**INFORME SOBRE PROYECTO DE SALUD MENTAL**

****

Integrantes:

D’Amore Eugenia

Elizalde Pedro

ÍNDICE

1) TEMA DE INVESTIGACIÓN Y DEFINICIONES PARA EL ANÁLISIS

2) ANÁLISIS DE VARIABLES

3) CONCLUSIONES

4) BIBLIOGRAFÍA

1) TEMA DE INVESTIGACIÓN Y DEFINICIONES PARA EL ANÁLISIS

Para el desarrollo del presente proyecto se eligió un tema que afecta a personas de todo el mundo: la salud mental. En los últimos años, y gracias a numerosas campañas de concientización, este aspecto ha adquirido mayor visibilidad, lo que impulsó el interés en comprender mejor los factores que la afectan. Se parte de la hipótesis de que diversas variables —como la edad, las horas de sueño, las horas de trabajo, el tiempo de exposición a pantallas, el tipo de dieta y el nivel de ejercicio, entre otras— podrían influir significativamente en la condición mental de las personas.

- **Objetivos:**

El objetivo general de este análisis es identificar posibles relaciones entre diferentes variables del estilo de vida y el estado de salud mental, con un enfoque especial en el score de felicidad.

Los objetivos específicos son:

1. Determinar los valores mínimo, máximo y promedio del score de felicidad.
2. Evaluar si existe correlación (lineal o no lineal) entre las variables seleccionadas y el estado de salud mental.
3. Analizar si existe una diferencia significativa en el nivel promedio de felicidad entre personas con y sin condiciones de salud mental, mediante el uso de pruebas de hipótesis.
4. Explorar la posibilidad de segmentar a la población mediante técnicas de *clustering*, agrupando a los individuos en función de su edad, nivel de estrés, horas de sueño, tiempo de exposición a pantallas y puntaje de felicidad.

* **Orígen y naturaleza de los datos:**

Los datos utilizados para este estudio provienen de un conjunto disponible en la plataforma Kaggle, bajo el título *“Mental health and lifestyle habits analysis”*. El dataset está conformado por 3.000 observaciones y variables tales como: país, edad, sexo, nivel de ejercicio, tipo de dieta, horas de sueño, nivel de estrés, condición de salud mental, horas laborales semanales, horas de uso de pantallas por día, interacción social y *happiness score*.

Si bien se considera que la fuente es confiable, fue necesario realizar una adecuada limpieza y ordenamiento de los datos antes de proceder al análisis.

- **Variables bajo análisis:**

**● Condición mental**

Se analiza la variable que indica si una persona presenta o no una condición de salud mental, observando la distribución de casos en la muestra. Se plantea la hipótesis de que las personas con condiciones mentales reportan un menor *happiness score* promedio, en comparación con quienes no las presentan. Para verificar esto, se aplicarán pruebas de hipótesis.

Adicionalmente, se buscará construir un modelo predictivo que permita estimar la probabilidad de que una persona presente una condición mental, en base a variables como edad, horas de sueño, horas de trabajo semanales, tiempo frente a pantallas, interacción social y felicidad.

#### **● Happiness Score**

Se evaluará el puntaje de felicidad en términos generales, incluyendo su media, mínimo, máximo y distribución por país. También se explorará si existen diferencias significativas en este puntaje en función del nivel de estrés reportado (alto vs. moderado), para determinar si el estrés tiene un efecto en la percepción de felicidad.

Las respuestas a estas preguntas permitirán realizar proyecciones y relaciones de un tema tan complejo y que afecta la calidad de vida de millones de habitantes en todo el mundo, y esto podría resultar de interés para diversos actores como autoridades gubernamentales, empleadores, entes de salud y público en general.

- **Metodología:**

El análisis fue realizado en lenguaje R utilizando el entorno RStudio, herramienta que permitió procesar, visualizar y modelar los datos de manera eficiente. En primer lugar, se llevó a cabo una limpieza del dataset, eliminando filas con valores faltantes en variables clave, convirtiendo variables categóricas a factores y escalando las variables numéricas para su posterior análisis.

Durante la etapa exploratoria se utilizaron gráficos (histogramas, boxplots y barras) y estadísticas descriptivas para examinar la distribución de las variables principales.

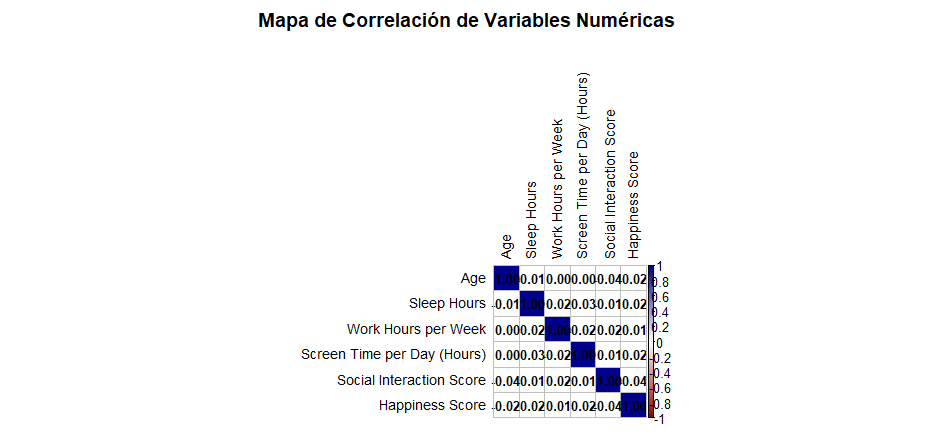
Con el fin de aplicar correctamente los tests estadísticos, se evaluó la normalidad de las variables mediante el test de Shapiro-Wilk. En función de estos resultados, como no se cumplía el supuesto de normalidad se prosiguió con pruebas no paramétricas como Kruskal-Wallis o Wilcox. Estas pruebas permitieron comparar el *happiness score* entre grupos definidos por características como la presencia de condiciones de salud mental, según el nivel de estrés.

También se analizaron correlaciones entre variables numéricas, aplicando los coeficientes de Pearson o Spearman según el caso, para identificar posibles relaciones entre las mismas.

Para explorar patrones ocultos en los datos, se aplicó un modelo de clustering K-means sobre variables estandarizadas, seleccionando el número óptimo de grupos mediante el método de la silueta. Los clusters resultantes fueron analizados e interpretados según las medias de las variables incluidas.

Además, con el objetivo de predecir la presencia o ausencia de una condición de salud mental, se construyeron modelos supervisados de clasificación: Random Forest y XGBoost. Se probaron distintos hiper parámetros y umbrales de clasificación, evaluando su desempeño mediante métricas como sensibilidad, especificidad y matriz de confusión. Si bien los resultados no fueron óptimos, permitieron explorar los límites de la capacidad predictiva del modelo a partir de las variables disponibles.

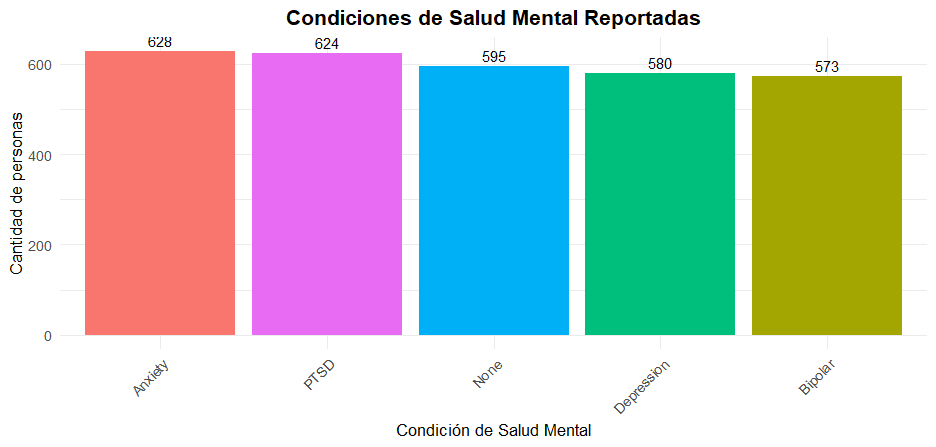
2) ANÁLISIS DE VARIABLES

****

Luego de realizar un análisis exploratorio preliminar, seleccionamos las variables numéricas del dataset —edad, horas de sueño, horas de trabajo por semana, tiempo de pantalla diario, nivel de interacción social y nivel de felicidad— para construir una matriz de correlación, visualizada mediante un heatmap.

El objetivo de este gráfico es identificar de forma visual posibles relaciones entre estas variables, como primer acercamiento a la comprensión de patrones relevantes en los datos. A su vez, se calcularon los coeficientes de correlación de Pearson y Spearman para complementar el análisis.

Sin embargo, los resultados no mostraron indicios de correlaciones fuertes ni lineales ni no lineales entre las variables seleccionadas, lo cual sugiere que, al menos en este conjunto de datos, no existen asociaciones significativas entre estas dimensiones del estilo de vida y la felicidad.

****

Se utilizó un gráfico de barras para representar la distribución de los diferentes tipos de condiciones de salud mental presentes en el conjunto de datos: ansiedad, PTSD (Trastorno de Estrés Postraumático), depresión, bipolaridad y ninguna.

El gráfico evidencia que no existe un predominio marcado de ninguna condición en particular, lo cual refleja una distribución relativamente homogénea entre las distintas categorías.

**Análisis de hipótesis**

Se realizaron dos pruebas de hipótesis con el objetivo de analizar si el nivel de felicidad difiere significativamente según dos factores:

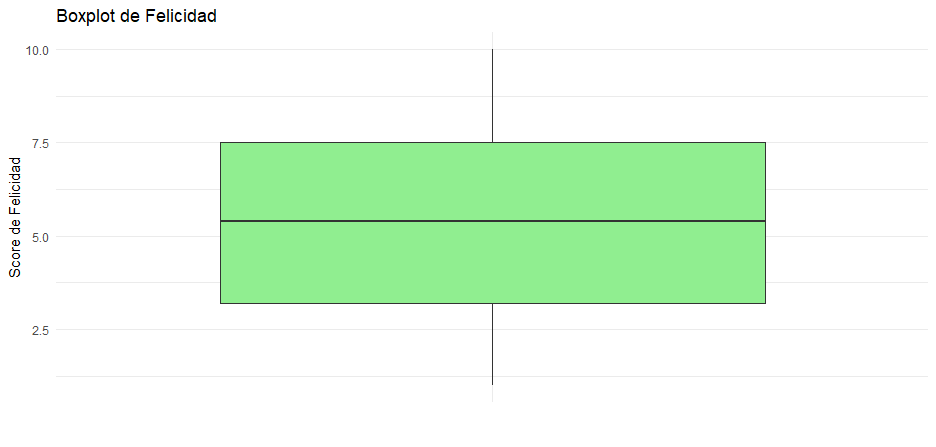
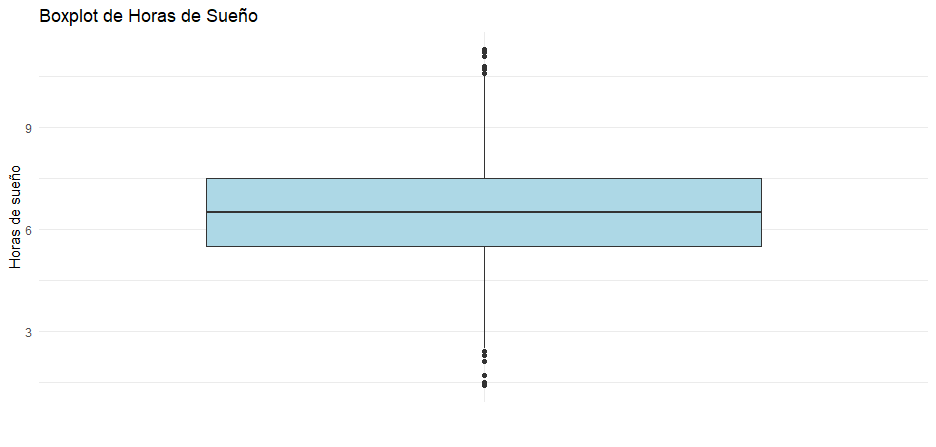
1. La presencia o ausencia de una condición de salud mental diagnosticada.
2. El nivel de estrés reportado por las personas.

Previo al análisis, se aplicó el test de Shapiro-Wilk para evaluar la normalidad de los datos. En ambos casos, los resultados indicaron que las variables no seguían una distribución normal (p < 0.05). Por esta razón, se utilizó el test de Wilcoxon, un método no paramétrico apropiado para comparar medias en este contexto.

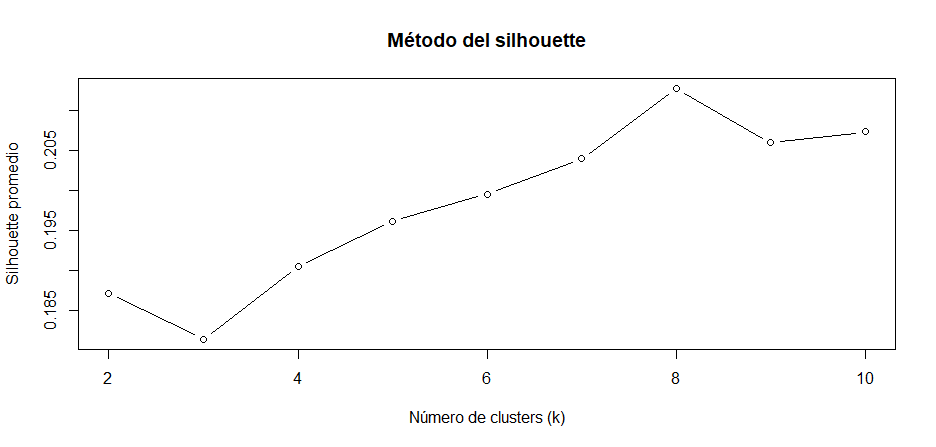
Resultados:

En la primera prueba, que comparó los niveles de felicidad entre personas con y sin diagnóstico de salud mental, no se encontró evidencia estadística suficiente para rechazar la hipótesis nula. Esto indica que, en esta muestra, no hay una diferencia significativa en la felicidad según la presencia o ausencia de una condición mental.

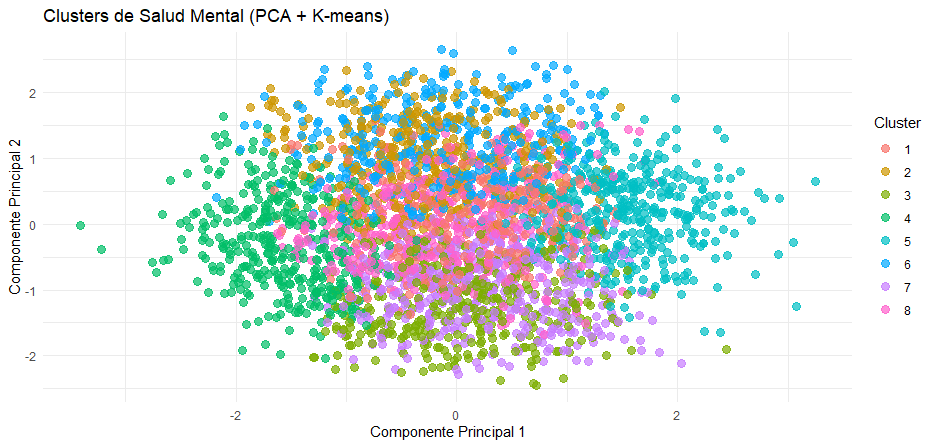
En la segunda prueba, que analizó si el nivel de felicidad varía según el nivel de estrés (alto vs. moderado), tampoco se observaron diferencias estadísticamente significativas. Por lo tanto, la felicidad no varía significativamente en función del nivel de estrés reportado.

 ****

Por otro lado, se analizaron posibles valores atípicos mediante boxplots de las variables "horas de sueño" y "score de felicidad". En el caso del sueño, se detectaron outliers tanto bajos como altos, reflejando una gran variabilidad en los hábitos de descanso. En cambio, el score de felicidad mostró una distribución más homogénea, sin outliers aparentes. Dado que los valores extremos del sueño eran relevantes, se decidió conservarlos para no perder información valiosa en los análisis posteriores.

****

Se aplicó k-means con 8 clusters, determinado a partir del método del codo y el análisis de silueta. Seguido se construyó un diagrama de dispersión con los mismos, que si bien la visualización en dos dimensiones no muestra una separación clara entre los grupos, el resumen numérico sugiere ciertas diferencias en variables como edad, sueño, felicidad y uso de pantallas.

****

A partir de este modelo se identificaron los siguientes perfiles distintos de personas según edad, horas de sueño, nivel de felicidad y tiempo de pantalla.

**Cluster 1:** Jóvenes con bajo sueño y baja felicidad, alto uso de pantallas.  
**Cluster 2:** Jóvenes con sueño moderado, baja felicidad y poco uso de pantallas.  
**Cluster 3:** Adultos mayores con buen descanso, baja felicidad y alto uso de pantallas.  
**Cluster 4:** Adultos mayores con poco sueño y baja felicidad, uso de pantalla moderado.  
**Cluster 5:** Jóvenes muy felices, con buen descanso y alto uso de pantallas.  
**Cluster 6:** Jóvenes muy felices con poco sueño y bajo uso de pantallas.  
**Cluster 7:** Adultos mayores muy felices con poco sueño y alto uso de pantallas.  
**Cluster 8:** Adultos mayores con buen descanso y alta felicidad, bajo uso de pantallas.

Si bien los grupos presentaron combinaciones interesantes de variables, los resultados no permitieron establecer conclusiones claras ni patrones consistentes entre las mismas y los niveles de felicidad. Esto sugiere que, al menos con la calidad y estructura de este dataset, no es posible definir perfiles de bienestar mental de forma concluyente, lo cual refuerza la complejidad del tema y la necesidad de contar con datos más completos y precisos para futuros análisis.

Por lo tanto, el clustering puede considerarse una herramienta exploratoria útil, pero sus resultados deben tomarse con cautela y no se deben sobreinterpretar.

**Modelos de clasificación:**

Por último, se desarrollaron modelos de clasificación utilizando algoritmos de aprendizaje automático para predecir la presencia de condiciones de salud mental en función de variables como edad, horas de sueño, nivel de estrés, interacción social, entre otras.

En primera instancia, se implementó un modelo de *Random Forest*, que si bien es robusto frente a relaciones no lineales y útil con datos mixtos, no alcanzó el nivel de desempeño esperado. Aunque el modelo fue capaz de identificar muchos casos positivos, la marcada desproporción entre clases afectó su rendimiento general. Se intentaron distintas estrategias para mejorar su rendimiento, incluyendo el balanceo de clases, ajuste de hiper parámetros y modificación del umbral de clasificación, pero los resultados seguían siendo poco satisfactorios.

|  | Predicho: No | Predicho: Sí |
| --- | --- | --- |
| Real: No | 2 (VN) | 8 (FP) |
| Real: Sí | 174 (FN) | 716 (VP) |

Ante esta situación, se optó por probar con el algoritmo *XGBoost*, conocido por su alta capacidad predictiva. Sin embargo, el modelo presentó una alta tasa de falsos negativos, es decir, no logró identificar correctamente a la mayoría de los casos positivos.

|  | Predicho: No | Predicho: Sí |
| --- | --- | --- |
| Real: No | 163 (VN) | 15 (FP) |
| Real: Sí | 652 (FN) | 69 (VP) |

Ambos modelos mostraron limitaciones importantes, posiblemente debido a la calidad y el desbalance del dataset. Aunque se realizaron múltiples pruebas y ajustes, no se logró alcanzar un desempeño predictivo satisfactorio. Este resultado invita a reflexionar sobre la necesidad de mejorar la recolección y balance de datos, así como considerar variables adicionales o enfoques complementarios en futuros análisis.

3)CONCLUSIÓN:

A pesar de aplicar múltiples técnicas estadísticas y de machine learning, los resultados obtenidos no permitieron establecer relaciones concluyentes entre las variables analizadas y la salud mental, lo que sugiere limitaciones en la calidad del dataset y refleja la complejidad del bienestar mental como fenómeno multifactorial.

4)BIBLIOGAFÍA

Dataset:[**https://www.kaggle.com/code/abhigaykwad/mental-health-and-lifestyle-habits-analysis**](https://www.kaggle.com/code/abhigaykwad/mental-health-and-lifestyle-habits-analysis)

[**https://mhanational.org/es/news/mental-health-america-releases-analysis-of-its**Noticias: **-2023-online-mental-health-screens-u-s-sees-continued-rise-of-anxiety-psychosis-and-adhd-risk/**](https://mhanational.org/es/news/mental-health-america-releases-analysis-of-its-2023-online-mental-health-screens-u-s-sees-continued-rise-of-anxiety-psychosis-and-adhd-risk/)

[**https://www.tiempoar.com.ar/ta\_article/adolescentes-y-celulares-cada-vez-mas-horas-de-uso-impacto-en-la-salud-mental-y-el-sueno/**](https://www.tiempoar.com.ar/ta_article/adolescentes-y-celulares-cada-vez-mas-horas-de-uso-impacto-en-la-salud-mental-y-el-sueno/)

[**https://www.nhlbi.nih.gov/es/health/sleep-deprivation/health-effects#:~:text=Beneficios%20en%20t%C3%A9rminos%20de%20salud,escuela%20y%20se%20sientan%20estresados**](https://www.nhlbi.nih.gov/es/health/sleep-deprivation/health-effects#:~:text=Beneficios%20en%20t%C3%A9rminos%20de%20salud,escuela%20y%20se%20sientan%20estresados)