCA, así como la varianza explicada por cada uno y la acumulada, se comprueba que se necesitan hasta 8 PCA para explicar el 90% de la varianza. Este hecho muestra la gran variabilidad que contienen los datos.

Otra razón podría ser que el contenido importante de los datos se distribuya a lo largo de muchas dimensiones, conteniendo relaciones complejas difíciles de capturar en la reducción de

dimensionalidad. Por último, este hecho también podría darse por una alta cantidad de ruido en los datos, explicación que podría ser la más probable para este conjunto, tal y como se ha comentado anteriormente dado la calidad de los mismos.

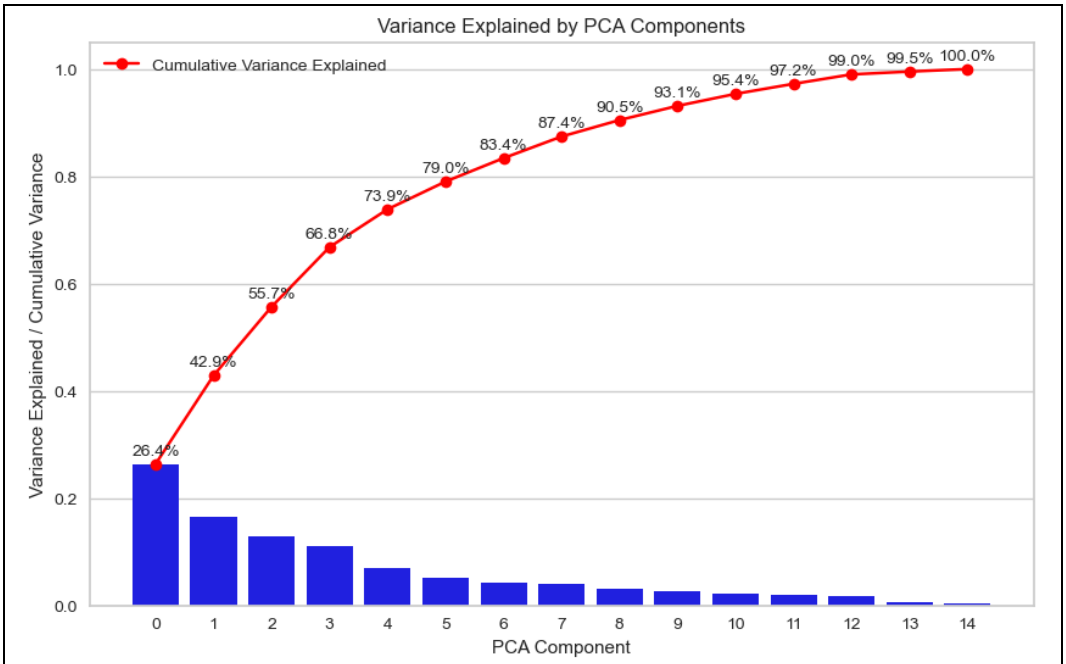



**Figura 5:** Análisis de PCA sobre el conjunto de datos. En azul, la varianza explicada por cada PCA. Se marca también la varianza acumulada con la línea roja.

# MODELOS DE APRENDIZAJE AUTOMÁTICO

Aunque la idea del aprendizaje automático no estaba propuesta en su inicio, a medida que hacía este trabajo surgió la siguiente pregunta: ¿Se puede predecir el nivel de atención a partir de todas las demás variables de este trabajo?

Esta pregunta dio paso a este apartado. Debido a que en la exploración de datos ya se podía entrever que los datos de la atención no eran buenos (no había ninguna correlación entre atención y otras variables que sí han sido identificadas por diversos grupos científicos), la idea inicial fue hacer una exploración rápida de diferentes modelos de aprendizaje automático (AA). Dicha exploración se hizo con LazyClassifier, una librería de Python que crea diferentes modelos de predicción de clase (o de regresión, según se desee) de manera automática. Esta librería es muy útil para realizar una primera exploración de modelos AA en la base de datos deseada (siempre que esta no sea muy grande), obteniendo el resultado de las métricas de diferentes modelos con una sola línea de código.

Como era de esperar, los resultados de los diferentes modelos de AA tienen una precisión extremadamente baja (<0.21), como se puede ver en la **Tabla 4**. Estos malos resultados impiden resolver la pregunta de partida y se decidió que no merecía la pena invertir más tiempo y esfuerzo en refinar modelos los modelos de aprendizaje automático, ya que una pequeña mejora no cambiaría mucho el resultado final.

**Tabla 4**: Resultados de la librería LazyClassifier sobre el conjunto de datos. Se muestra la precisión, la precisión balanceada, el valor F1 y el tiempo tomado para construir el modelo.

| Modelo | Precisión | Precisión balanceada | Valor F1 | Tiempo (s) |
|---|---|---|---|---|
| **RandomForestClassifier** | 0,222 | 0,191 | 0,193 | 0,309 |
| **QuadraticDiscriminantAnalysis** | 0,238 | 0,187 | 0,181 | 0,028 |
| **LogisticRegression** | 0,206 | 0,173 | 0,208 | 0,038 |
| **BaggingClassifier** | 0,206 | 0,172 | 0,197 | 0,068 |
| **LinearDiscriminantAnalysis** | 0,206 | 0,170 | 0,206 | 0,027 |
| **KNeighborsClassifier** | 0,190 | 0,167 | 0,168 | 0,030 |
| **NearestCentroid** | 0,190 | 0,166 | 0,200 | 0,027 |
| **PassiveAggressiveClassifier** | 0,175 | 0,165 | 0,135 | 0,030 |
| **SVC** | 0,206 | 0,163 | 0,122 | 0,034 |
| **CalibratedClassifierCV** | 0,190 | 0,154 | 0,105 | 0,181 |
| **ExtraTreesClassifier** | 0,175 | 0,152 | 0,151 | 0,273 |
| **DecisionTreeClassifier** | 0,143 | 0,149 | 0,129 | 0,025 |
| **DummyClassifier** | 0,143 | 0,125 | 0,036 | 0,023 |
| **ExtraTreeClassifier** | 0,111 | 0,125 | 0,114 | 0,024 |
| **RidgeClassifier** | 0,159 | 0,120 | 0,148 | 0,025 |
| **LinearSVC** | 0,159 | 0,120 | 0,149 | 0,080 |
| **RidgeClassifierCV** | 0,159 | 0,120 | 0,148 | 0,034 |
| **LGBMClassifier** | 0,127 | 0,112 | 0,105 | 0,337 |
| **AdaBoostClassifier** | 0,127 | 0,111 | 0,034 | 0,174 |
| **BernoulliNB** | 0,127 | 0,104 | 0,132 | 0,024 |
| **LabelSpreading** | 0,095 | 0,102 | 0,092 | 0,027 |
| **LabelPropagation** | 0,095 | 0,102 | 0,092 | 0,027 |
| **GaussianNB** | 0,079 | 0,094 | 0,085 | 0,026 |
| **Perceptron** | 0,111 | 0,085 | 0,073 | 0,032 |
| **SGDClassifier** | 0,079 | 0,063 | 0,073 | 0,041 |

# MAPAS AUTOORGANIZATIVOS

## Introducción

En el ámbito de las redes neuronales, los Mapas Autoorganizativos (SOM, por sus siglas en inglés) destacan como redes no supervisadas y autoorganizativas diseñadas para simplificar patrones complejos en espacios de entrada de alta dimensión. Estas redes resultan valiosas al analizar grandes conjuntos de datos para identificar grupos con características similares, siendo ideales para explorar relaciones entre varios factores que influyen en la atención.

## Principios de los SOM

Los SOM operan bajo el principio del aprendizaje no supervisado, centrándose en agrupar patrones de datos mediante la inspección de similitudes entre las entradas. A diferencia de las redes supervisadas, los SOM no requieren clasificaciones predefinidas, lo que les permite revelar estructuras naturales dentro de los datos. La agrupación cumple múltiples propósitos, desde identificar valores atípicos y corregir errores hasta revelar estructuras grupales emergentes para un análisis más perspicaz.

La base del análisis SOM es la construcción de una transformación de la información de entrada multidimensional en un formato bidimensional (2D). El análisis SOM define un mapa de características mediante el mapeo de un espacio de datos de entrada multidimensional en una matriz de nodos 2D.

## Arquitectura y algoritmo de los SOM

Los Mapas Autoorganizativos constituyen una red neuronal que adopta la forma de rejilla, típicamente bidimensional, compuesta por nodos o neuronas. Esta arquitectura única permite a la red aprender y establecer relaciones topográficas entre los vectores de entrada y los vectores asociados a las neuronas a lo largo del mapa.

Cada nodo en la rejilla ocupa una posición fija y está intrínsecamente vinculado a un vector prototípico de igual dimensión que los datos de entrada. Para ilustrarlo, si estamos trabajando con datos que incluyen atributos como "sueño", "edad", "minutos de atención" y "horas de deporte", cada nodo en la rejilla contendrá valores asociados a cada uno de estos atributos, es decir, un vector de longitud cuatro.

Al entrenar al algoritmo, cada nodo recibe la información de entrada (un vector con tantas dimensiones como variables haya, en nuestro caso, cuatro). El vector de entrada (instancia) se compara con los vectores de referencia de los nodos, y la ubicación en la que el vector de entrada y el vector de referencia son más similares se activa junto con su vecindad (algoritmo de competición), es decir, se asigna el dato de entrada dimensional a un nodo concreto, representado por dos coordenadas y un vector de referencia dimensional en el espacio de datos de entrada. Por tanto, los datos de entrada dimensionales se asignan a un nodo concreto, representado por dos coordenadas y un vector de referencia dimensional en el espacio de datos de entrada.

Durante el aprendizaje, los nodos que se encuentran geográficamente cerca de la instancia dada (hasta una cierta distancia geométrica que puede ser definida por el radio de vecindad) se activan entre sí para aprender algo del mismo vector de datos de entrada. Esto es, se modificarán levemente los vectores prototipo de los nodos como resultado del aprendizaje, teniendo un efecto mayor en el nodo ganador (el más similar al dato de entrada) y un efecto cada vez menor a medida que nos alejamos del mismo.
Por tanto, el aprendizaje es un proceso plástico y la precisión estadística final del análisis SOM depende del número de vectores de datos de entrada introducidos en la red.

Tras el entrenamiento del SOM, cada nodo alberga un subconjunto (que puede ser vacío) de instancias de los datos de entrada. Cada instancia, por su parte, está vinculada a un nodo específico en el mapa SOM. Estos nodos son disjuntos entre sí, lo que significa que un nodo puede representar varias instancias de los datos de entrada.

La propiedad distintiva de un SOM radica en la preservación de las propiedades topológicas de los datos de entrada en el mapa resultante. La proximidad entre inputs, definida en función de atributos como "sueño" o "edad", se refleja en la disposición de los nodos en la rejilla SOM. Este algoritmo de agrupación considera una estructura topológica de vecindad sobre los grupos, donde datos cercanos comparten nodos adyacentes.

En este trabajo, por ende, se pretende utilizar los mapas organizativos para analizar y revelar cualquier tipo de patrón que puedan tener los diferentes factores estudiados, haciendo especial hincapié en aquellos que tienen relación con la capacidad de atención.

## Detalles del análisis

Para la construcción del SOM de este estudio, se utilizó el paquete 'Kohonen' de R, creado para síntesis y visualización de mapas autoorganizativos.
Se empezó con una optimización de los hiperparámetros (tamaño de rejilla, tasa de aprendizaje y radio de la vecindad), con la finalidad de que el SOM final sea lo más reproducible y fiable posible. Para ello, se construyeron diferentes SOM y se analizaron las medidas de calidad con el paquete 'somQuality' de R. Estas son:

- El **error de cuantización** mide la distancia promedio al cuadrado entre los puntos de datos y los prototipos del mapa, buscando minimizar esta métrica para lograr una mejor representación de los datos en el mapa.

- El **porcentaje de varianza explicada** es similar a otros métodos de agrupación y representa la proporción de la varianza total explicada por el agrupamiento. Un valor más alto indica una mayor capacidad del mapa para explicar la variabilidad en los datos.

- El **error topográfico** evalúa cuán bien se preserva la estructura topográfica de los datos en el mapa. Se calcula como la proporción de observaciones para las cuales el nodo mejor coincidente no es vecino del segundo mejor nodo coincidente en el mapa. Un valor más bajo indica una mejor representación topográfica, con 0 indicando una excelente preservación topográfica.

- El **Error Kaski-Lagus** combina aspectos del error de cuantización y el error topográfico. Se calcula como la suma de la distancia media entre puntos y sus prototipos mejor coincidentes, y de la distancia geodésica media entre los puntos y sus prototipos segundo mejor coincidentes. Un valor más bajo indica una mejor calidad del mapa en términos de cuantización y preservación topográfica.

Como es de esperar, a mayor número de nodos mejores son las métricas de calidad, dado que los datos se separan más, llegando a tener varios nodos con solo una instancia.
Los hiperparámetros fueron escogidos por las métricas de calidad mencionadas anteriormente. Se buscó maximizar el porcentaje de variabilidad explicada y minimizar el error topográfico.

Finalmente, los hiperparámetros escogidos fueron una rejilla de hexagonal de 10x10, una tasa de aprendizaje en 2 fases (la primera con un valor de 0.1 y la segunda con 0.05) y un radio de vecindad de 3.
Se escogió el algoritmo de "suma de cuadrados" para evaluar el mejor nodo (algoritmo de competición).

Posteriormente, se utilizaron las librerías de 'kmeans', 'silhouette', y 'cluster' para analizar el número de grupos que podían formarse.

Por último, las librerías 'hclust' y 'cutree' fueron usadas para dividir el SOM en el número de grupos que había resultado del análisis anterior y poder visualizarlo en el mapa.

## Resultados

Después de realizar varias pruebas y ver la distribución de los datos a lo largo del SOM, se descartaron las variables “Alcohol”, “Alcance_móvil” y “Género”. Esto se debe a que no se veía ningún tipo de patrón como se puede observar más adelante con otras variables. Los puntos aparecían distribuidos a lo largo del mapa sin una distribución concreta, por lo que no aportan información relevante a este estudio.

Las métricas de calidad para el SOM construido fueron: rejilla de 10x10 de tipo hexagonal, 500 presentaciones de los datos al mapa, tasa de aprendizaje inicial de 0.1 y posterior de 0.05, un radio de vecindad de 3 y como función de distancia la suma de cuadrados.

En la **Figura 6** se muestran los “SOM changes”, en la que se representa cómo varía la distancia media a la unidad más cercana. En ella, se puede comprobar como este valor decrece con el paso de las iteraciones, con una bajada drástica alrededor de las 175 iteraciones y una posterior convergencia después de la misma hasta estabilizarse en unas 0.025 unidades de distancia media a la unidad más cercana. Este valor parece constante después de probar con distintos hiperparámetros y variando las veces que se presenta la información al algoritmo (500 y 1000).
Esta gráfica, por tanto, sugiere una adaptación progresiva y mejora en la capacidad del SOM para organizar y representar de manera efectiva la información, indicando un proceso de aprendizaje exitoso y convergencia hacia una representación óptima de los datos en el espacio SOM. Esto se debe a que la disminución de la “distancia media a la unidad más cercana” indica una mejora en la representación de los datos y un ajuste a patrones y relaciones subyacentes.

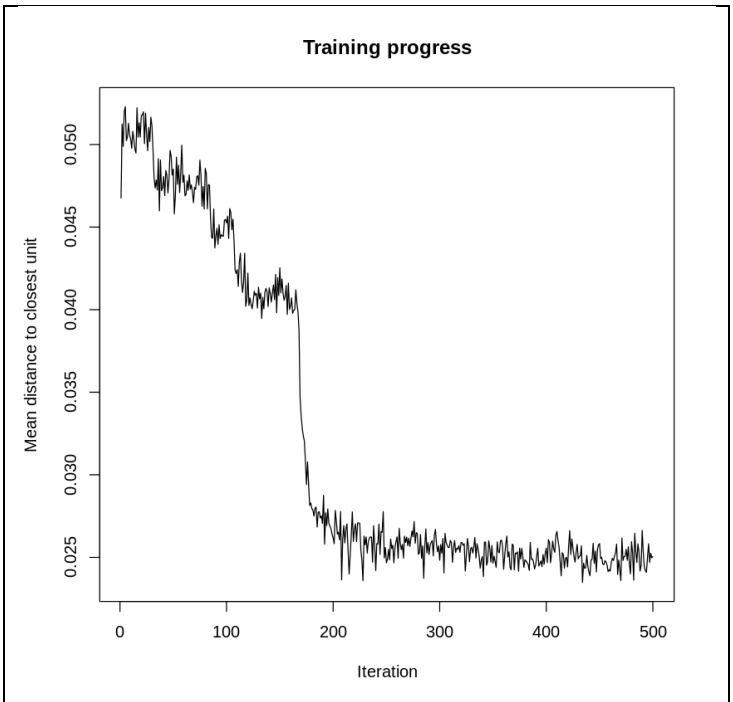



**Figura 6:** Cambios en la distancia media a la unidad más cercana a medida que se entrenaba el SOM

Posteriormente se analizó el número de instancias que contenía cada nodo. Como se puede ver en la **Figura 7**, una gran parte de los nodos contienen sólo una o dos instancias. Esto era de esperar debido

al alto número de nodos en la rejilla (necesarios para maximizar las métricas de calidad). Solo se observa un nodo totalmente vacío.

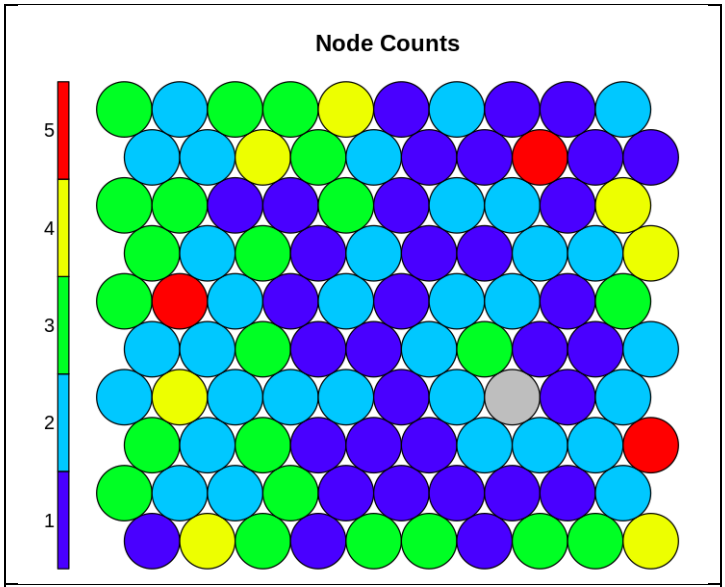



**Figura 7:** Conteo de las instancias que hay en cada nodo de la rejilla. A la izquierda se muestra la leyenda de colores

En la **Figura 8** se pueden apreciar los códigos (vectores prototípicos) de cada neurona del mapa. Estos códigos son esenciales porque representan las características centrales aprendidas por el SOM durante el proceso de entrenamiento. Sin embargo, este gráfico es muy complicado de interpretar sin otros que lo complementen, debido a la abundante información que proporciona.
Para abordar esta complejidad y obtener una comprensión más profunda, se llevó a cabo un enfoque más detallado y específico para cada variable individual (**Figura 9**), así como un estudio de agrupación final para ver si dichas neuronas se podrían agrupar mostrando patrones específicos.

En esta **Figura 9**, por ejemplo, sí que podemos observar ciertas tendencias (como las mencionadas en la matriz de correlación). Por ejemplo, si vemos la distribución de edad (**9.a**) y de tiempo en RRSS (**9.k**) parecen inversas. De igual forma pasa con estrés (**9.f**) y sueño (**9.i**) o estrés (**9.f**) y felicidad (**9.c**). Donde se ven zonas de más estrés (color naranja), en las otras distribuciones se denotan colores más morados. Esto, como decía, ya se había observado en la matriz de correlación, con la diferencia que este mapa autoorganizativo no solo tiene en cuenta las asociaciones entre 2 variables (como la matriz de correlación), sino que las integra todas a la vez para crear un mapa como el mostrado en la **Figura 9**.
La variable atención (**9.b**), no muestra ninguna distribución concreta, sino que se pueden ver varias zonas con alta atención (color naranja) y otras con baja atención (zonas moradas). Curiosamente, esas zonas de mayor atención, especialmente en personas mayores (parte derecha), se sobrepone con las zonas de lectura (**9.h**) y meditación (**9.d**).
En cuanto a los grupos más jóvenes (parte izquierda), la atención se sobrepone con buenas relaciones sociales, un estrés moderado y un uso limitado de las redes sociales.
Por tanto, y aunque los datos no sean del todo fiables, se podría plantear la cuestión de si la magnitud con la que diversos factores afectan a nuestra capacidad de atención va cambiando a medida que nos hacemos mayores, o si estos resultados son un reflejo de cómo tendemos a cambiar ciertos hábitos a

lo largo de nuestra vida. Por ejemplo, cambio de momentos de relaciones sociales por momentos tranquilos de lectura o meditación.

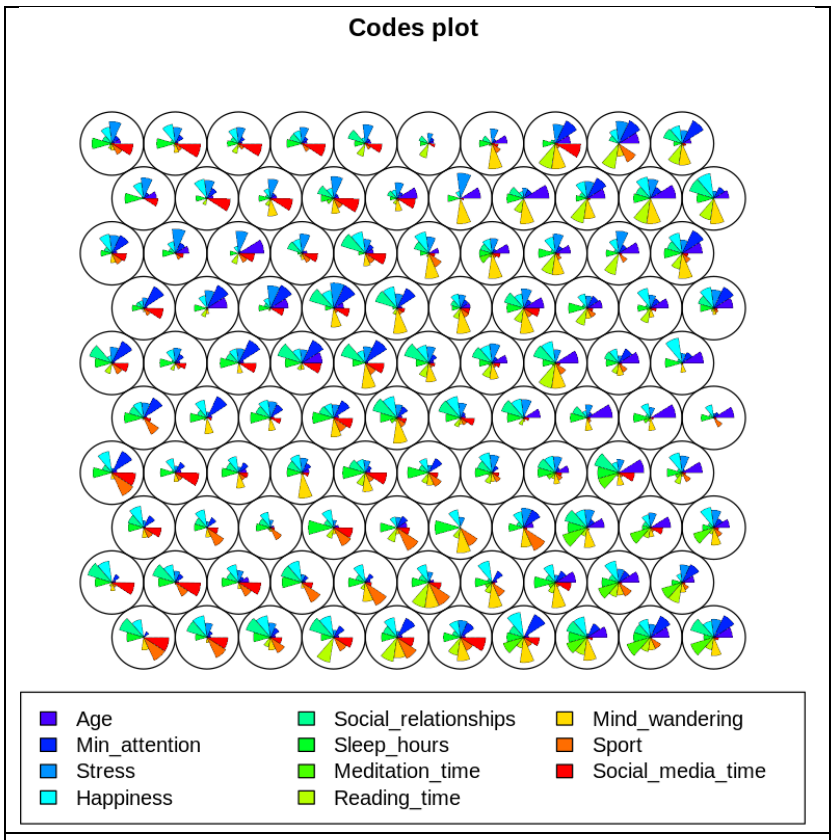
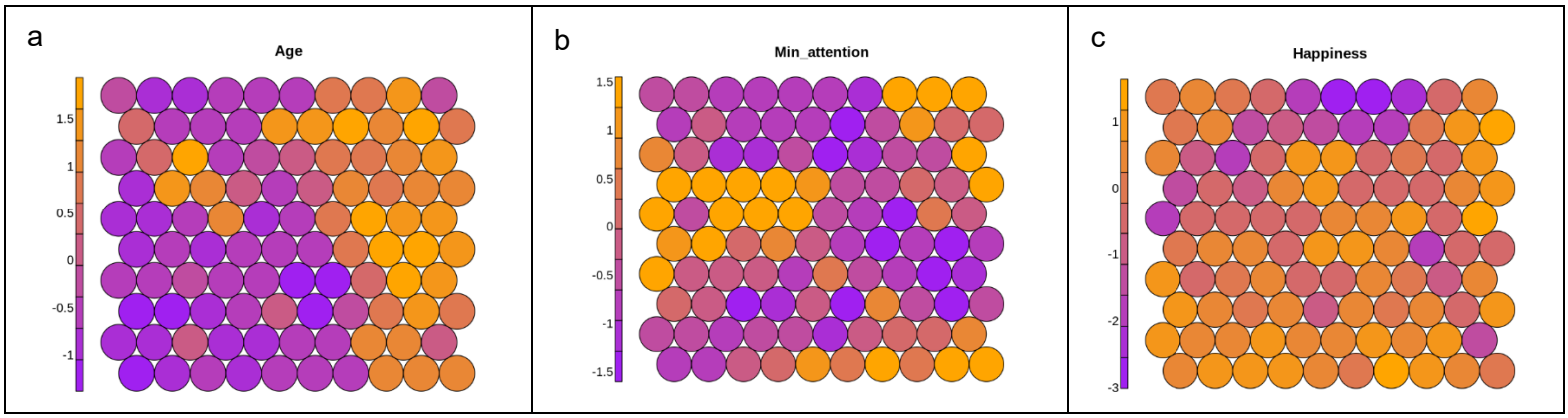



**Figura 8:** Representación de los vectores prototípicos de cada nodo. Abajo se muestra la leyenda de colores que corresponde a cada variable.



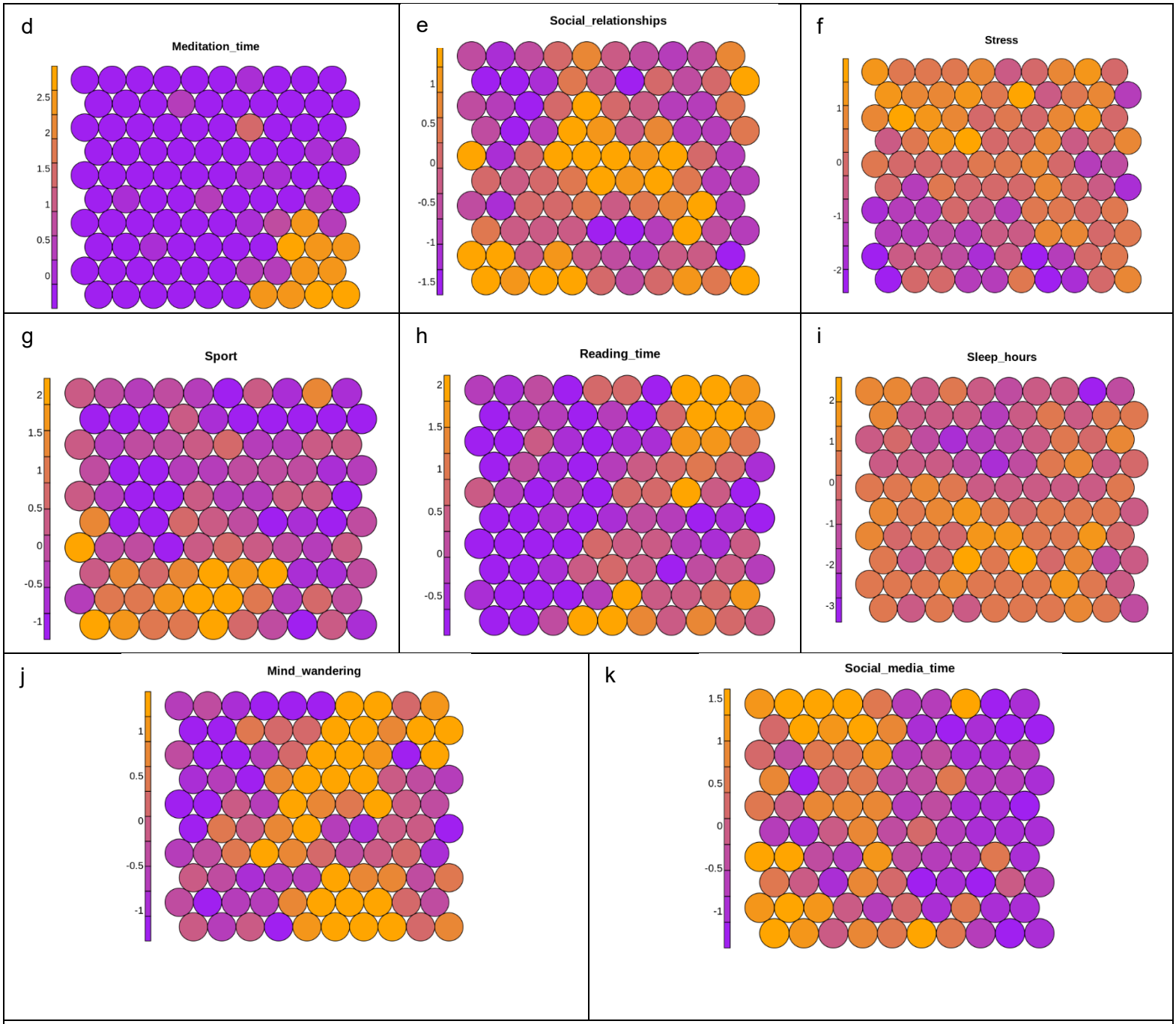



**Figura 9**: Distribución de las variables en el mapa SOM. La escala de colores, de morado a naranja, indica la magnitud de las diferentes clases. Debido a que los datos tuvieron que ser escalados para llevar a cabo el estudio del SOM, las escalas difieren las vistas en figuras anteriores. Por ejemplo, en la variable edad (a), el color morado corresponde a gente joven (números más pequeños), mientras que cuanto más naranja se vuelve el color, mayor edad (mayores es el número al escalar los datos). Se representa la distribución en el SOM de a) edad, b) minutos atención, c) felicidad, d) meditación, e) tiempo de relaciones sociales, f) estrés, g) deporte, h) tiempo de lectura, i) horas de sueño, j) divagación mental y k) tiempo en redes sociales

Para finalizar este estudio, se planteó buscar el número óptimo de grupos que contienen los datos mediante “Kmeans”, otro algoritmo de aprendizaje automático no supervisado. Los resultados, mostrados en la **Figura 10**, revelaron que los datos se ajustaban bien a una distribución con 4 grupos. Dichos grupos se visualizaron en el mapa de los vectores prototípicos del SOM, mostrado en la **Figura 11**.

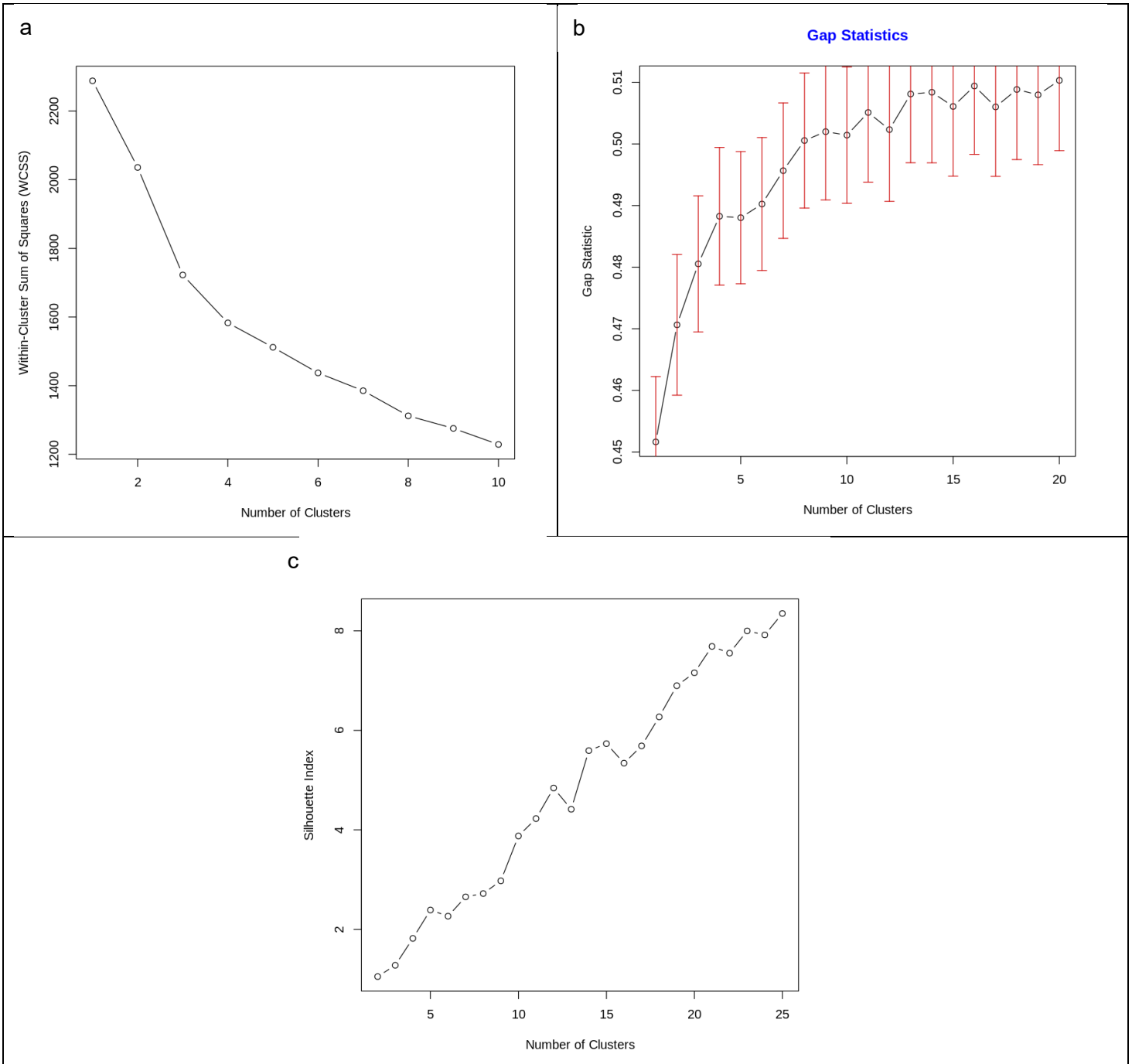



**Figura 10:** Resultado de la agrupación del conjunto de datos mediante diferentes algoritmos. a) kmeans, b) gap statistics, c) silhouette index

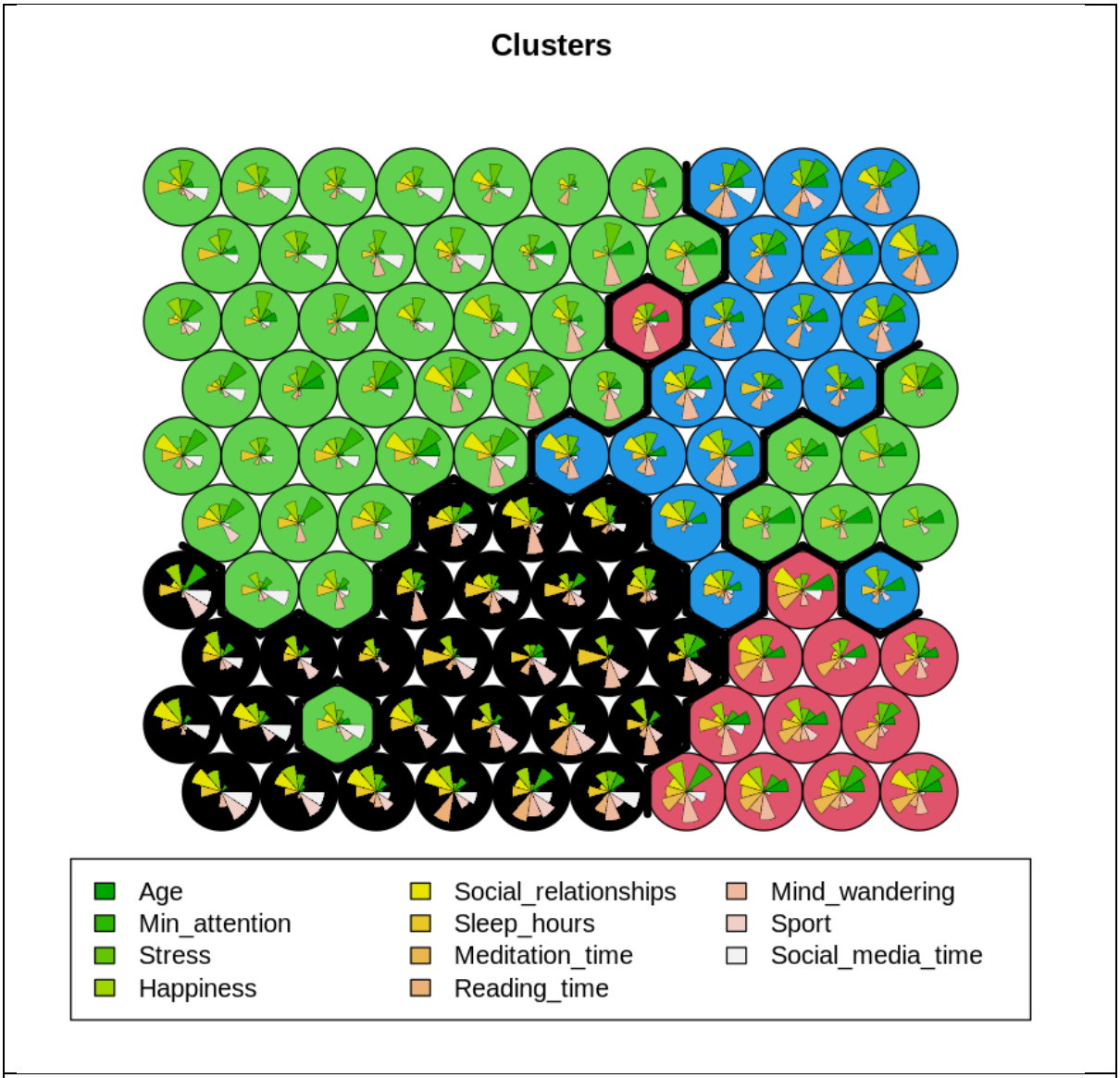



**Figura 11:** Distribución de los grupos el mapa de los vectores prototípicos del SOM.

Sin embargo, se comprobó que, a pesar de que la estructura del SOM es robusta con diferentes hiperparámetros, la agrupación no lo es, siendo muy dependiente y sensible de los mismos. Por tanto, esta agrupación muestra tener poca robustez y ser poco fiable, exponiendo que los resultados de este paso del proyecto no pueden ser considerados generalizables de manera confiable, impidiendo hacer un estudio más exhaustivo de las variables que caracterizan a cada grupo.

# CONCLUSIÓN

A partir de los resultados mostrados, los datos recolectados en este trabajo no muestran las correlaciones entre diferentes factores que afectan a la atención (como el sueño, la actividad física…) y esta misma que sí han sido demostradas en otras investigaciones de mayor nivel. Esto podría deberse, principalmente, a la calidad de los datos, siendo estos subjetivos y poco reproducibles, lo que se puede traducir en mucho ruido.

Este hecho puede ser la causa principal que explica la gran variabilidad que podemos ver en los datos (como demostró el análisis de PCA) y que ningún modelo de aprendizaje automático construido con

LazyClassifier tenga métricas aceptables. Dicha magnitud de variabilidad también afectó al SOM, que necesitaba un gran número de nodos para explicar la varianza de los datos (llegando a necesitar casi un nodo por instancia en algunos casos). Por último, también podría explicar la poca robustez y reproducibilidad de las agrupaciones creadas a partir del SOM (no tanto así de la topografía del SOM, que sí que muestra mayor robustez), ya que una pequeña variación de los hiperparámetros del SOM daba lugar a agrupaciones muy diferentes.

Por ende, los datos recopilados en este proyecto no han revelado patrones definidos que influyan de manera clara en la atención, ya sea de forma positiva o negativa. A pesar de ello, se han identificado tendencias actuales, destacándose la correlación negativa entre la edad y el uso de redes sociales, deporte, así como entre el sueño y el estrés. La interpretación de este resultado podría atribuirse a la simplicidad de las preguntas formuladas, muchas de las cuales no requieren de pruebas específicas, como el examen de la atención, para obtener respuestas más precisas.

No obstante, los datos sugieren ciertas tendencias que podrían estar vinculadas con nuestra capacidad de atención, dando lugar a interrogantes específicos, como se plantea alrededor de la **Figura 9**. Tras analizar la distribución de los factores que afectan a la atención en el SOM, se vislumbra una propensión a que esta se refuerce con momentos de lectura y meditación en personas de edad avanzada, mientras que, en individuos más jóvenes, la atención parece beneficiarse de momentos de interacción social, estrés moderado y un uso limitado de las redes sociales.

# LIMITACIONES

De partida, este cuestionario partía con un gran número de limitaciones. Se sabía que para que el conjunto de preguntas fuera reproducible y fiable, se debía de llevar a cabo un estudio minucioso y lleno de diferentes tipos de test (como atención, sueño, meditación…), los cuales no se iban a hacer por problemas económicos y de tiempo.
Este proyecto partía como la fase final del curso de 'Data analytics' de Google, por lo que primaba tener un gran número de respuestas antes que estas fueran totalmente reproducibles.
Así pues, las limitaciones de este trabajo fueron las siguientes:

- **Falta de medición objetiva**: El cuestionario se basó en autorreportes sin incluir pruebas objetivas para medir la atención o la calidad del sueño (entre otros test posibles). La ausencia de mediciones objetivas podría afectar la validez de los datos, ya que las respuestas pueden estar sujetas a sesgos de percepción y memoria.

- **Distribución a través de redes sociales**: La obtención de datos a través de redes sociales puede introducir sesgos significativos. La muestra obtenida fue mayoritariamente conformada por individuos jóvenes cercanos al investigador, lo que podría afectar la representatividad de la muestra y limitar la generalización de los resultados a otras poblaciones demográficas.

- **Énfasis en la cantidad de respuestas**: La estrategia de distribución del cuestionario priorizó la cantidad de respuestas sobre la calidad de las mismas. Esto podría haber contribuido a la

inclusión de respuestas apresuradas o incompletas, comprometiendo la precisión de los datos recopilados.

- **Sesgo de autoselección**: La naturaleza voluntaria de la participación puede haber resultado en un sesgo de autoselección, donde individuos con mayor interés en los temas abordados podrían haber sido más propensos a participar. Esto podría distorsionar los resultados al reflejar las experiencias y opiniones de aquellos más motivados por el tema.

- **Homogeneidad de la muestra**: Dada la proximidad a la red de contactos del investigador, la muestra exhibió una homogeneidad demográfica, limitando la diversidad y generalización de los resultados a poblaciones más amplias.

- **Falta de control experimental**: La ausencia de un diseño experimental controlado limita la capacidad de establecer relaciones causales entre los factores investigados y la atención sostenida. La presencia de variables no controladas podría afectar la validez interna del estudio.

- **Posible sesgo de respuesta social**: La autopresentación positiva o sesgo de respuesta social puede haber influido en las respuestas, especialmente al abordar temas como la meditación o prácticas de sueño, donde los participantes podrían sentir la presión de proporcionar respuestas socialmente aceptables.

Al considerar estas limitaciones, es crucial interpretar los resultados con cautela y reconocer las restricciones inherentes al diseño y la implementación del estudio. Se recomienda la realización de investigaciones futuras que aborden estas limitaciones para obtener una comprensión más completa y robusta de los factores que afectan la atención sostenida.