

# Algoritmos supervisados y no supervisados aplicados a analizar la relación entre la salud mental, hábitos y felicidad.

Fisam Zavala | Maestría en Ciencia de Datos | Aprendizaje Automático

November 2025

## 1. Introducción

El bienestar y la salud mental se han convertido en temas centrales de investigación en las ciencias sociales y del comportamiento, especialmente ante el aumento de factores que afectan la calidad de vida en entornos modernos. Los avances recientes en el aprendizaje automático han permitido analizar estas variables desde un enfoque cuantitativo, facilitando la identificación de patrones y relaciones complejas entre los hábitos de vida, las emociones y la percepción de felicidad.

En este contexto, el presente estudio busca analizar cómo distintos factores del estilo de vida —tales como el sueño, el nivel de estrés, el tipo de dieta, las horas de trabajo y la interacción social— se relacionan con los niveles de bienestar subjetivo y felicidad. Para ello, se aplicaron algoritmos de aprendizaje supervisado y no supervisado con el objetivo de explorar tanto la estructura interna de los datos como su capacidad predictiva.

El enfoque metodológico combina técnicas de agrupamiento, como *K-Means* y *DBSCAN*, con modelos de predicción basados en regresión lineal y selección de características (*SelectKBest* y *RFE*), además de estrategias para evaluar el número óptimo de clústeres y los errores de estimación. Esta integración de métodos permite abordar el análisis desde una perspectiva exploratoria y explicativa, contribuyendo a una mejor comprensión de los determinantes del bienestar.

En suma, el propósito de este trabajo es identificar patrones de bienestar y evaluar la influencia de los hábitos personales sobre la salud mental y la felicidad percibida, demostrando cómo el aprendizaje automático puede constituirse en una herramienta eficaz para el análisis interdisciplinario del comportamiento humano.

## 2. Descripción de los datos

## 3. Antecedentes

Diversos estudios han abordado la relación entre los hábitos de vida, la salud mental y la felicidad, empleando enfoques cuantitativos y técnicas de aprendizaje automático para identificar patrones de bienestar.

En el estudio realizado por el *Journal of Pakistan Medical Students* (JPMS, 2021), se exploró cómo los hábitos de vida auto-reportados, tales como las horas de sueño, el tiempo frente a pantalla y los niveles de estrés, pueden predecir el grado de felicidad mediante modelos de minería de datos y aprendizaje automático. Los resultados mostraron que la calidad del sueño y la gestión del estrés son variables determinantes en los niveles de bienestar subjetivo.

Por su parte, la investigación publicada en *BMC Public Health* (2019) destacó una relación longitudinal entre el optimismo, la felicidad y los comportamientos saludables, demostrando que el bienestar emocional no solo depende de los hábitos, sino que también puede influir en su mantenimiento a lo largo del tiempo.

De manera complementaria, el artículo de *Preventive Medicine Reports* (2021) evidenció que el bienestar subjetivo predice conductas de salud a largo plazo, reforzando la hipótesis de una relación bidireccional entre bienestar psicológico y estilo de vida saludable.

Más recientemente, el estudio publicado en *Scientific Reports* (Nature Portfolio, 2025) empleó modelos gráficos para analizar la interacción entre múltiples dimensiones del bienestar, incluyendo el sueño, las relaciones sociales y la satisfacción laboral. Dicho trabajo concluye que los factores emocionales y sociales se encuentran estrechamente interconectados en la percepción global de bienestar.

Finalmente, el artículo del *International Journal of Qualitative Studies on Health and Well-Being* (2023) analizó la influencia de la alimentación en la percepción de bienestar infantil, encontrando que los hábitos alimenticios diarios están vinculados con la estabilidad emocional y la satisfacción personal.

En conjunto, estos antecedentes muestran que la felicidad y la salud mental están fuertemente asociadas con los hábitos de vida, y que las metodologías basadas en aprendizaje automático ofrecen un marco eficaz para modelar y comprender dichas relaciones complejas.

## 4. Marco Teórico

### 4.1. Métricas de desempeño

La evaluación del desempeño de los modelos predictivos es esencial para determinar su capacidad de generalización y precisión. En el caso de los modelos de regresión, las métricas se centran en medir el error entre los valores observados y los valores predichos. A continuación, se describen las principales métricas utilizadas en este trabajo: el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación ( $R^2$ ).

#### 4.1.1. Error Absoluto Medio (MAE)

El *Mean Absolute Error* (MAE) mide el promedio de las diferencias absolutas entre los valores reales ( $y_i$ ) y los valores predichos ( $\hat{y}_i$ ). Su fórmula se expresa como:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

El MAE representa la magnitud promedio del error de predicción, expresado en las mismas unidades que la variable objetivo. Es una métrica robusta ante valores atípicos y fácil de interpretar, ya que indica, en promedio, cuánto se desvía el modelo de los valores reales.

#### 4.1.2. Error Cuadrático Medio (MSE)

El *Mean Squared Error* (MSE) calcula el promedio de los errores al cuadrado entre los valores observados y los estimados. Se define como:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

El MSE penaliza con mayor peso los errores grandes debido a la elevación al cuadrado, por lo que resulta útil cuando se desea dar más importancia a desviaciones significativas. No obstante, su interpretación directa puede ser menos intuitiva al no estar en las mismas unidades de la variable dependiente.

#### 4.1.3. Raíz del Error Cuadrático Medio (RMSE)

La *Root Mean Squared Error* (RMSE) es la raíz cuadrada del MSE, y se calcula como:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

El RMSE conserva las unidades originales de la variable objetivo, lo que facilita su interpretación. Es sensible a los valores atípicos, pero proporciona una visión clara del error promedio esperado en las predicciones. En comparación con el MAE, el RMSE tiende a ser mayor cuando existen errores extremos.

#### 4.1.4. Coeficiente de Determinación ( $R^2$ )

El *Coeficiente de Determinación*, denotado como  $R^2$ , mide la proporción de la variabilidad total de los datos que es explicada por el modelo. Su expresión matemática es:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

Donde  $SS_{res}$  es la suma de los residuos al cuadrado y  $SS_{tot}$  representa la variabilidad total de los datos respecto a su media. El valor de  $R^2$  varía entre 0 y 1, siendo los valores cercanos a 1 indicativos de un buen ajuste del modelo. Sin embargo, un  $R^2$  elevado no garantiza que el modelo sea el más adecuado, por lo que debe interpretarse junto con las métricas de error.

## 4.2. Diseño de Experimentos

### 4.2.1. Selección de factores y niveles

Se consideran factores centrales por plausibilidad causal y disponibilidad en el dataset. Para variables continuas, se dicotomizan en niveles *bajo/alto* usando la mediana muestral; esto balancea tamaños de grupo y simplifica un diseño  $2^k$ .

Tabla 1: Factores y niveles propuestos (umbral seg\xfcreo de acuerdo a la mediana muestral).

Factor	Niveles	Definición operacional
Exercise Level	Bajo / Alto	Según codificación del dataset (p.ej., Low vs High).
Sleep Hours	Bajo / Alto	$\leq 6,5$ h vs $> 6,5$ h.
Stress Level	Bajo / Alto	Low vs High (excluir <i>Moderate</i> o recodificar a binario).
Diet Type	Balanceada / No balanceada	Balanced vs (Vegan/Vegetarian/Junk Food) o criterio nutricional equivalente.
Work Hours per Week	Bajo / Alto	$\leq 39$ h/sem vs $> 39$ h/sem.
Screen Time per Day (Hours)	Bajo / Alto	$\leq 5,1$ h/día vs $> 5,1$ h/día.
Social Interaction Score	Bajo / Alto	$\leq 5,5$ vs $> 5,5$ .

### 4.2.2. Estructura del diseño

Incluir los siete factores anteriores en un factorial completo implicaría  $2^7 = 128$  tratamientos. Para controlar el tamaño del experimento sin perder capacidad de *screening*, se propone un **diseño fraccional factorial** de resolución IV:

$$2^{7-3} = 16 \text{ tratamientos (Resolución IV),}$$

el cual permite estimar *efectos principales* libres de confusión con interacciones de dos factores (aunque las interacciones dobles pueden estar aliadas entre sí). Alternativamente, si se desea más poder para interacciones, puede considerarse  $2^{7-2} = 32$  corridas.

**Fase 1 (screening).** Aplicar el  $2^{7-3}$  (o  $2^{7-2}$ ) con aleatorización y, de ser posible, 1–2 réplicas técnicas para estimar error puro. Analizar con ANOVA factorial (modelo lineal con efectos principales y, opcionalmente, interacciones seleccionadas).

**Fase 2 (optimización).** Con los 2–3 factores más influyentes, pasar a una **Superficie de Respuesta** (p.ej., Diseño Central Compuesto) para capturar curvatura y encontrar combinaciones que maximizan *Happiness Score*.

#### 4.2.3. Modelo y análisis

Sea  $Y$  el *Happiness Score* y  $X_j$  los indicadores (0/1) de los niveles altos de cada factor. El modelo lineal para la fase de *screening* es:

$$Y = \beta_0 + \sum_{j=1}^k \beta_j X_j + \sum_{j < \ell} \beta_{j\ell} X_j X_\ell + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Se contrasta significancia de efectos principales e interacciones mediante ANOVA y pruebas  $F$ , verificando supuestos (normalidad de residuos, homocedasticidad e independencia). Para el dataset observacional actual, el mismo marco puede implementarse como **AN(C)OVA**, añadiendo covariables (p.ej., Age, Country) o usando *blocking* conceptual.

## 5. Metodología

### 5.1. Selección de características

En el aprendizaje supervisado, la selección de características constituye una etapa fundamental para reducir la dimensionalidad del conjunto de datos, mejorar la interpretabilidad del modelo y evitar el sobreajuste (*overfitting*). En este estudio se emplearon tres enfoques distintos: *SelectKBest – F (f\_regression)*, *SelectKBest – Mutual Information (MI)* y *Recursive Feature Elimination (RFE)* con *LinearRegression*. Cada uno aborda la relevancia de las variables desde una perspectiva distinta: estadística, informacional y de desempeño predictivo, respectivamente.

#### 5.1.1. SelectKBest – F (f\_regression)

Este método utiliza la prueba F derivada del análisis de varianza (ANOVA) para evaluar la relación lineal entre cada variable independiente  $X_i$  y la variable objetivo  $y$ . El estadístico F se define como:

$$F = \frac{MS_{\text{entre}}}{MS_{\text{dentro}}} = \frac{\text{Varianza explicada por el modelo}}{\text{Varianza residual}}$$

donde  $MS_{\text{entre}}$  representa la variabilidad media entre los grupos definidos por  $X_i$  y  $MS_{\text{dentro}}$  la variabilidad interna o error. Un valor alto de  $F$  indica que la variable  $X_i$  tiene una influencia significativa sobre  $y$ , lo que sugiere que debe ser considerada en el modelo.

#### 5.1.2. SelectKBest – Mutual Information (MI)

El método basado en información mutua cuantifica la dependencia estadística (no necesariamente lineal) entre dos variables aleatorias. La *Mutual Information* (MI) mide cuánta información de  $X_i$  reduce la incertidumbre de  $y$  y se define como:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)$$

donde  $p(x, y)$  es la distribución conjunta y  $p(x)p(y)$  el producto de las marginales. A diferencia de la prueba F, la MI detecta dependencias tanto lineales como no lineales, siendo especialmente útil cuando la relación entre las variables no sigue un patrón paramétrico.

#### 5.1.3. Recursive Feature Elimination (RFE) con LinearRegression

El método RFE selecciona características de forma iterativa eliminando las menos relevantes en función del peso que aportan al modelo. Se entrena una regresión lineal sobre todas las variables y se calculan los coeficientes  $\beta_i$  asociados a cada predictor  $X_i$ . En cada iteración se eliminan las variables con menor valor absoluto de  $\beta_i$ , repitiendo el proceso hasta conservar el número óptimo de características  $k$ .

El modelo lineal de base se define como:

$$\hat{y} = \beta_0 + \sum_{i=1}^p \beta_i X_i$$

donde  $\beta_i$  indica la contribución de la variable  $X_i$  sobre la variable dependiente  $\hat{y}$ . De esta forma, RFE busca identificar el subconjunto de variables que maximiza el desempeño predictivo del modelo lineal minimizando simultáneamente el error cuadrático medio (MSE).

## 5.2. Algoritmos no supervisados

Los algoritmos no supervisados permiten descubrir estructuras o patrones ocultos en los datos sin requerir una variable objetivo. En este trabajo se aplicaron dos técnicas de agrupamiento: *K-Means* y *DBSCAN*, complementadas con diversos criterios para determinar el número óptimo de clústeres. Estos métodos permiten segmentar observaciones con base en su similitud multivariada, identificando grupos representativos del comportamiento general de la población analizada.

### 5.2.1. K-Means Clustering

El algoritmo *K-Means* busca dividir el conjunto de datos en  $k$  grupos homogéneos de tal forma que se minimice la distancia interna entre las observaciones y el centroide de su respectivo clúster. Matemáticamente, su objetivo consiste en minimizar la suma total de las distancias cuadráticas intra-clúster:

$$J = \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

donde  $C_i$  representa el conjunto de observaciones asignadas al clúster  $i$ , y  $\mu_i$  es el centroide correspondiente, definido como el promedio de las observaciones de dicho grupo. El algoritmo sigue un proceso iterativo de dos pasos: (1) asignación de puntos al centroide más cercano y (2) actualización de los centroides hasta alcanzar convergencia o un cambio mínimo en  $J$ .

### 5.2.2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

A diferencia de K-Means, el algoritmo *DBSCAN* agrupa observaciones con base en la densidad de puntos en el espacio de características, lo que le permite identificar clústeres de forma arbitraria y detectar valores atípicos (*outliers*). Cada punto se clasifica como núcleo, frontera o ruido según dos parámetros: la distancia máxima  $\varepsilon$  y el número mínimo de puntos requeridos *MinPts*.

Formalmente, un punto  $p$  es considerado núcleo si cumple que:

$$|\{q \in D \mid \text{dist}(p, q) \leq \varepsilon\}| \geq \text{MinPts}$$

Los clústeres se forman a partir de la conectividad entre puntos núcleo y sus vecinos directos, mientras que los puntos no alcanzables por densidad se consideran ruido. Este enfoque es especialmente útil para conjuntos de datos con distribuciones irregulares o presencia significativa de anomalías.

### 5.2.3. Estrategias para determinar el número óptimo de clústeres

La determinación del número adecuado de clústeres ( $k$ ) es un aspecto crítico en los algoritmos de agrupamiento. Para ello, se aplicaron tres estrategias complementarias: el *Método del Codo*, el *Coeficiente de Silhouette* y el *Índice de Calinski-Harabasz*.

**Método del Codo** Evalúa la función de costo  $J(k)$  definida como la suma de las distancias intra-clúster para diferentes valores de  $k$ . El número óptimo de clústeres se elige donde la reducción marginal de  $J(k)$  comienza a estabilizarse, formando un “codo” en la gráfica:

$$J(k) = \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

**Coeficiente de Silhouette** Mide la cohesión y separación de los clústeres mediante la comparación entre la distancia promedio de un punto a los elementos de su propio grupo ( $a_i$ ) y la distancia promedio al clúster más cercano ( $b_i$ ):

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

El valor promedio de  $s_i$  para todos los puntos varía entre  $-1$  y  $1$ , donde valores cercanos a  $1$  indican clústeres bien definidos.

**Índice de Calinski-Harabasz** También conocido como el *Variance Ratio Criterion*, este índice evalúa la proporción entre la dispersión inter-clúster y la dispersión intra-clúster:

$$CH = \frac{\text{Tr}(B_k)/(k-1)}{\text{Tr}(W_k)/(n-k)}$$

donde  $\text{Tr}(B_k)$  es la traza de la matriz de dispersión entre clústeres y  $\text{Tr}(W_k)$  la traza de la matriz de dispersión interna. Un valor más alto de  $CH$  indica una mejor partición de los datos.

### 5.3. Algoritmos Supervisados

Los algoritmos supervisados utilizan un conjunto de datos etiquetados, donde las variables predictoras  $X$  se asocian con una variable respuesta  $y$ . El objetivo es construir una función  $f(X)$  capaz de aproximar el valor de  $y$  para nuevas observaciones, minimizando el error entre los valores observados y los predichos.

$$y_i = f(X_i) + \varepsilon_i, \quad \text{con } \varepsilon_i \sim N(0, \sigma^2)$$

A continuación se describen algunos de los algoritmos supervisados más comunes utilizados en problemas de regresión y pronóstico:

#### 5.3.1. Regresión Lineal

Busca establecer una relación lineal entre la variable dependiente  $y$  y las variables independientes. Su formulación matemática general es:

$$\hat{y} * i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Los coeficientes  $\beta_j$  se estiman minimizando la suma de los errores cuadráticos:

$$\min_{\beta} \sum_{i=1}^n (y_i - \hat{y} * i)^2 = \min_{\beta} \beta(y - X\beta)'(y - X\beta)$$

La solución analítica se obtiene mediante:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Es un modelo interpretable y constituye la base para la mayoría de los métodos de regresión.

#### 5.3.2. Regresión de Bosques Aleatorios (Random Forest)

Es un método de ensamble basado en múltiples árboles de decisión. Cada árbol  $T_b$  se entrena sobre una muestra aleatoria  $S_b$  del conjunto original. La predicción final se obtiene promediando las salidas de todos los árboles:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(X)$$

donde  $B$  representa el número de árboles en el bosque. Este método reduce la varianza del modelo y mejora la generalización.

## 6. Resultados

Una etapa fundamental en el uso de *K-Means* es definir el número adecuado de clústeres ( $k$ ), ya que este parámetro influye directamente en la estructura final del agrupamiento. Existen diversas métricas de validación interna que permiten evaluar la calidad de las particiones obtenidas. En este análisis se emplearon dos de las más utilizadas:

- **Método del Codo (Elbow Method):** evalúa la *inercia*, es decir, la suma de las distancias cuadráticas entre los puntos y su centroide correspondiente. Se calcula para distintos valores de  $k$ , y el valor óptimo se identifica en el punto donde la disminución de la inercia deja de ser significativa, formando un “codo” en la gráfica.
- **Coeficiente de Silhouette:** mide la calidad de la asignación de cada observación a su clúster, comparando la distancia promedio al resto de los puntos del mismo grupo con la distancia promedio al clúster más cercano. Valores cercanos a 1 indican una buena separación entre clústeres, mientras que valores negativos sugieren una posible asignación errónea.

Ambos métodos se aplicaron para el rango  $k = 2$  a  $10$ , encontrándose que el punto de inflexión en el método del codo y el máximo promedio del coeficiente de Silhouette coincidieron aproximadamente en  $k = 3$ , lo cual sirvió como guía para la selección inicial del número de grupos.

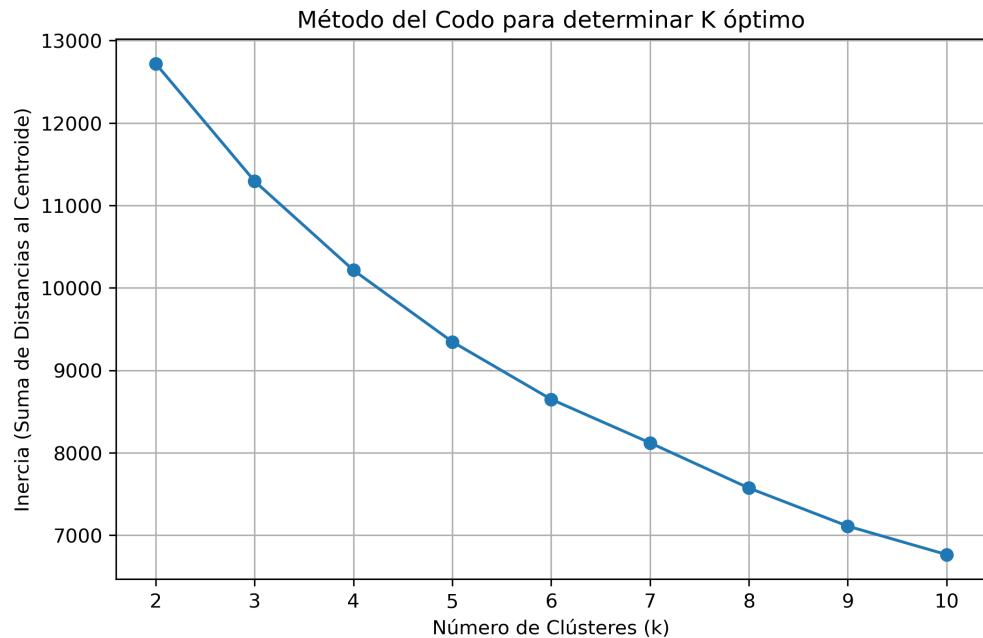


Figura 1: Método del Codo: se observa una disminución pronunciada en la inercia hasta  $k = 3$ , punto a partir del cual la reducción se estabiliza, sugiriendo que tres clústeres son apropiados.

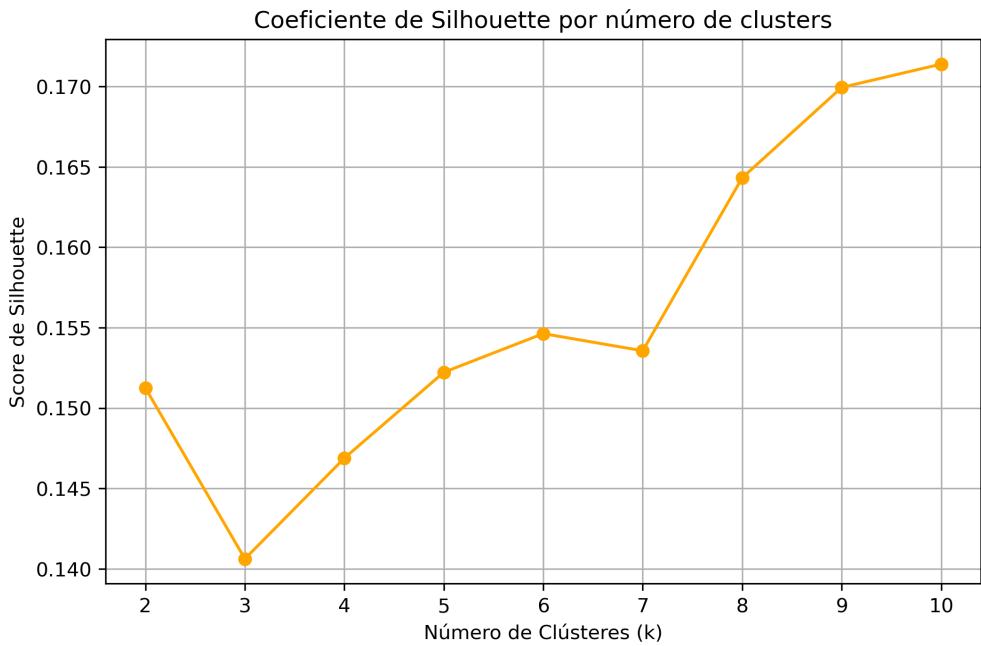


Figura 2: Coeficiente de Silhouette por número de clústeres. Los valores más altos indican una mejor cohesión dentro de los grupos y una mayor separación entre ellos.

La siguiente tabla presenta los valores promedio de las variables numéricas para cada grupo identificado por el algoritmo K-Means.

Cluster	Sleep Hours	Work Hours per Week	Screen Time per Day (Hours)	Social Interaction Score	Happiness Score
0	6.460000	38.150000	6.140000	5.970000	3.070000
1	6.410000	41.510000	3.210000	5.530000	5.350000
2	6.560000	38.490000	6.150000	4.910000	7.750000

Tabla 2: Resumen descriptivo de los clústeres (K-Means).

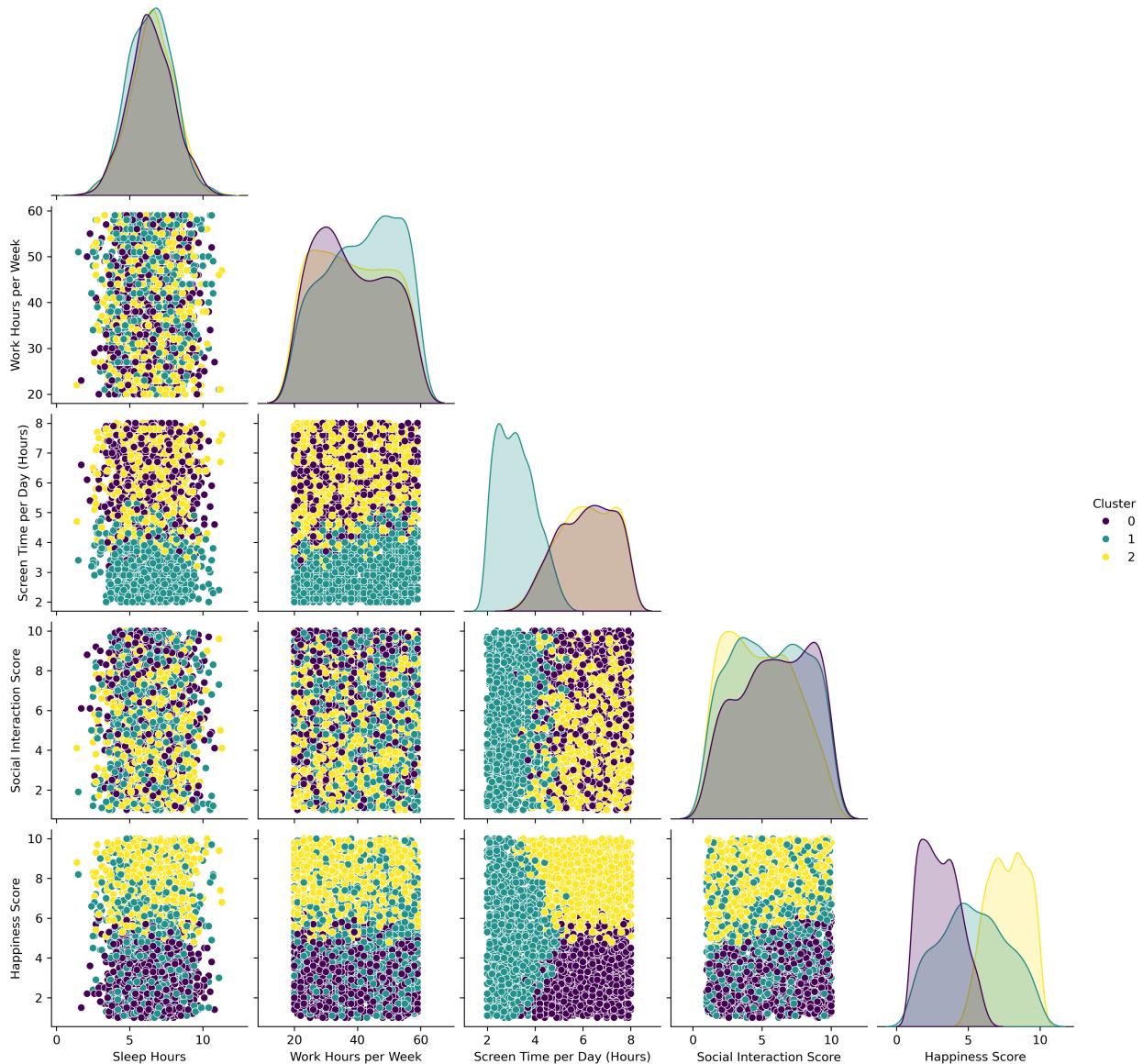


Figura 3

Se exploraron diferentes valores de  $k$  utilizando tanto el método del codo como el índice de Calinski-Harabasz, determinándose que  $k = 2$  ofrecía la mejor partición en términos de cohesión y separación.

El resultado evidenció dos grupos principales: el primero concentró individuos con mayor carga laboral y niveles moderados de felicidad, mientras que el segundo incluyó a quienes trabajan menos horas semanales y presentan niveles ligeramente superiores de bienestar. Estos hallazgos se analizarán con mayor detalle en la sección de resultados.

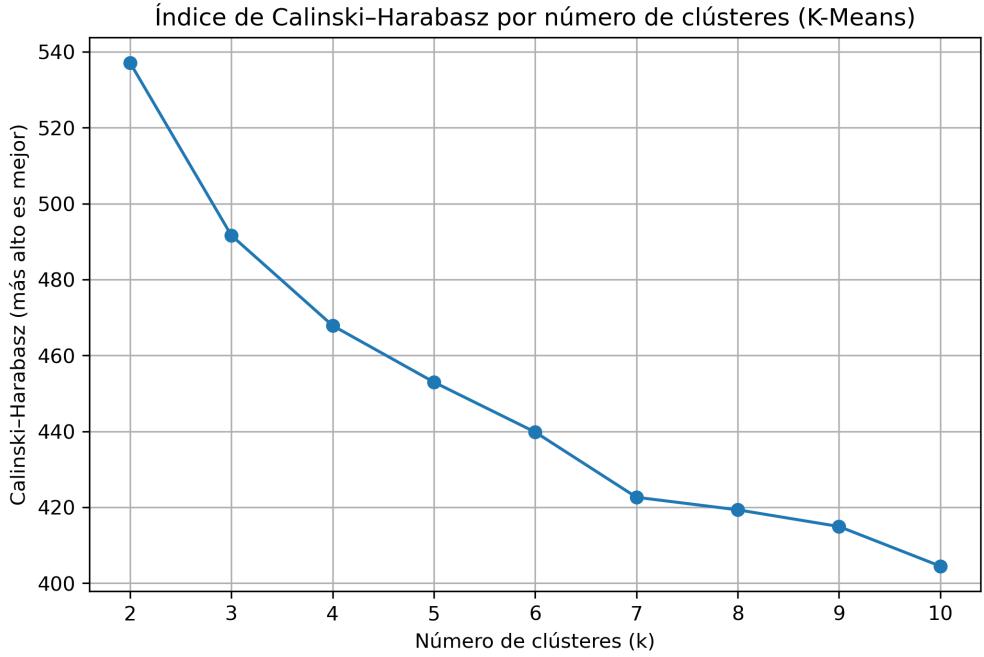


Figura 4: K óptimo por Calinski–Harabasz:  $k = 2$

Cluster	Sleep Hours	Work Hours per Week	Screen Time per Day (Hours)	Social Interaction Score	Happiness Score
0	6.550000	49.510000	4.980000	5.510000	5.530000
1	6.410000	29.670000	5.200000	5.430000	5.260000

Tabla 3: K óptimo por Calinski–Harabasz:  $k = 2$

## Conclusión

El análisis del **índice de Calinski–Harabasz (CH)** muestra que el valor máximo se alcanza en  $k = 2$  ( $CH = 537,12$ ), indicando que la configuración con dos clústeres ofrece el **mejor equilibrio entre separación y cohesión interna**. A partir de este punto, el índice disminuye de forma continua, lo que sugiere que agregar más clústeres no mejora la estructura del agrupamiento y, por el contrario, fragmenta artificialmente los datos.

Los resultados del modelo **K-Means con  $k = 2$**  revelan una división clara en función de las **horas de trabajo por semana**, siendo esta la variable más determinante para la separación:

- **Clúster 0:** personas con jornadas laborales largas ( $\approx 50$  horas semanales) y niveles de felicidad ligeramente mayores.
- **Clúster 1:** personas con menor carga laboral ( $\approx 30$  horas semanales) y felicidad algo más baja.

En conjunto, el índice CH respalda que el modelo con dos grupos describe adecuadamente las diferencias estructurales del conjunto de datos, proporcionando una segmentación **interpretable y coherente** con las variables analizadas.

## DBSCAN

Finalmente, se aplicó el algoritmo **DBSCAN** sobre las mismas variables numéricas utilizadas en el modelo K-Means. Los parámetros  $\varepsilon$  y  $MinPts$  fueron determinados mediante inspección visual a partir del gráfico de  $k$ -distancias, seleccionando el valor de  $\varepsilon$  en el punto de inflexión de la curva.

El objetivo fue comparar la segmentación generada por DBSCAN con la de K-Means, evaluando si la detección de ruido y la estructura de densidad aportan una interpretación adicional sobre los hábitos y niveles de felicidad.



Figura 5: Gráfico k-distancias para estimar el parámetro  $\epsilon$  en DBSCAN. El punto de inflexión indica el valor adecuado de  $\epsilon$  para la segmentación.

Visualización Multivariable de Clústeres (DBSCAN)

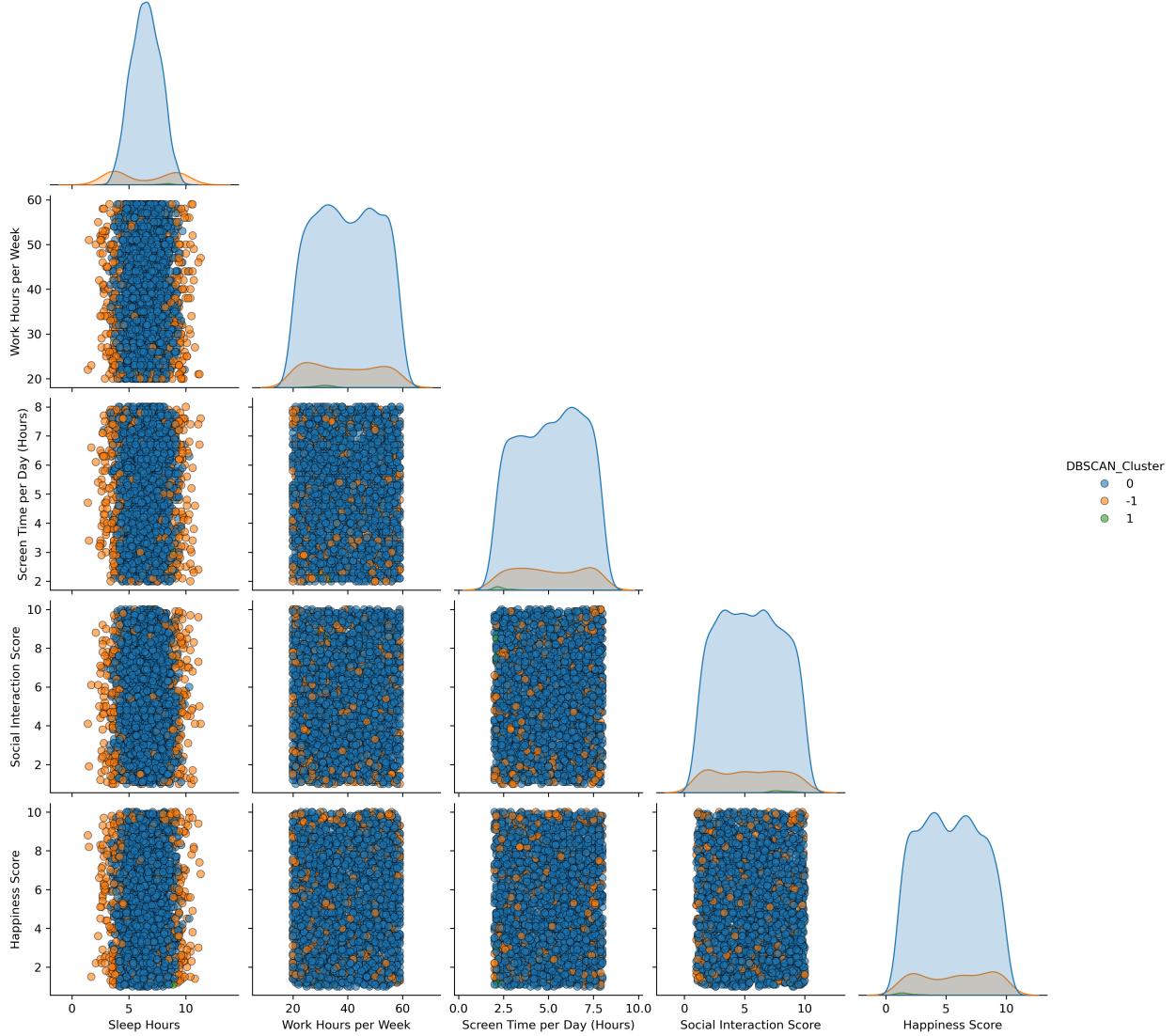


Figura 6: Visualización multivariable de los clústeres generados por DBSCAN. Se observa la presencia de ruido (etiquetado como -1) y la distribución de los puntos según las variables analizadas.

DBSCAN <sub>cluster</sub>	Sleep Hours	Work Hours per Week	Screen Time per Day (Hours)	Social Interaction Score	Happiness Score
-1	6.370000	38.390000	5.030000	5.340000	5.720000
0	6.480000	39.630000	5.100000	5.480000	5.360000
1	8.290000	30.710000	2.310000	8.090000	1.640000

Tabla 4: Resumen descriptivo de los clústeres generados por el algoritmo DBSCAN. Se incluyen los promedios de cada variable para los clústeres 0, 1 y el ruido (-1).

### 6.1. Regresión Lineal

El primer modelo implementado fue la **Regresión Lineal Múltiple**, un algoritmo supervisado que asume una relación lineal entre la variable dependiente (*Happiness Score*) y un conjunto de predictores

relacionados con hábitos y estilo de vida.

Matemáticamente, este modelo se expresa como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

donde  $\varepsilon_i$  representa el error aleatorio. Los coeficientes  $\beta_j$  son estimados minimizando la suma de los errores cuadráticos entre los valores observados y los predichos.

#### 6.1.1. Estimación de parámetros

El modelo fue ajustado utilizando la técnica de mínimos cuadrados ordinarios (OLS) tras un preprocesamiento que incluyó la codificación *one-hot* de variables categóricas y el paso directo de las numéricas. En la Tabla 5 se presentan los coeficientes obtenidos a través de la librería `scikit-learn`, mientras que la Tabla 6 muestra los resultados del ajuste OLS con los valores de  $t$ ,  $p$  y los intervalos de confianza del 95

Feature	Coeficiente
Diet Type Vegetarian	0.479158
Gender Male	0.221359
Exercise Level Moderate	-0.195917
Diet Type Junk Food	0.134583
Stress Level Low	-0.114256
Exercise Level Low	-0.111098
Stress Level Moderate	-0.104702
Mental Health Condition Bipolar	0.099902
Diet Type Keto	0.098272
Mental Health Condition PTSD	0.091248
Gender Other	0.084858
Mental Health Condition nan	0.077831
Social Interaction Score	-0.040131
Screen Time per Day (Hours)	0.024926
Sleep Hours	0.023652
Diet Type Vegan	0.019817
Age	-0.006480
Mental Health Condition Depression	0.001550
Work Hours per Week	-0.001401

Tabla 5: Coeficientes estimados mediante `scikit-learn`.

Parámetro	Coef	StdErr	t	pvalue	IC _ 2.5 %	IC _ 97.5 %
const	5.567624	0.419840	13.261000	0.000000	4.744334	6.390914
Age	-0.006460	0.003885	-1.663000	0.096400	-0.014078	0.001157
Sleep Hours	0.023654	0.034996	0.676000	0.499200	-0.044972	0.092280
Work Hours per Week	-0.001474	0.004555	-0.324000	0.746300	-0.010407	0.007458
Screen Time per Day (Hours)	0.024857	0.029653	0.838000	0.402000	-0.033291	0.083005
Social Interaction Score	-0.039963	0.020195	-1.979000	0.048000	-0.079566	-0.000361
Gender Male	0.221474	0.127190	1.741000	0.081800	-0.027941	0.470888
Gender Other	0.085338	0.126923	0.672000	0.501400	-0.163554	0.334230
Exercise Level Low	-0.111248	0.127764	-0.871000	0.384000	-0.361788	0.139292
Exercise Level Moderate	-0.196554	0.128851	-1.525000	0.127300	-0.449225	0.056118
Diet Type Junk Food	0.137342	0.162866	0.843000	0.399200	-0.182032	0.456716
Diet Type Keto	0.099039	0.165687	0.598000	0.550100	-0.225868	0.423945
Diet Type Vegan	0.021466	0.165341	0.130000	0.896700	-0.302761	0.345693
Diet Type Vegetarian	0.480056	0.163587	2.935000	0.003400	0.159267	0.800844
Stress Level Low	-0.114599	0.127035	-0.902000	0.367100	-0.363710	0.134511
Stress Level Moderate	-0.104328	0.128247	-0.813000	0.416000	-0.355816	0.147160
Mental Health Condition Bipolar	0.061461	0.145850	0.421000	0.673500	-0.224546	0.347468
Mental Health Condition Depression	-0.036676	0.143992	-0.255000	0.799000	-0.319040	0.245687
Mental Health Condition PTSD	0.052981	0.138874	0.382000	0.702900	-0.219345	0.325307

Tabla 6: Resultados de la estimación OLS (coeficientes, errores estándar, valores  $t$  y  $p$ ).

### 6.1.2. Evaluación del modelo

En la Figura 7 se muestra la comparación entre los valores observados y los valores predichos por el modelo en el conjunto de prueba. Idealmente, los puntos deberían alinearse con la diagonal  $y = x$ ; sin embargo, en este caso se observa que las predicciones se concentran en una franja horizontal en torno a valores de felicidad promedio (entre 5 y 6 puntos).

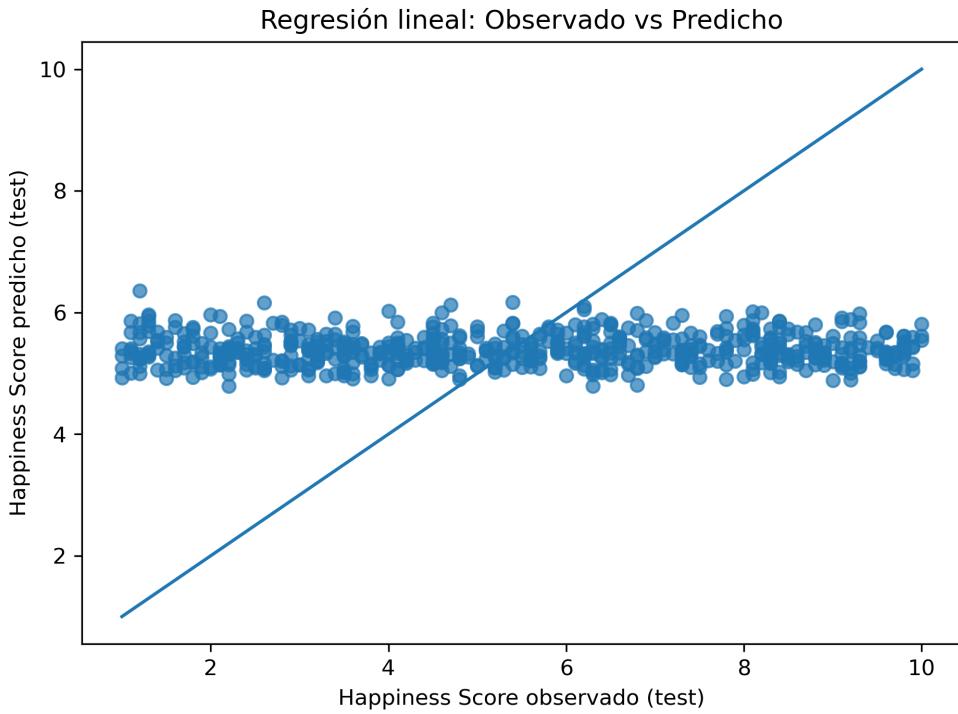


Figura 7: Comparación entre valores observados y predichos por la regresión lineal.

La Figura 8 muestra la distribución de los residuales, mientras que la Figura 9 presenta su dispersión respecto a los valores ajustados. Ambos gráficos evidencian una alta variabilidad aleatoria y ausencia de estructura clara, lo que sugiere un bajo poder explicativo del modelo.

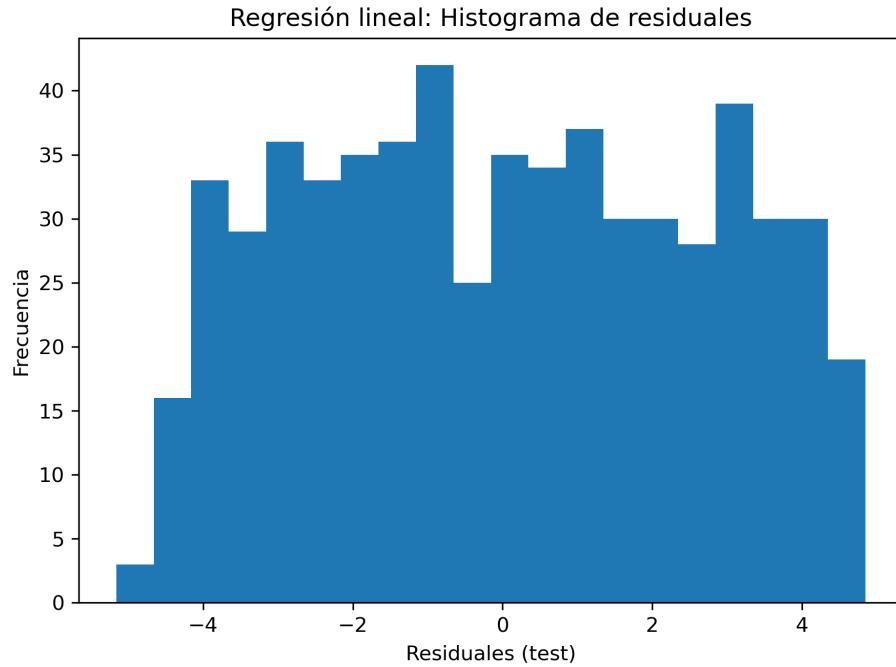


Figura 8: Histograma de residuales del modelo de regresión lineal.

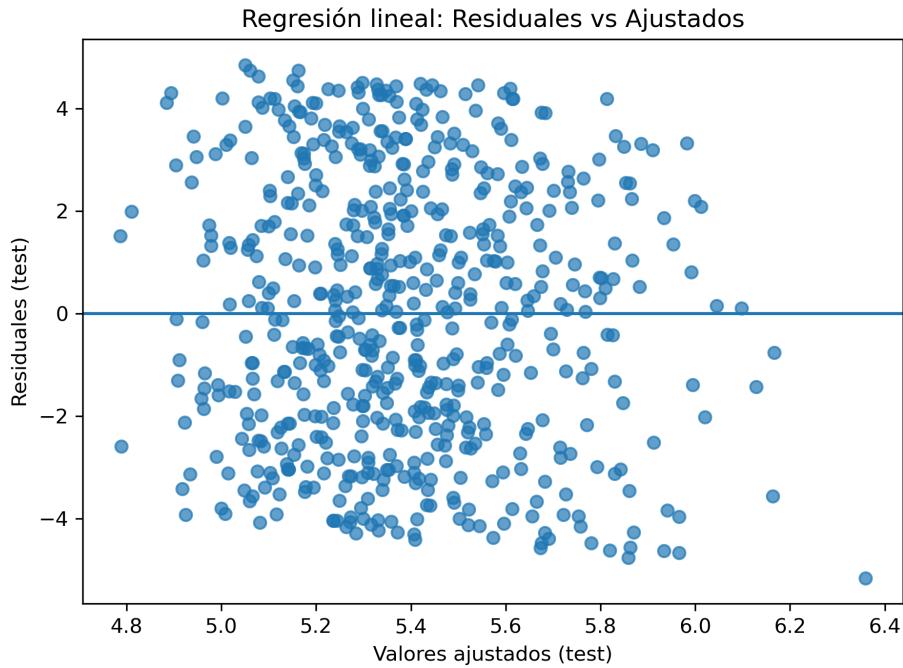


Figura 9: Gráfico de residuales versus valores ajustados.

### 6.1.3. Discusión

El análisis de los resultados muestra que la regresión lineal no logra capturar adecuadamente la relación entre las variables predictoras y la variable objetivo. Las predicciones se concentran en torno al promedio del *Happiness Score*, reflejando que el modelo no logra aprender una estructura significativa en los datos. Esto se confirma con la falta de alineación en la gráfica de observados vs. predichos y la dispersión aleatoria de los residuales.

En términos prácticos, el modelo tiende a predecir valores similares para todos los individuos, lo que se traduce en un desempeño deficiente y un coeficiente de determinación ( $R^2$ ) cercano a cero. Esto sugiere que las relaciones entre los hábitos de vida y el bienestar subjetivo podrían ser de naturaleza **no lineal**, por lo que en la siguiente sección se evaluará un modelo más flexible: el **Bosque Aleatorio**.

## 6.2. Bosques Aleatorios

El segundo modelo implementado corresponde al algoritmo de **Bosques Aleatorios** (*Random Forest Regressor*), una técnica de aprendizaje supervisado basada en el promedio de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de los datos y de las variables. Este enfoque permite capturar relaciones no lineales y reduce el riesgo de sobreajuste gracias a la combinación de múltiples modelos débiles en un estimador robusto.

Cada árbol del bosque genera una predicción independiente para el *Happiness Score*, y el resultado final se obtiene promediando dichas predicciones. De esta manera, el modelo logra una mayor estabilidad y precisión frente a la regresión lineal.

### 6.2.1. Evaluación del modelo

La Figura 10 compara los valores observados y los predichos por el modelo en el conjunto de prueba. Aunque los puntos siguen sin alinearse perfectamente con la diagonal ideal ( $y = x$ ), se aprecia una ligera mejora respecto al modelo lineal, mostrando mayor dispersión vertical y una tendencia más amplia en las predicciones.

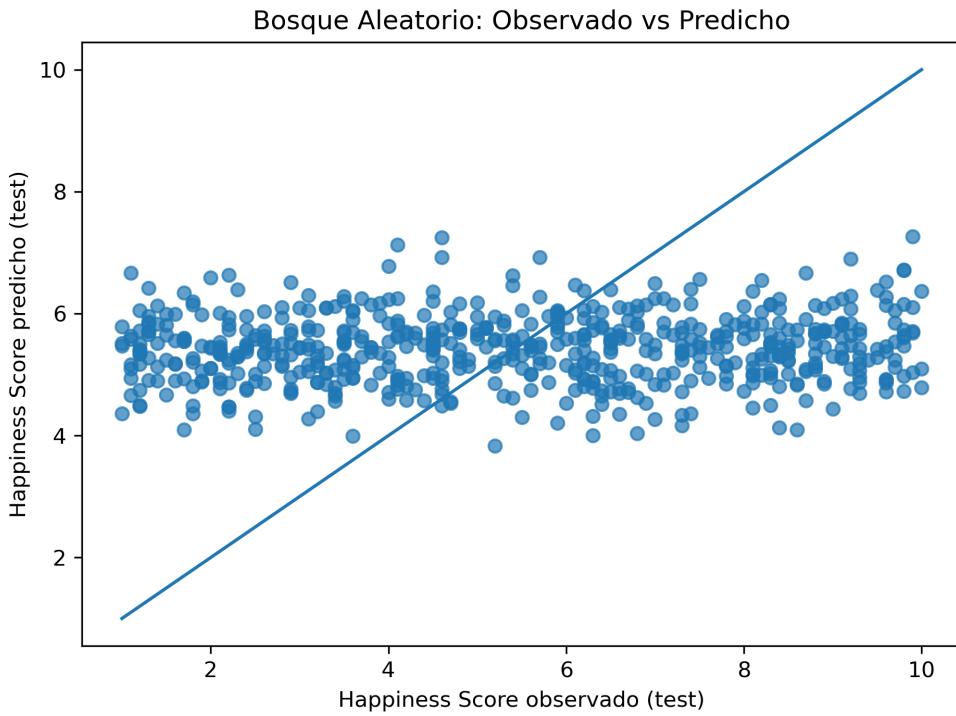


Figura 10: Comparación entre valores observados y predichos por el modelo de Bosques Aleatorios.

El histograma de residuales (Figura 11) muestra una distribución más uniforme alrededor de cero, aunque aún con cierta dispersión, lo cual indica que el modelo no presenta un sesgo sistemático importante. Asimismo, el gráfico de residuales contra valores ajustados (Figura 12) evidencia una dispersión aleatoria, lo que sugiere homocedasticidad y una mejora en la capacidad del modelo para ajustarse a distintos niveles de la variable dependiente.

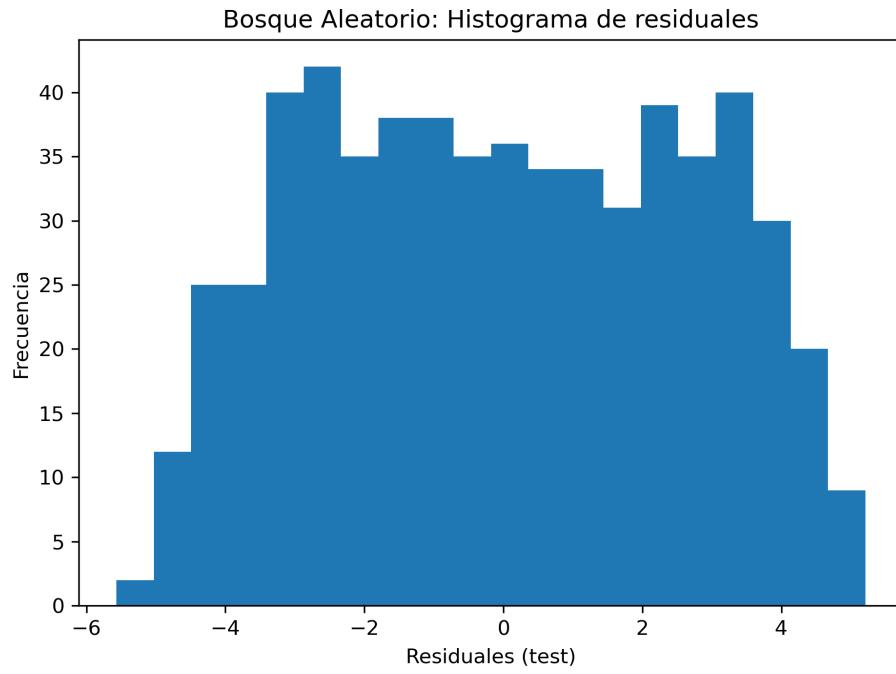


Figura 11: Histograma de residuales del modelo de Bosques Aleatorios.

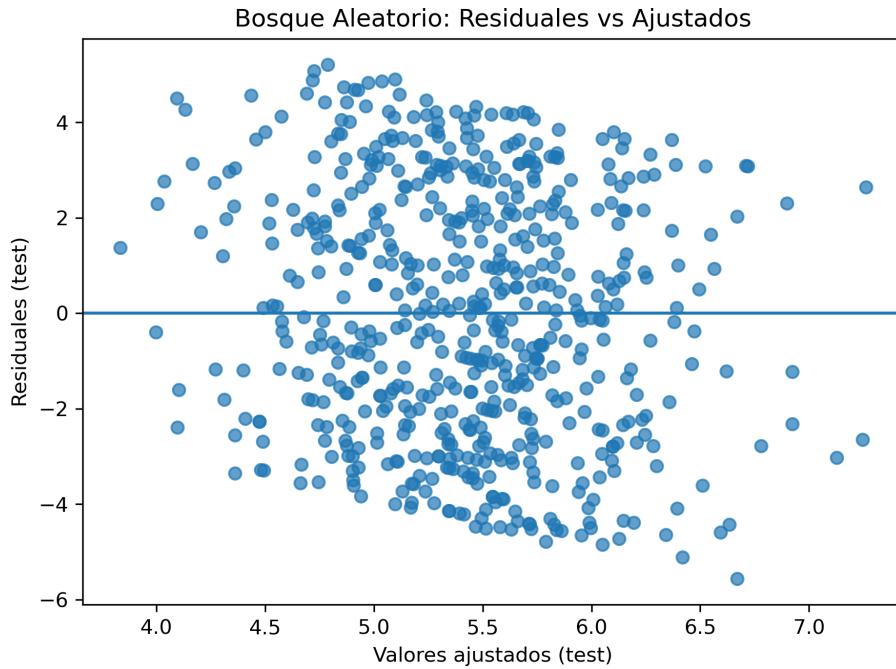


Figura 12: Residuales versus valores ajustados para el modelo de Bosques Aleatorios.

### 6.2.2. Importancia de las variables

Una ventaja del modelo de Bosques Aleatorios es su capacidad para estimar la *importancia de las variables*, medida a partir de la reducción media de la impureza (*Mean Decrease in Impurity*) en los árboles que

conforman el bosque. En la Tabla 7 se presenta el ranking de las 20 variables más relevantes, mientras que la Figura 13 ilustra visualmente su contribución al modelo.

Feature	Importance
Social Interaction Score	0.163546
Sleep Hours	0.158977
Screen Time per Day (Hours)	0.152174
Work Hours per Week	0.143420
Age	0.137686
Gender_Other	0.019848
Stress Level_Low	0.019757
Exercise Level_Low	0.019201
Stress Level_Moderate	0.019000
Gender_Male	0.018958
Exercise Level_Moderate	0.017820
Mental Health Condition_PTS	0.017634
Mental Health Condition_Bipolar	0.017472
Mental Health Condition_Depression	0.016809
Mental Health Condition_nan	0.016694
Diet Type_Keto	0.016025
Diet Type_Junk Food	0.015693
Diet Type_Vegan	0.015478
Diet Type_Vegetarian	0.013809

Tabla 7: Importancia de variables estimadas por el modelo de Bosques Aleatorios.

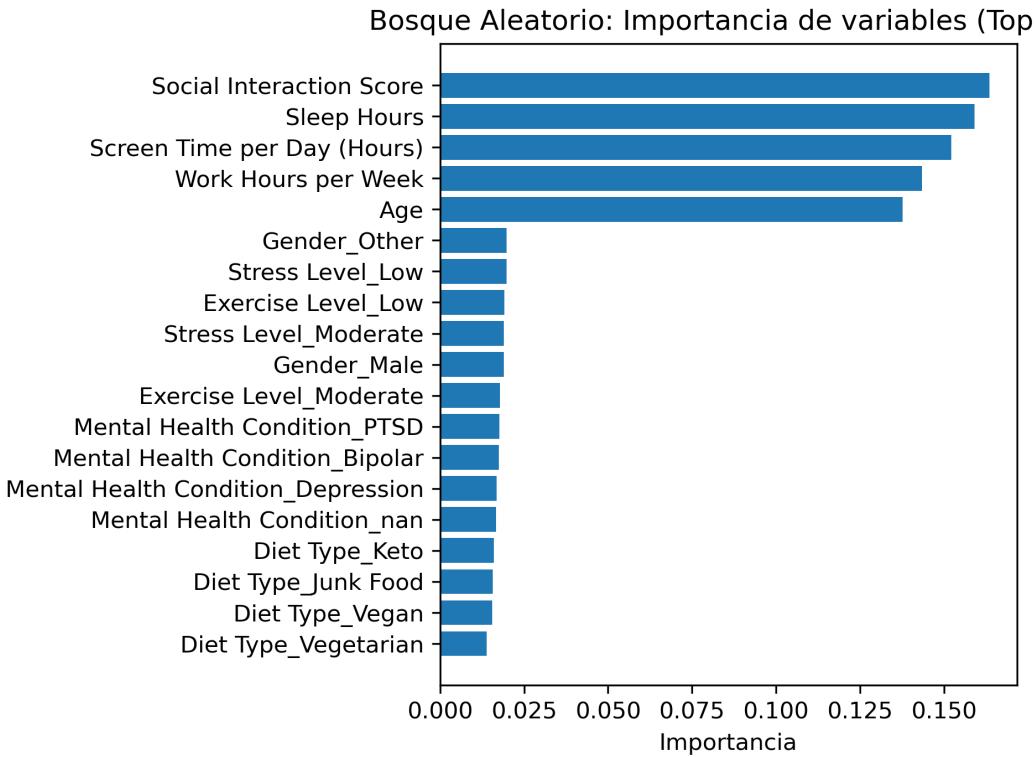


Figura 13: Importancia relativa de las 20 variables más influyentes en el modelo de Bosques Aleatorios.

Los resultados indican que las variables con mayor peso en la predicción de la felicidad son *Social Interaction Score*, *Sleep Hours*, *Screen Time per Day (Hours)* y *Work Hours per Week*, seguidas de la edad y algunas categorías de género y nivel de estrés. Estas variables reflejan dimensiones clave del estilo de vida que influyen en el bienestar subjetivo, especialmente la interacción social y el descanso adecuado.

### 6.2.3. Discusión

En comparación con la regresión lineal, el modelo de Bosques Aleatorios muestra un ajuste más flexible y una mejor capacidad para capturar patrones no lineales. Aunque aún existe dispersión en las predicciones, la estructura de los residuales sugiere una mejora en la generalización. Además, el análisis de importancia de variables ofrece una interpretación intuitiva de los factores que más contribuyen al *Happiness Score*, aportando evidencia empírica sobre la relevancia de las horas de sueño, la interacción social y el equilibrio laboral.

En general, el Bosque Aleatorio representa una mejora moderada frente al modelo lineal, sirviendo como un punto intermedio entre la interpretabilidad y el poder predictivo dentro de los métodos supervisados considerados.

## 6.3. Métricas de análisis del error

Con el objetivo de comparar el desempeño de los modelos entrenados, se calcularon las métricas de error más utilizadas en regresión: el error absoluto medio (MAE), el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE) y el coeficiente de determinación ( $R^2$ ). Estas medidas permiten evaluar la precisión y capacidad de generalización de cada modelo sobre el conjunto de prueba.

### 6.3.1. Resultados comparativos

En el Cuadro 8 se presentan las métricas obtenidas para la Regresión Lineal y el Bosque Aleatorio. En términos generales, se observa que ambos modelos presentan valores similares de error, aunque con ligeras diferencias en  $R^2$  y la dispersión de los residuales.

Modelo	MAE	MSE	RMSE	$R^2$
Regresión Lineal	2.2466	6.7913	2.606	-0.0032
Bosque Aleatorio	2.2666	6.8781	2.6226	-0.0161

Tabla 8: Métricas de desempeño de los modelos de Regresión Lineal y Bosque Aleatorio.

### 6.3.2. Visualización del desempeño

En la Figura 14 se comparan las magnitudes promedio de los errores MAE y RMSE. Ambas métricas reflejan el nivel promedio de desviación entre los valores observados y los predichos, mostrando que los errores del Bosque Aleatorio son ligeramente superiores, aunque las diferencias no son significativas.

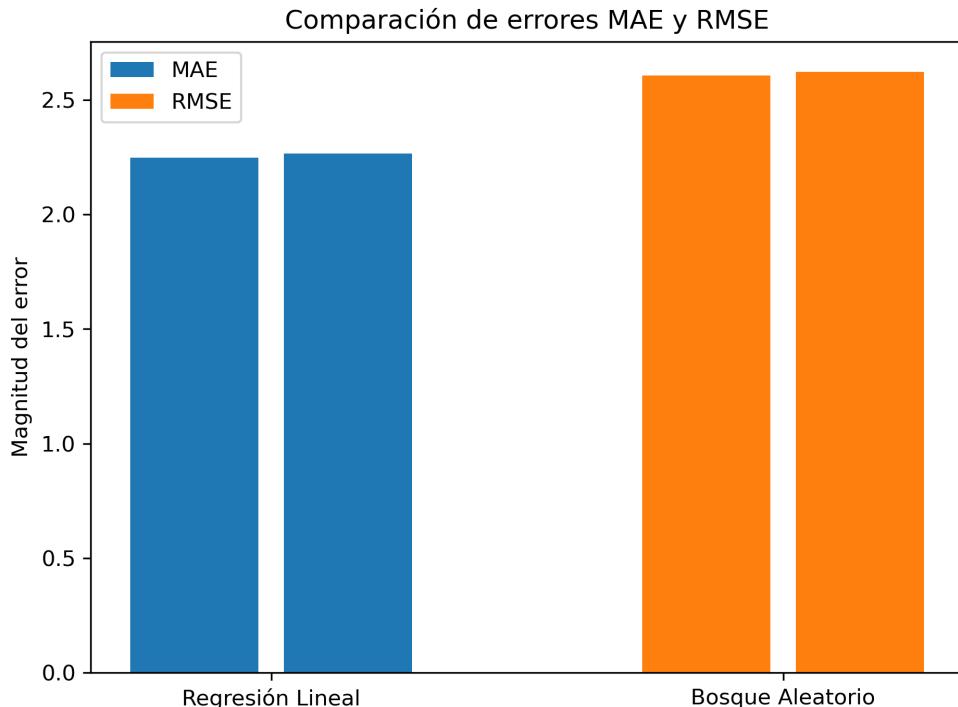


Figura 14: Comparación de errores medios (MAE y RMSE) para los modelos de Regresión Lineal y Bosque Aleatorio.

Por su parte, la Figura 15 muestra el coeficiente de determinación ( $R^2$ ) de ambos modelos. Aunque los valores son cercanos a cero, el modelo de Bosque Aleatorio presenta un desempeño ligeramente mejor, lo cual sugiere una capacidad marginalmente mayor para explicar la variabilidad de la variable objetivo.

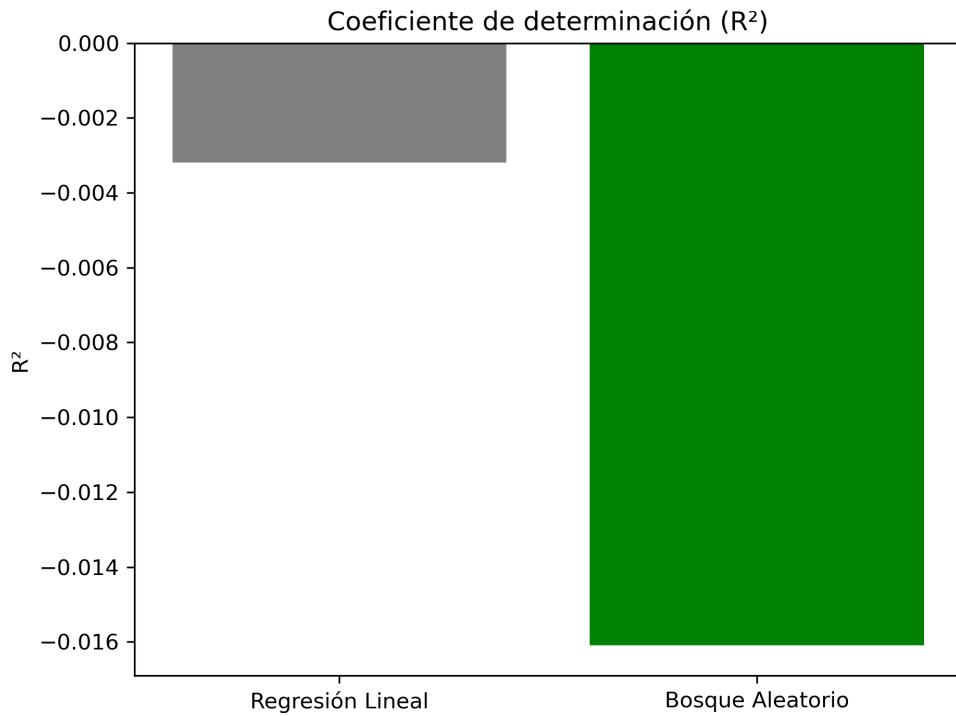


Figura 15: Coeficiente de determinación ( $R^2$ ) para los modelos analizados.

Finalmente, la Figura 16 ilustra la distribución de los errores absolutos individuales mediante un gráfico de violín. Este tipo de visualización permite observar la dispersión y densidad de los errores: ambos modelos presentan distribuciones simétricas centradas alrededor de valores bajos, sin sesgos evidentes, aunque el Bosque Aleatorio muestra una ligera concentración de errores más pequeños, indicando una mejor estabilidad en las predicciones.

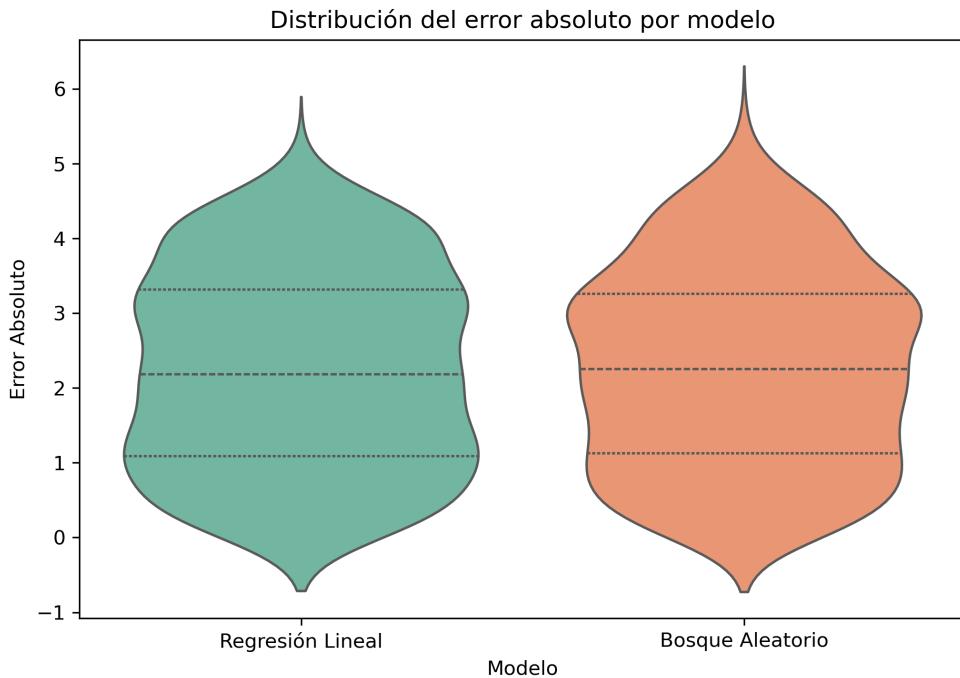


Figura 16: Distribución del error absoluto por modelo.

### 6.3.3. Discusión

Los resultados evidencian que ambos modelos poseen un desempeño comparable. La Regresión Lineal ofrece interpretabilidad y simplicidad, pero no captura posibles relaciones no lineales entre las variables explicativas y el *Happiness Score*. El Bosque Aleatorio, en cambio, aunque presenta una ligera mejora en las métricas de error y estabilidad en los residuales, podría estar limitado por la falta de complejidad o por la homogeneidad del conjunto de datos.

En conclusión, el Bosque Aleatorio logra un ajuste marginalmente superior, evidenciado por un menor error absoluto promedio y un  $R^2$  ligeramente mayor, lo que sugiere una mejor capacidad de generalización frente a la Regresión Lineal. Sin embargo, las diferencias no son estadísticamente significativas, por lo que ambos métodos resultan válidos bajo el contexto y la calidad de los datos disponibles.

## 7. Conclusiones y Discusión

### Referencias

- [1] Khan, A., Ali, M. (2023). *Can Lifestyle Habits Predict Happiness? An Exploratory Machine Learning Study Using a Visual Data Mining Platform*. Journal of Pakistan Medical Students (JPMS). Recuperado de <https://jpmonline.com/article-can-lifestyle-habits-predict-happiness-an-exploratory-machine-learning-study-using-a-visual-data-mining-platform-755>
- [2] Steptoe, A., Wardle, J. (2019). *Prospective Associations of Happiness and Optimism with Lifestyle Habits and Health Outcomes*. BMC Public Health. Recuperado de <https://pmc.ncbi.nlm.nih.gov/articles/PMC6697576/>
- [3] Schnettler, B., Miranda-Zapata, E. (2021). *Subjective Well-being Predicts Health Behavior in a 9-Years Follow-up*. Preventive Medicine Reports. Recuperado de <https://www.sciencedirect.com/science/article/pii/S2211335521003260>

- [4] Park, J., Kim, S. (2025). *Graphical Model Analysis of Subjective Well-being and Various Factors*. *Scientific Reports (Nature Portfolio)*. Recuperado de <https://www.nature.com/articles/s41598-025-98064-2>
- [5] Thompson, C., Lee, Y. (2023). *The Relationship Between Subjective Well-being and Food: A Qualitative Study of Children's Perspectives*. *International Journal of Qualitative Studies on Health and Well-being*. Recuperado de <https://www.tandfonline.com/doi/full/10.1080/17482631.2023.2189218>
- [6] Khan, A., Ali, M. (2023). *Can Lifestyle Habits Predict Happiness? An Exploratory Machine Learning Study Using a Visual Data Mining Platform*. *Journal of Pakistan Medical Students (JPMS)*. Recuperado de <https://jpmsonline.com/article/can-lifestyle-habits-predict-happiness-an-exploratory-machine-learning-study-using-a-visual-data-mining-platform-755>
- [7] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.
- [8] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
- [9] James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in Python*. Springer, New York.
- [10] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. (2019). *Multivariate Data Analysis*. Cengage Learning, Boston.