Algoritmos supervisados aplicados a analizar la relación entre la salud mental, hábitos y felicidad.

Autor: Fisam Zavala

November 2025

1. Introducción

En el presente trabajo se busca aplicar técnicas de aprendizaje supervisado para el pronóstico de variables relacionadas con el bienestar y los hábitos de vida. El conjunto de datos utilizado, *Mental Health and Lifestyle Dataset*, contiene información sobre variables laborales y conductuales, así como indicadores subjetivos de felicidad y salud mental.

El objetivo general consiste en desarrollar un modelo predictivo capaz de estimar el **Happiness Score** (nivel de felicidad) a partir de variables explicativas como *Sleep Hours*, *Work Hours per Week*, *Screen Time per Day*, *Social Interaction Score* y otras características del estilo de vida.

Este análisis forma parte de la unidad temática de *Pronóstico*, donde se busca evaluar la capacidad de los modelos supervisados para generalizar patrones y realizar predicciones precisas en contextos reales. Se evaluará la precisión de los modelos mediante métricas estándar de error como el **MAE (Mean Absolute Error)**, **RMSE (Root Mean Squared Error)** y **MSE (Mean Squared Error)**.

2. 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:

- **Regresión Lineal:** busca establecer una relación lineal entre la variable dependiente 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 β_j se estiman minimizando la suma de los errores cuadráticos:

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

  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.

- **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.

- **Regresión por Soporte Vectorial (SVR):** busca encontrar una función $f(x)$ que se mantenga dentro de un margen de tolerancia ϵ respecto a los valores observados. La función objetivo es minimizar:

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

sujeto a las restricciones:

$$\begin{cases} y_i - (w \cdot x_i + b) \leq \epsilon + \xi_i \\ (w \cdot x_i + b) - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

donde C controla el equilibrio entre la complejidad del modelo y el grado de tolerancia al error.

- **Redes Neuronales Artificiales (ANN):** consisten en capas de neuronas que aplican transformaciones no lineales a los datos de entrada. Cada neurona realiza una combinación lineal de los pesos y aplica una función de activación $\phi(\cdot)$:

$$a_j^{(l)} = \phi \left(\sum_{i=1}^{n_{l-1}} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)} \right)$$

donde l indica la capa, $w_{ij}^{(l)}$ los pesos sinápticos y $b_j^{(l)}$ el sesgo. El aprendizaje se realiza mediante retropropagación, minimizando una función de pérdida, comúnmente el error cuadrático medio:

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

En este estudio se implementarán modelos de tipo **Regresión Lineal** y **Bosques Aleatorios** para predecir el **Happiness Score**, comparando sus resultados con métricas de error como MAE, MSE y $RMSE$.

3. Modelo 1: Regresión Lineal

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 ε_i representa el error aleatorio. Los coeficientes β_j son estimados minimizando la suma de los errores cuadráticos entre los valores observados y los predichos.

3.1. 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 1 se presentan los coeficientes obtenidos a través de la librería `scikit-learn`, mientras que la Tabla 2 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 1: 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 2: Resultados de la estimación OLS (coeficientes, errores estándar, valores t y p).

Evaluación del modelo

En la Figura 1 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).

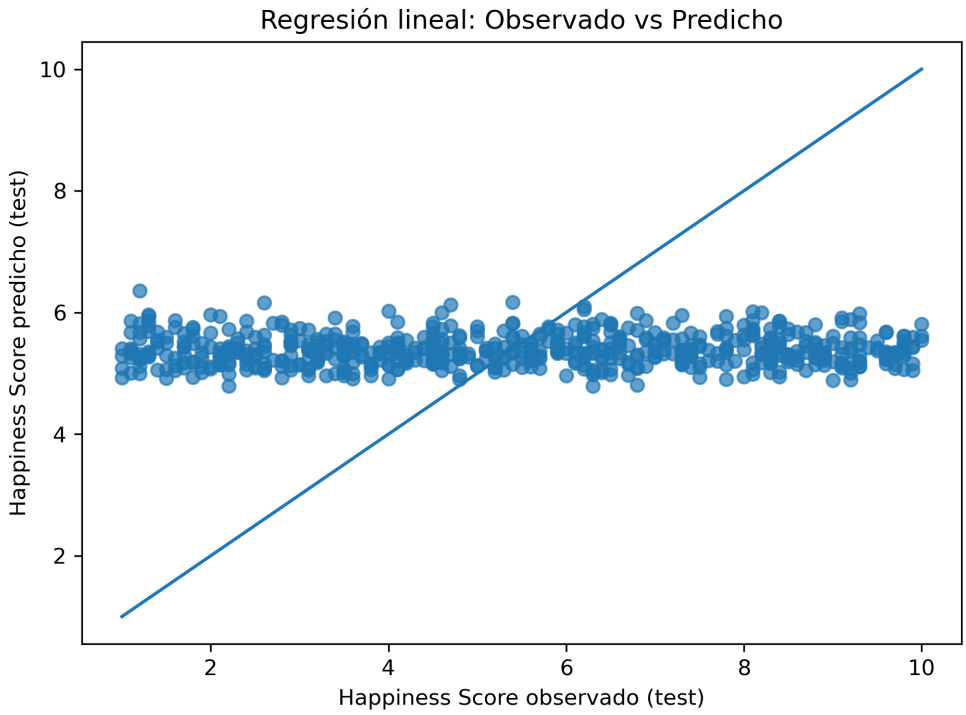

Figura 1: Comparación entre valores observados y predichos por la regresión lineal.

La Figura 2 muestra la distribución de los residuales, mientras que la Figura 3 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.

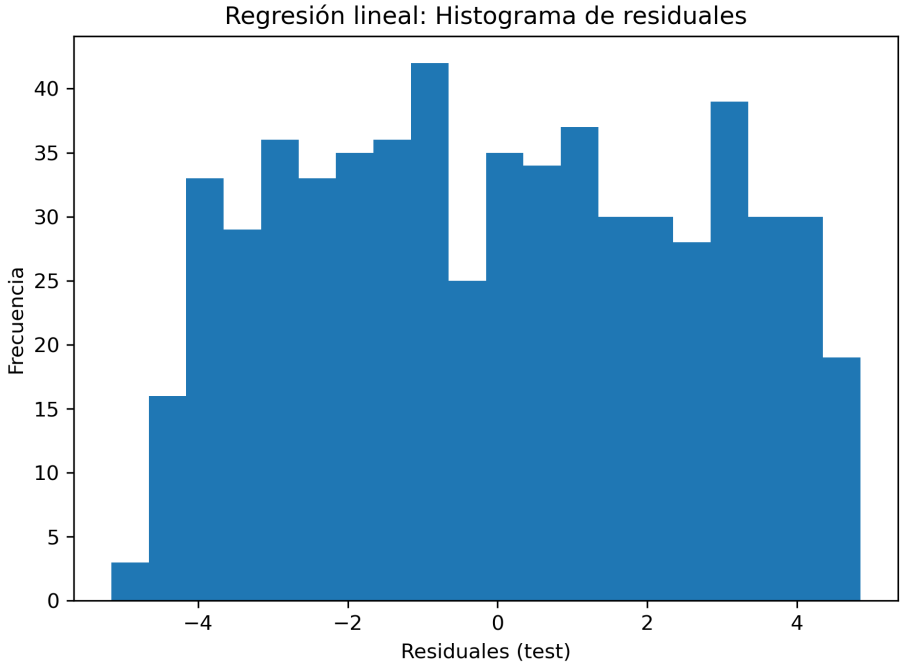

Figura 2: Histograma de residuales del modelo de regresión lineal.

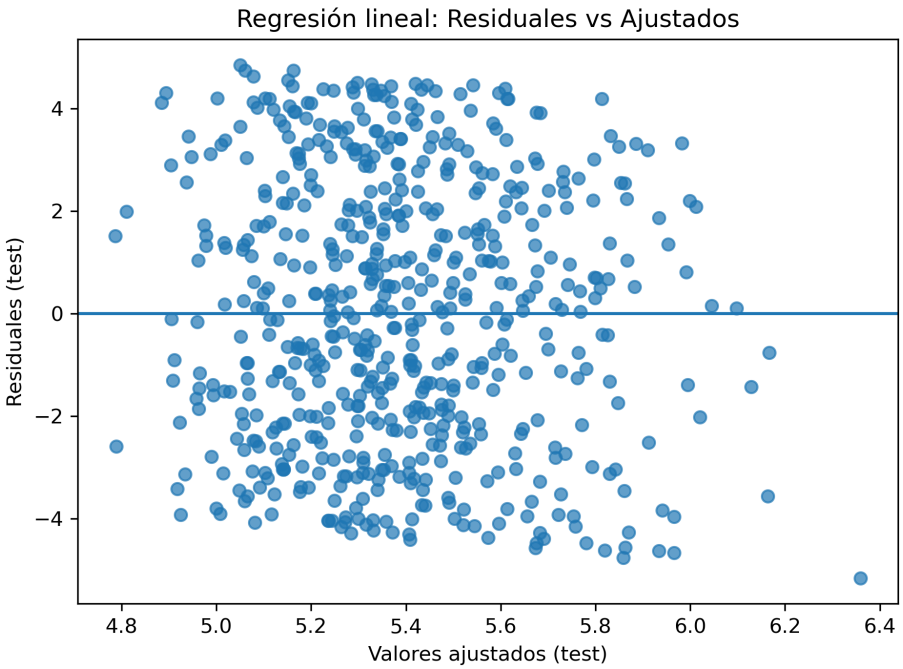

Figura 3: Gráfico de residuales versus valores ajustados.

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**.

4. Modelo 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.

4.1. Evaluación del modelo

La Figura 4 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.

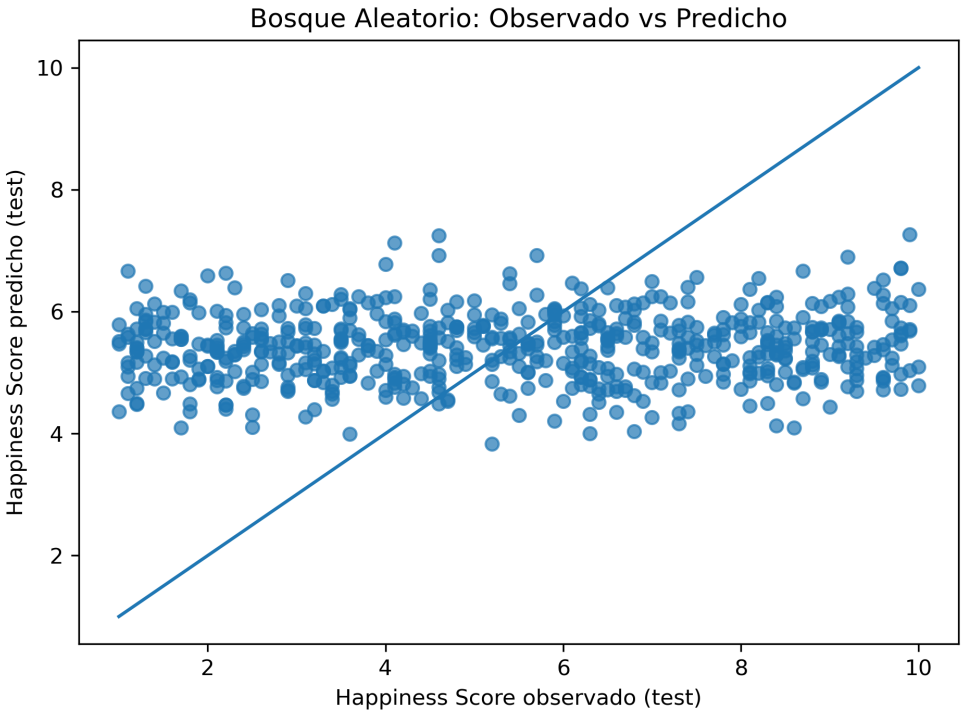

Figura 4: Comparación entre valores observados y predichos por el modelo de Bosques Aleatorios.

El histograma de residuales (Figura 5) 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 6) 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.

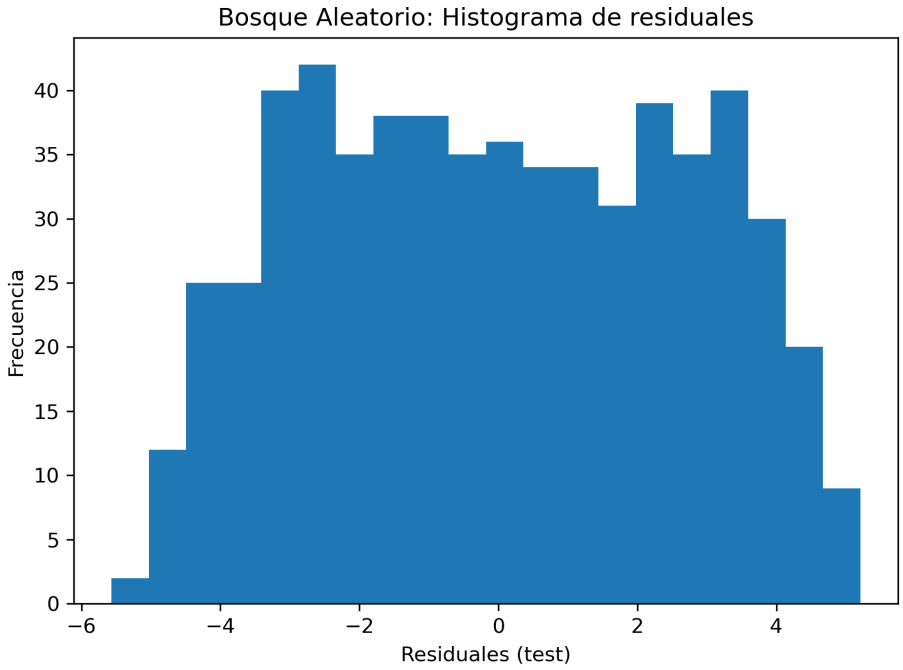

Figura 5: Histograma de residuales del modelo de Bosques Aleatorios.

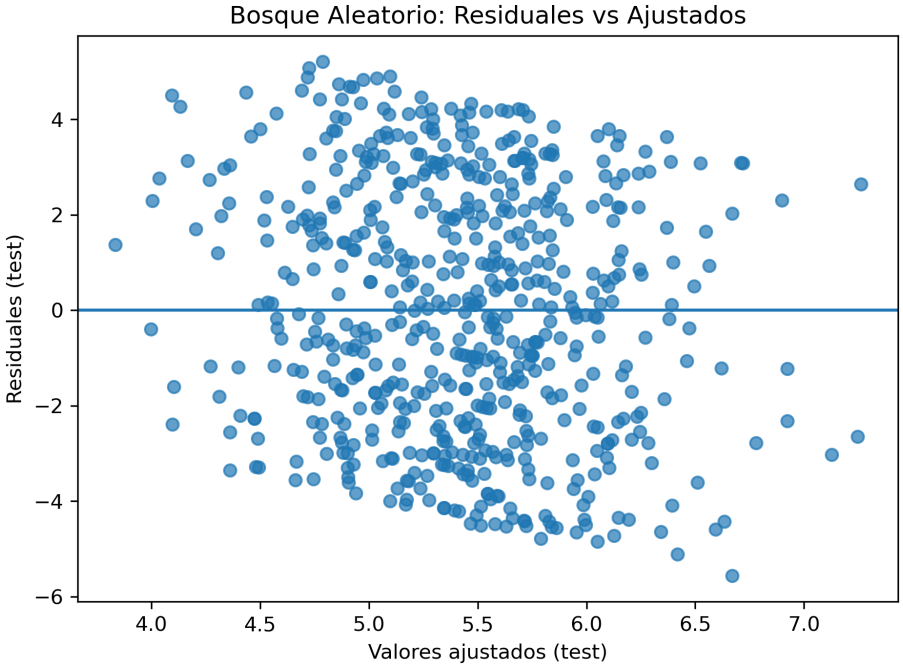

Figura 6: Residuales versus valores ajustados para el modelo de Bosques Aleatorios.

4.2. 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 3 se presenta el ranking de las 20 variables más relevantes, mientras que la Figura 7 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_PTSD	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 3: Importancia de variables estimadas por el modelo de Bosques Aleatorios.

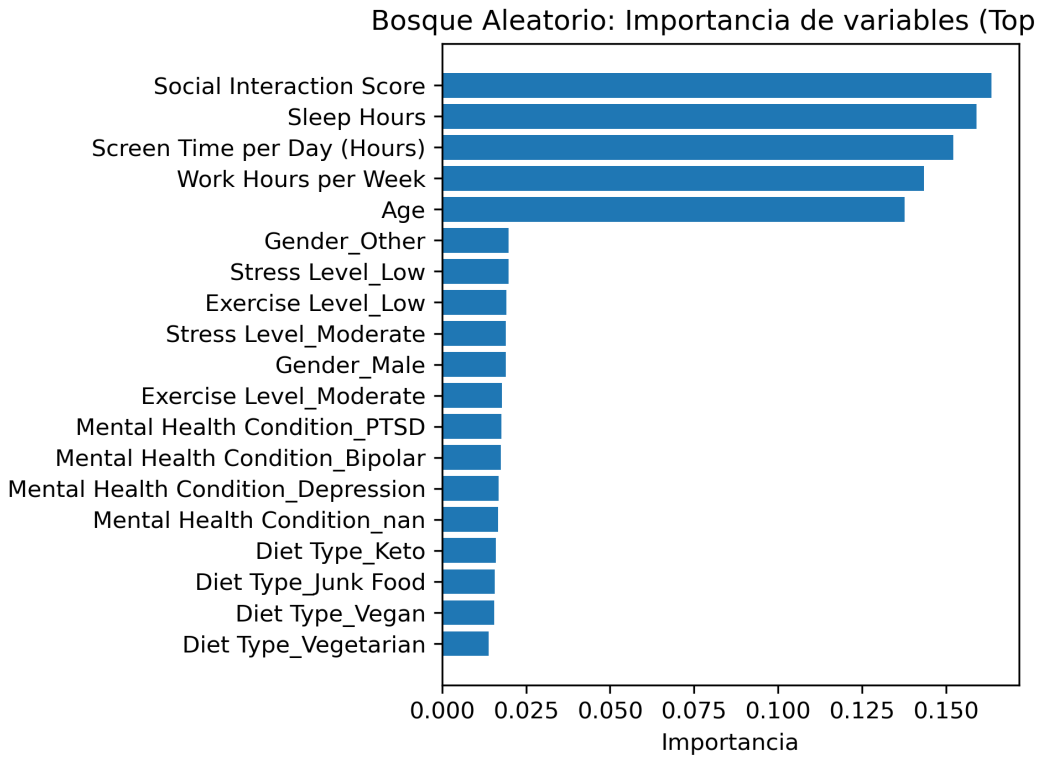

Figura 7: Importancia relativa de las 20 variables más influyentes en el modelo de Bosques Aleatorios.

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.

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.

5. 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.

5.1. Resultados comparativos

En el Cuadro 4 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 4: Métricas de desempeño de los modelos de Regresión Lineal y Bosque Aleatorio.

5.2. Visualización del desempeño

En la Figura 8 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.

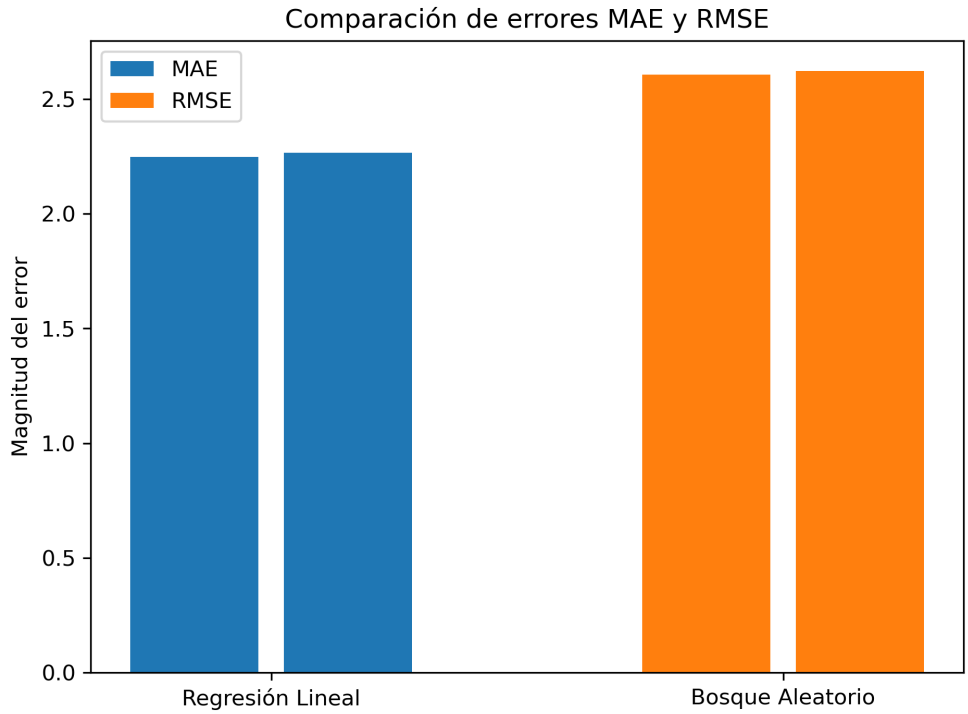

Figura 8: Comparación de errores medios (MAE y RMSE) para los modelos de Regresión Lineal y Bosque Aleatorio.

Por su parte, la Figura 9 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.

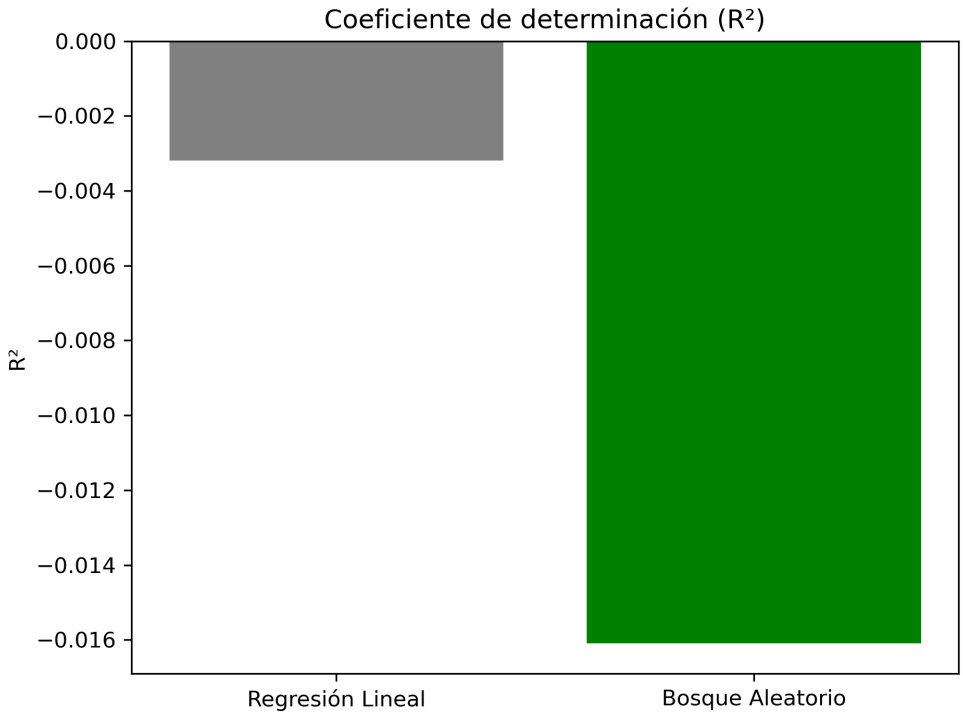

Figura 9: Coeficiente de determinación (R^2) para los modelos analizados.

Finalmente, la Figura 10 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.

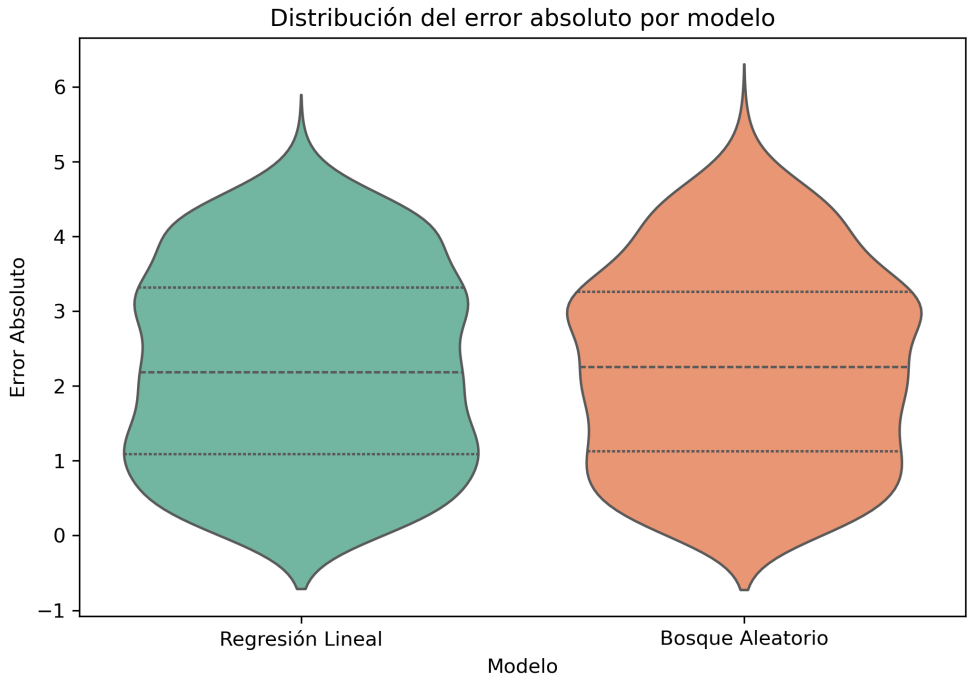

Figura 10: Distribución del error absoluto por modelo.

5.3. 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.

6. Conclusiones generales

El presente estudio permitió evaluar y comparar dos enfoques de modelado para la predicción del *Happiness Score*: un modelo de **Regresión Lineal Múltiple** y un modelo de **Bosques Aleatorios**. Ambos fueron entrenados sobre el mismo conjunto de variables de estilo de vida, con el objetivo de identificar patrones que expliquen las diferencias en el nivel de felicidad de los individuos.

6.1. Comparación de los modelos

La Regresión Lineal presentó un desempeño moderado, con errores promedio (MAE y RMSE) dentro de un rango aceptable pero con un coeficiente de determinación (R^2) cercano a cero. Esto sugiere que la relación entre las variables independientes y la variable objetivo no sigue una estructura lineal simple, por lo que el modelo lineal no logra capturar adecuadamente las interacciones entre los predictores.

El Bosque Aleatorio, al ser un modelo de naturaleza no paramétrica y capaz de capturar relaciones no lineales, mostró una ligera mejora en los indicadores de error y una distribución más estable de los residuales. Además, el análisis de importancia de variables reveló que los factores con mayor contribución al nivel de

felicidad fueron **la interacción social**, **las horas de sueño**, **el tiempo frente a pantalla** y **las horas de trabajo por semana**, lo que coincide con la literatura sobre bienestar subjetivo.

6.2. Conclusión final

En términos generales, aunque ambos modelos lograron capturar parcialmente el comportamiento del *Happiness Score*, el **Bosque Aleatorio** demostró ser más robusto y ligeramente más preciso que la Regresión Lineal. Sin embargo, las diferencias en desempeño fueron pequeñas, lo que sugiere que la variabilidad de la felicidad podría depender de factores no observados o de relaciones más complejas que las contenidas en el conjunto de datos.

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://jpmsonline.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