
Grupo 10

Katherin Escobar

Heberth Martínez

Diana Mazuera

Natalia Santamaría



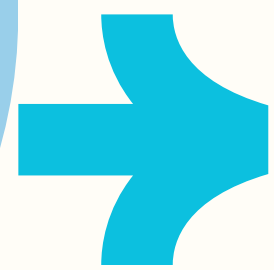
ANÁLISIS EXPLORATORIO DE DATOS

CICLO DE SUEÑO Y PRODUCTIVIDAD

Introducción

Nombre de la columna	Descripción
Fecha	La fecha de recopilación de datos
ID de persona	Identificador único para cada individuo
Edad	Edad de la persona (18-60 años)
Género	Masculino, Femenino u Otro
Hora de inicio del sueño	Hora en la que la persona se fue a dormir (en formato de 24 horas)
Hora de finalización del sueño	Hora en que la persona se despertó (en formato de 24 horas)
Horas totales de sueño	Duración total del sueño (en horas)
Calidad del sueño	Calidad del sueño autoinformada (escala: 1-10)
Ejercicio (minutos/día)	Minutos dedicados a hacer ejercicio al día
Ingesta de cafeína (mg)	Cantidad de cafeína consumida en mg
Tiempo frente a la pantalla antes de acostarse (minutos)	Tiempo dedicado al uso de pantallas antes de dormir
Horas de trabajo (hrs/día)	Total de horas de trabajo en un día
Puntuación de productividad	Puntuación de productividad autoinformada (escala: 1-10)
Puntuación del estado de ánimo	Puntuación del estado de ánimo autoinformado (escala: 1-10)
Nivel de estrés	Nivel de estrés auto-reportado (escala: 1-10)

- Analiza los hábitos de sueño y su impacto en la productividad, el estado de ánimo y los niveles de estrés.
- 5000 registros de personas entre los 18 y 60 años de edad y sus distintos estilos de vida.



Análisis Exploratorio



RangeIndex: 5000 entries, 0 to 4999

Data columns (total 15 columns):

#	Column	Non-Null Count	Dtype
0	Date	5000 non-null	object
1	Person_ID	5000 non-null	int64
2	Age	5000 non-null	int64
3	Gender	5000 non-null	object
4	Sleep Start Time	5000 non-null	float64
5	Sleep End Time	5000 non-null	float64
6	Total Sleep Hours	5000 non-null	float64
7	Sleep Quality	5000 non-null	int64
8	Exercise (mins/day)	5000 non-null	int64
9	Caffeine Intake (mg)	5000 non-null	int64
10	Screen Time Before Bed (mins)	5000 non-null	int64
11	Work Hours (hrs/day)	5000 non-null	float64
12	Productivity Score	5000 non-null	int64
13	Mood Score	5000 non-null	int64
14	Stress Level	5000 non-null	int64

dtypes: float64(4), int64(9), object(2)



PREGUNTAS SMART



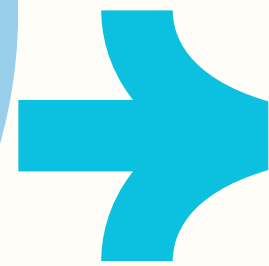
¿Qué impacto tiene el total de horas de sueño en la productividad de los trabajadores?



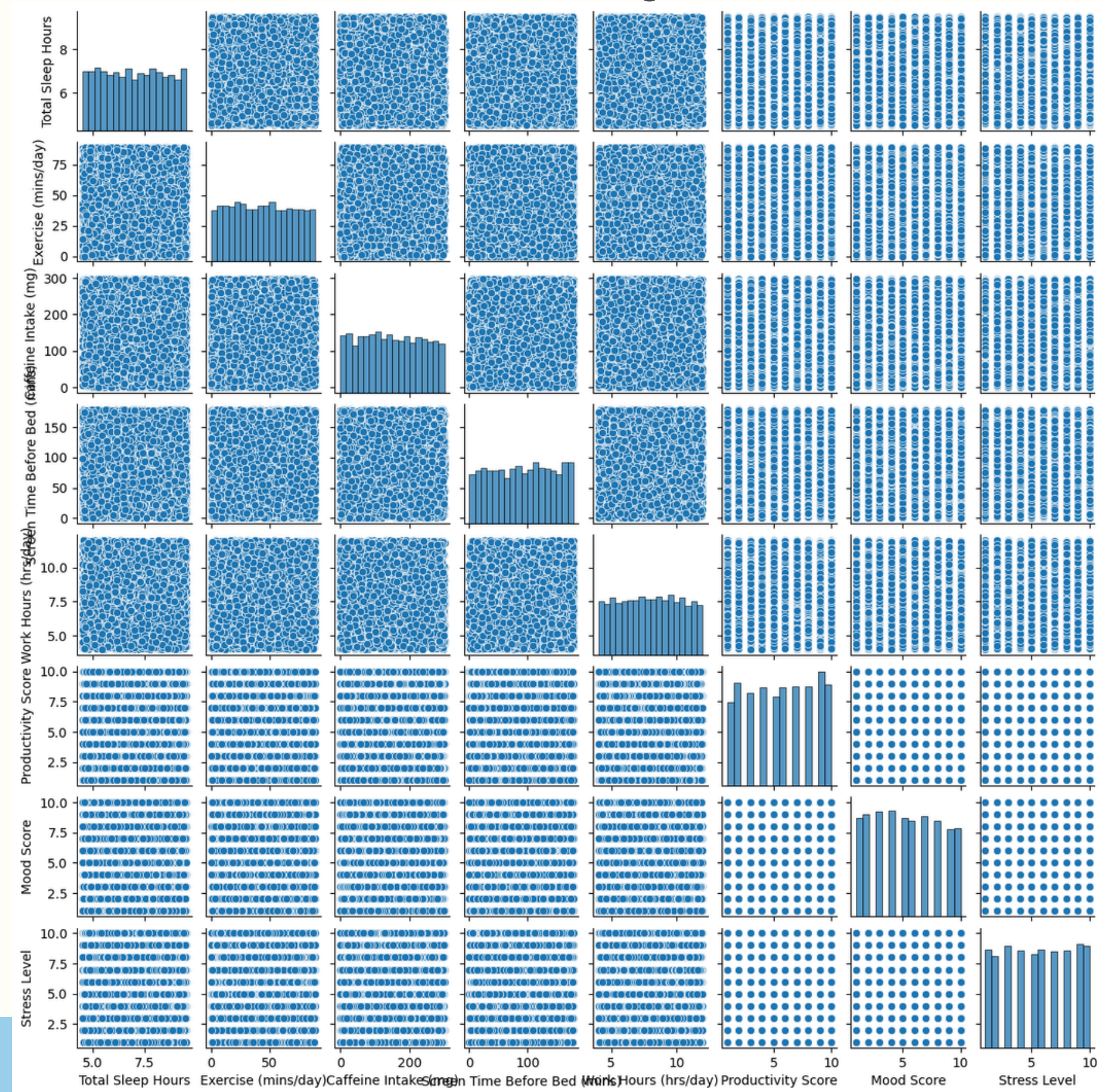
Análisis Exploratorio

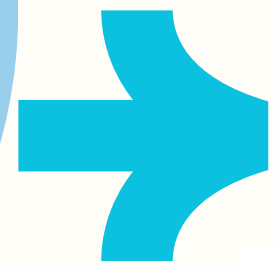


	Total Sleep Hours	Exercise (mins/day)	Caffeine Intake (mg)	Screen Time Before Bed (mins)	Work Hours (hrs/day)	Productivity Score	Mood Score	Stress Level
0	5.28	86	87	116	8.808920	8	3	6
1	5.41	32	21	88	6.329833	10	3	7
2	5.35	17	88	59	8.506306	10	9	10
3	7.55	46	34	80	6.070240	8	4	2
4	6.75	61	269	94	11.374994	8	7	9

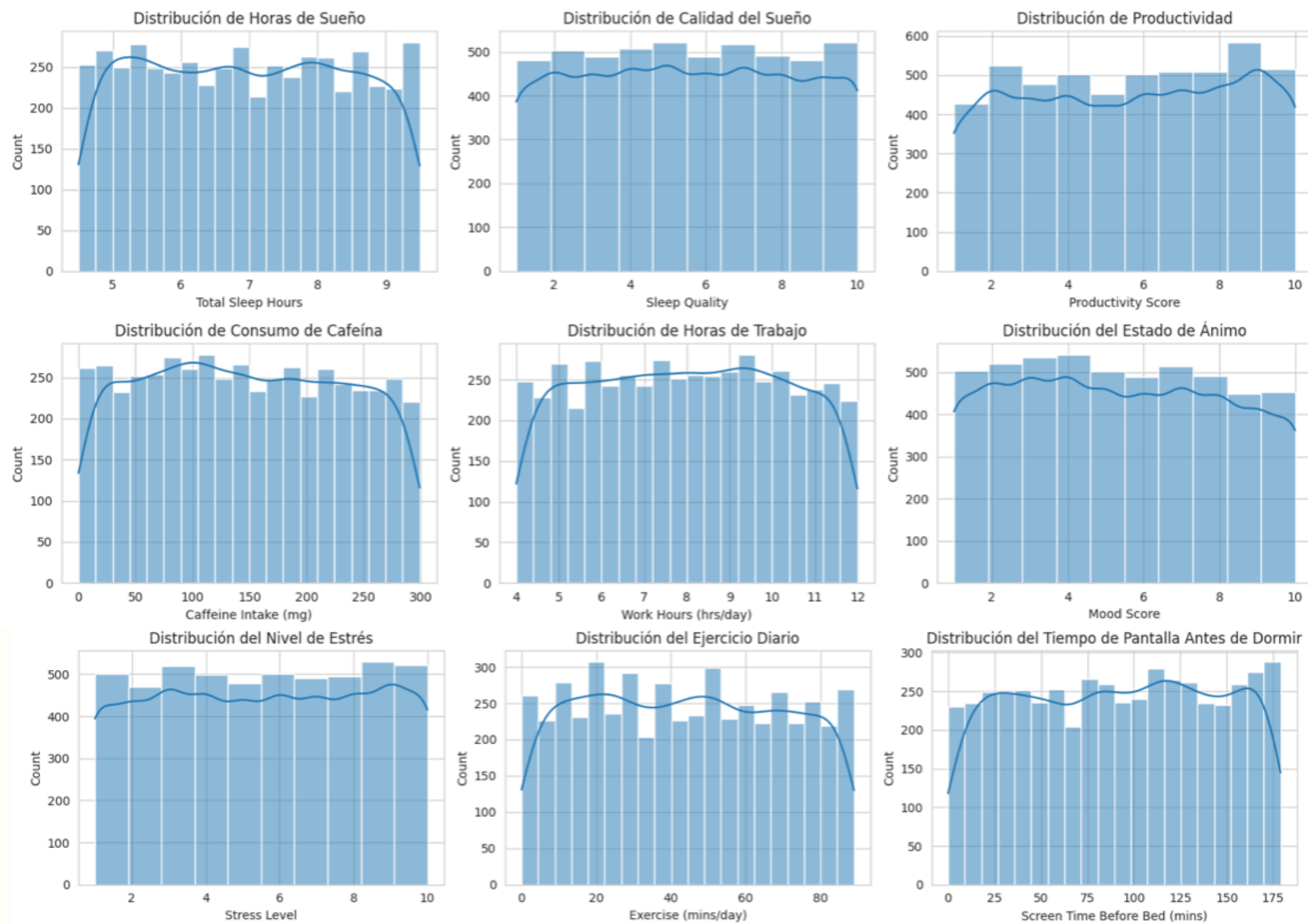


Análisis Exploratorio



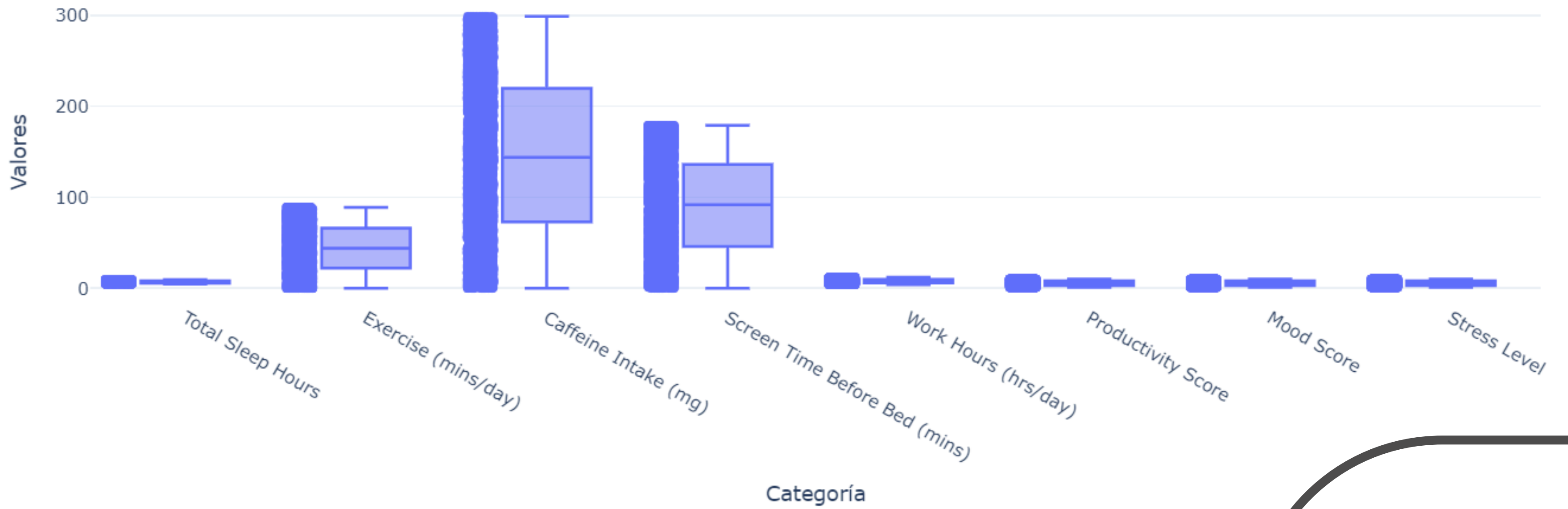


Análisis Exploratorio



➔ Análisis Exploratorio ➔

Boxplot: Sleep Cycle & Productivity

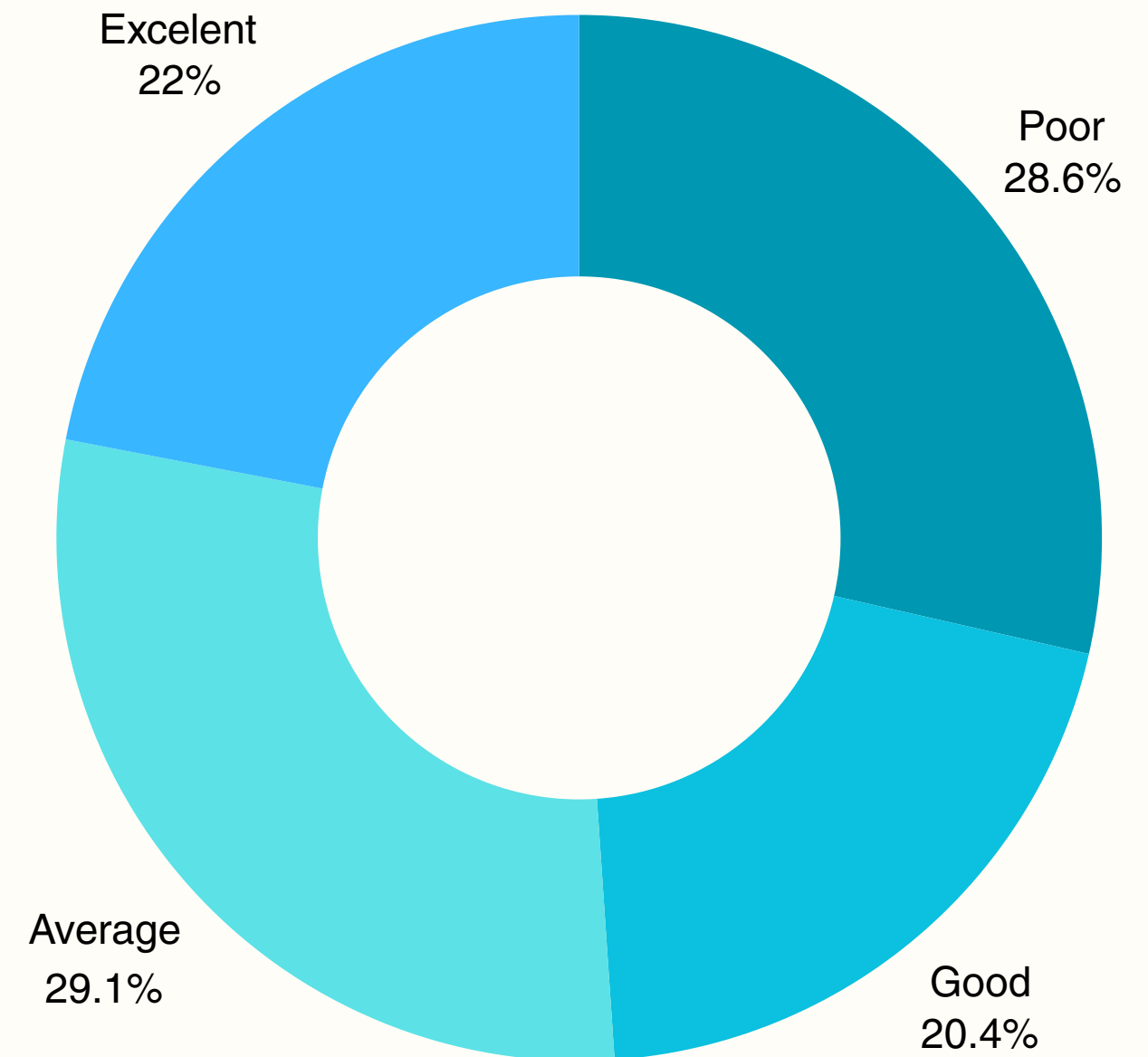


Selección de variables

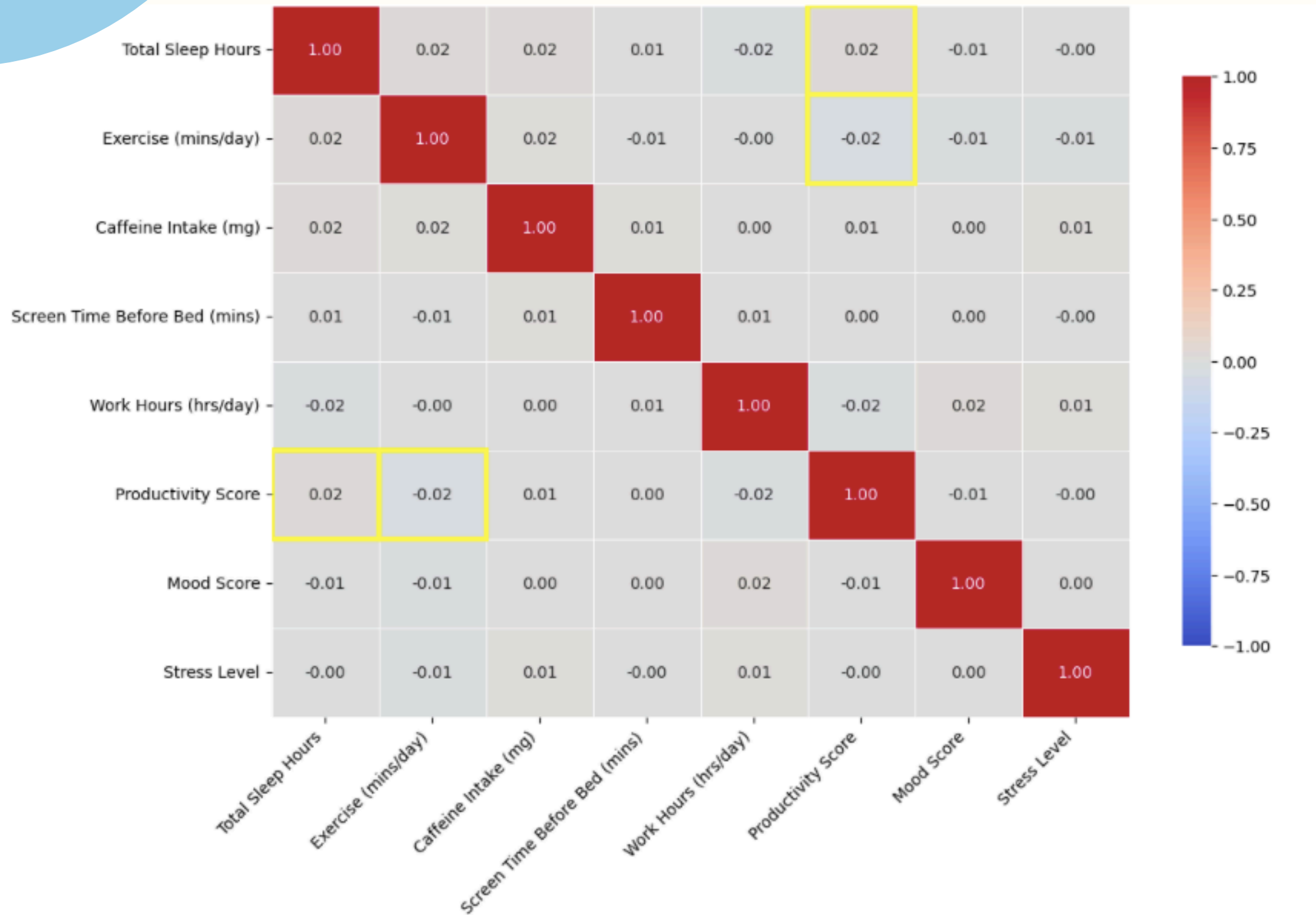
- Revisión Bibliografía
- Matriz de correlación

¿A qué tipo de problema nos enfrentamos?

Distribucion Productivity Score

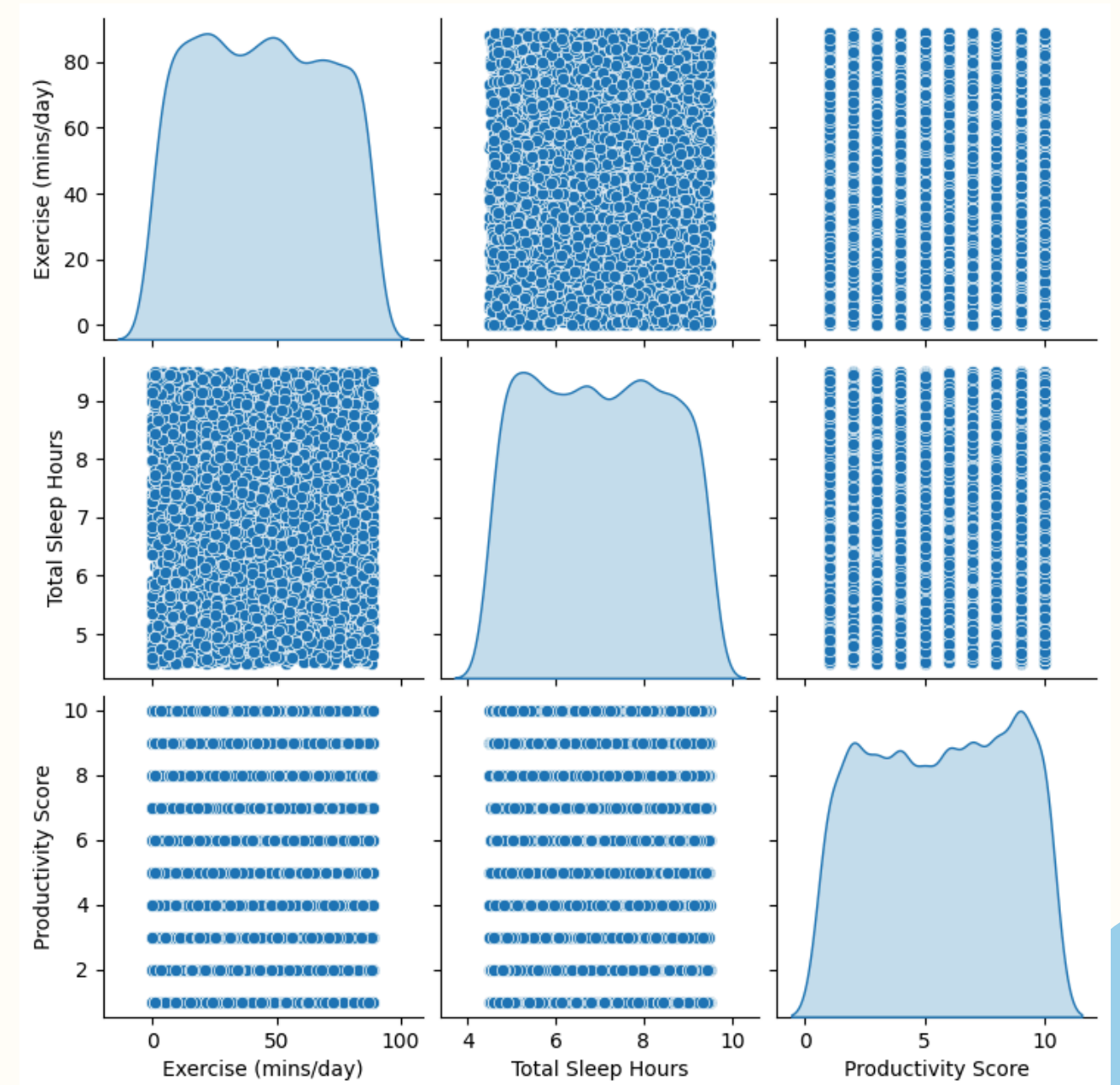
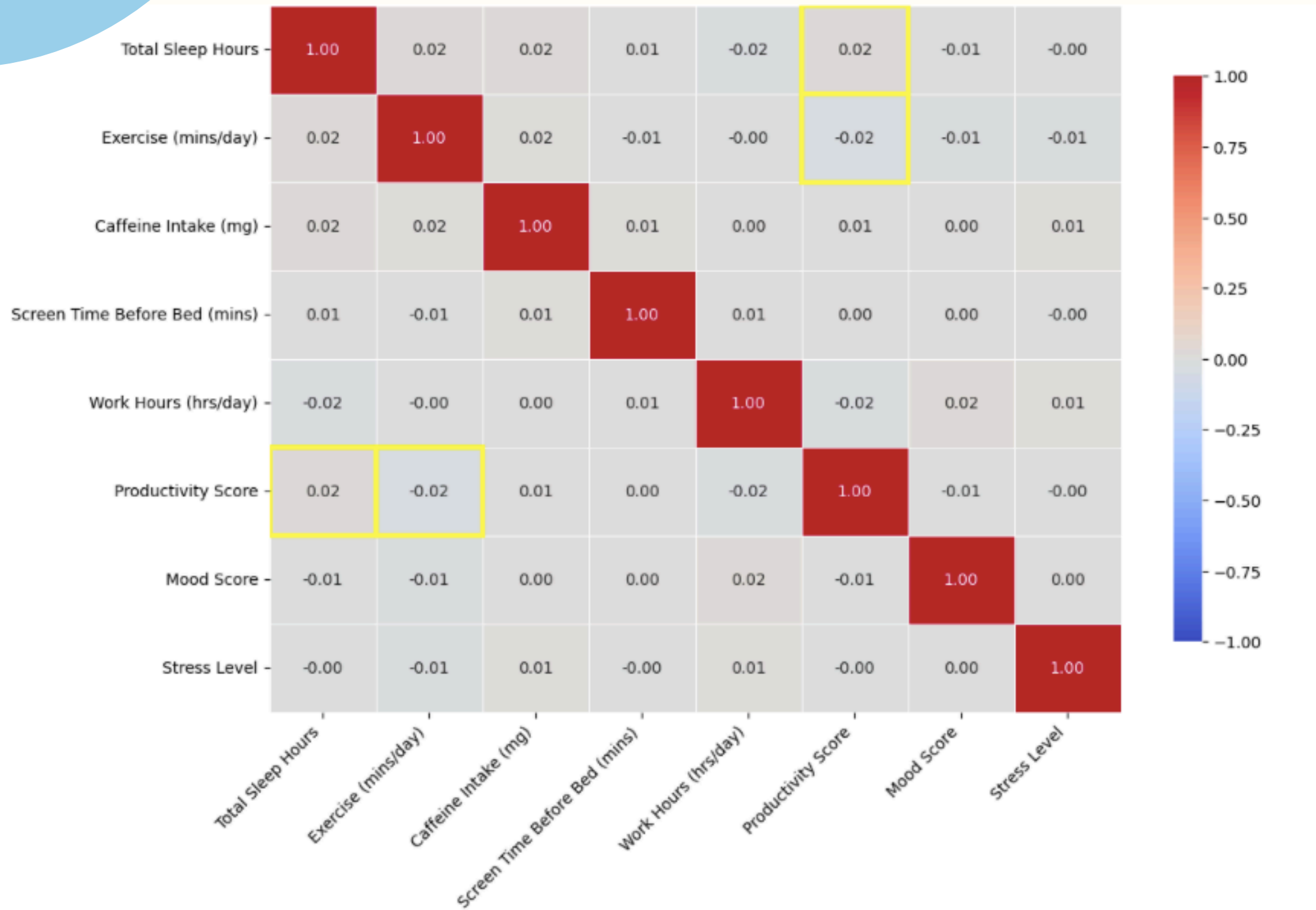


➔ Análisis de Correlación ➔

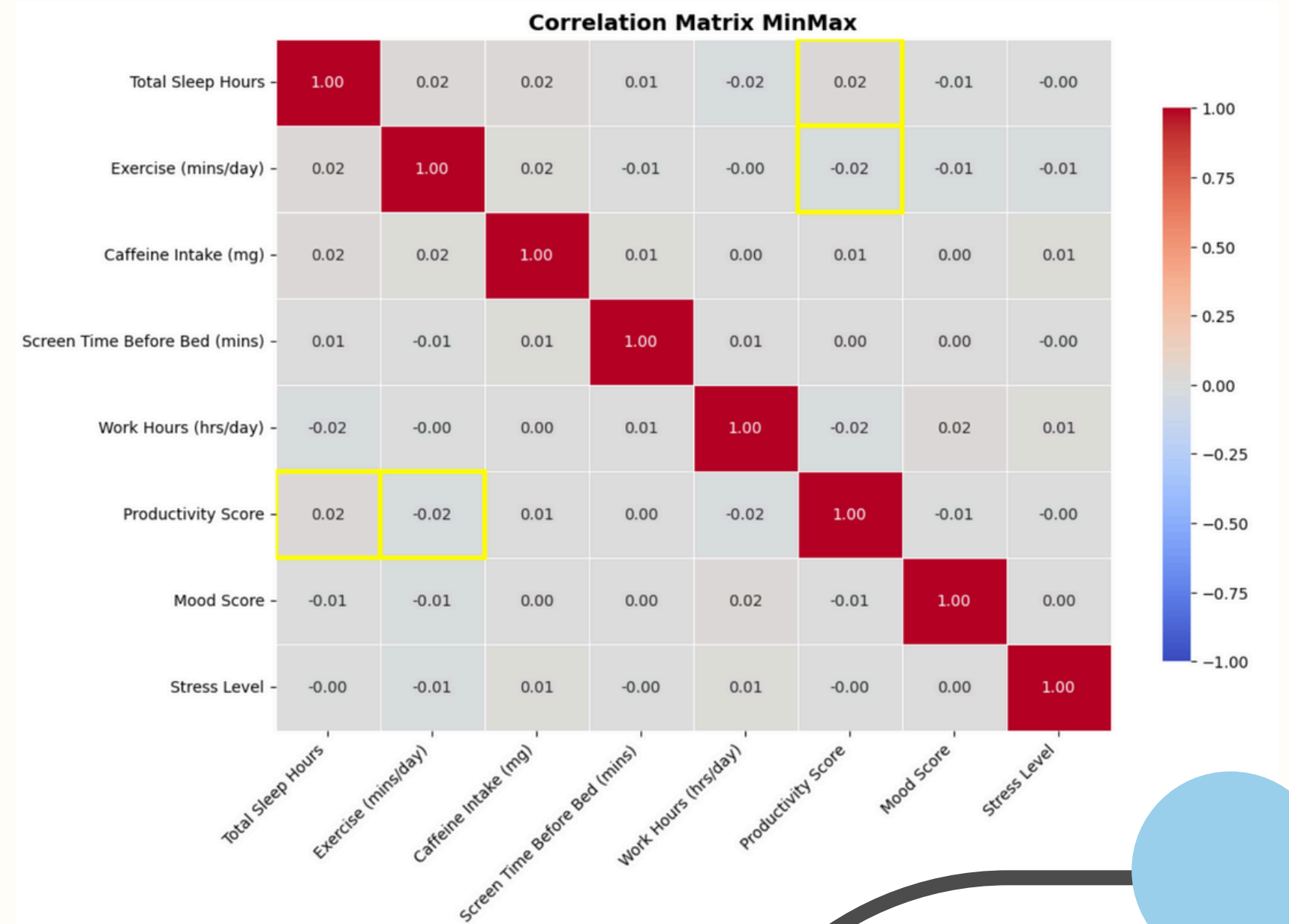
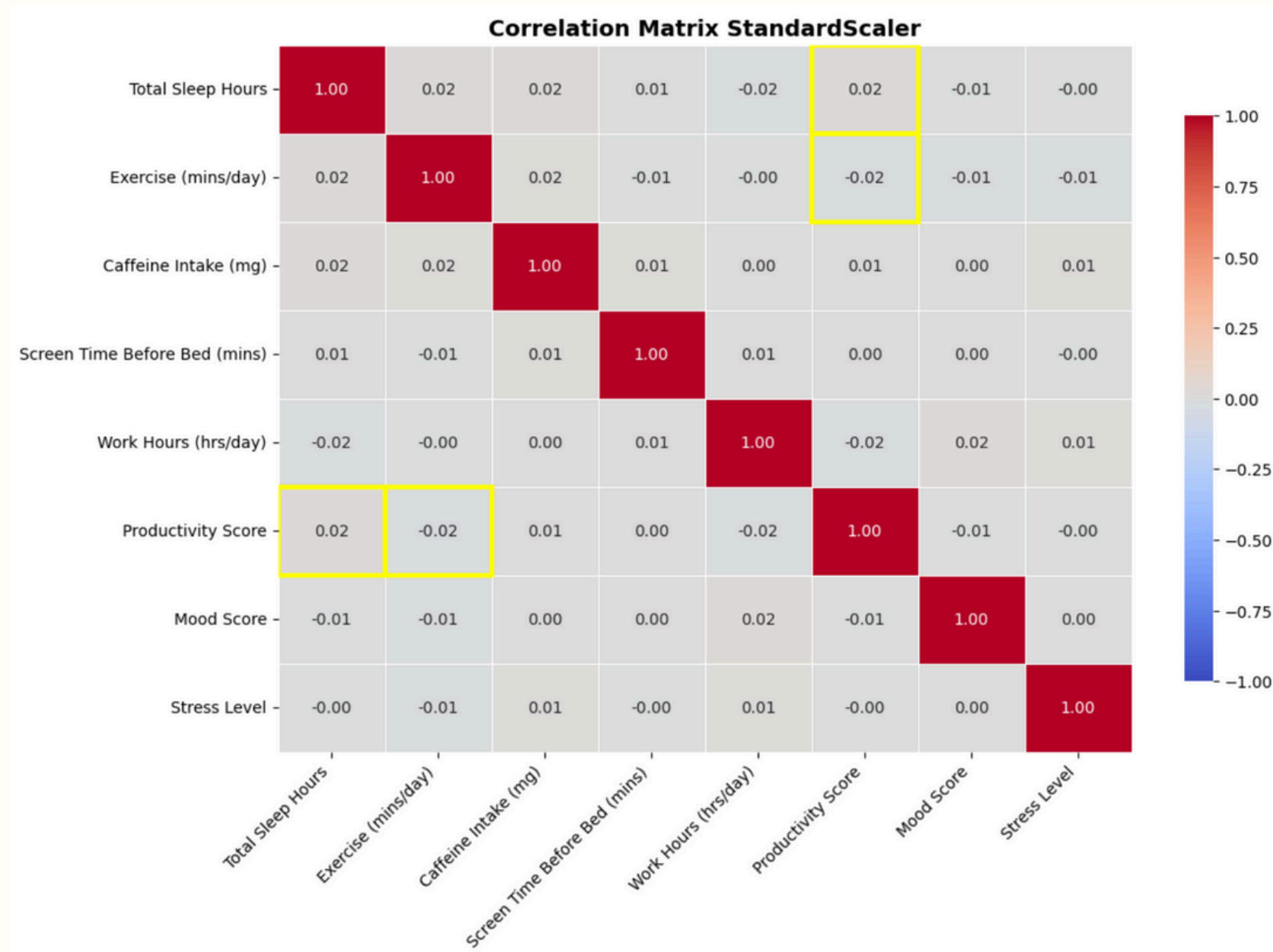


- Cálculo de la matriz de correlación
- Visualización de la matriz de correlación
- Identificación de las dos variables con mayor correlación.

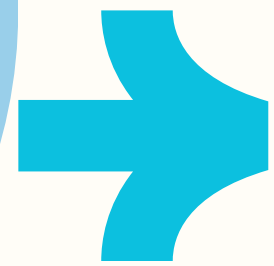
➔ Análisis de Correlación ➔



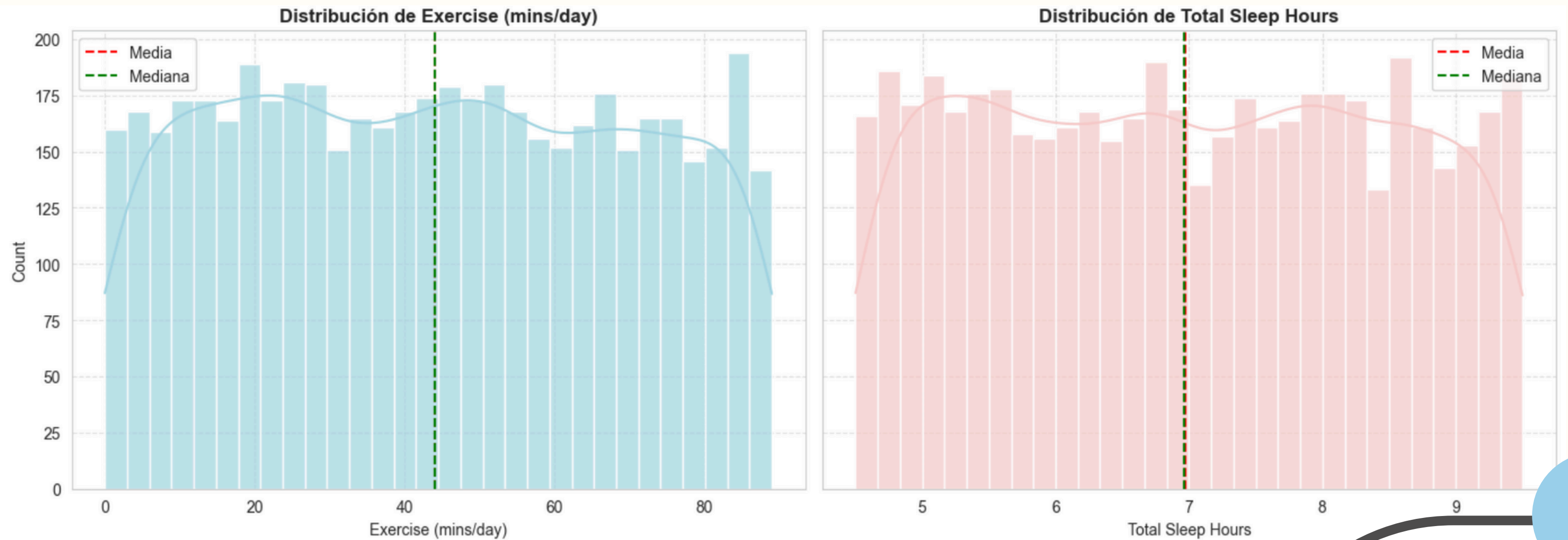
Normalización y reevaluación de Correlación



No hay cambios en la matriz de correlación.

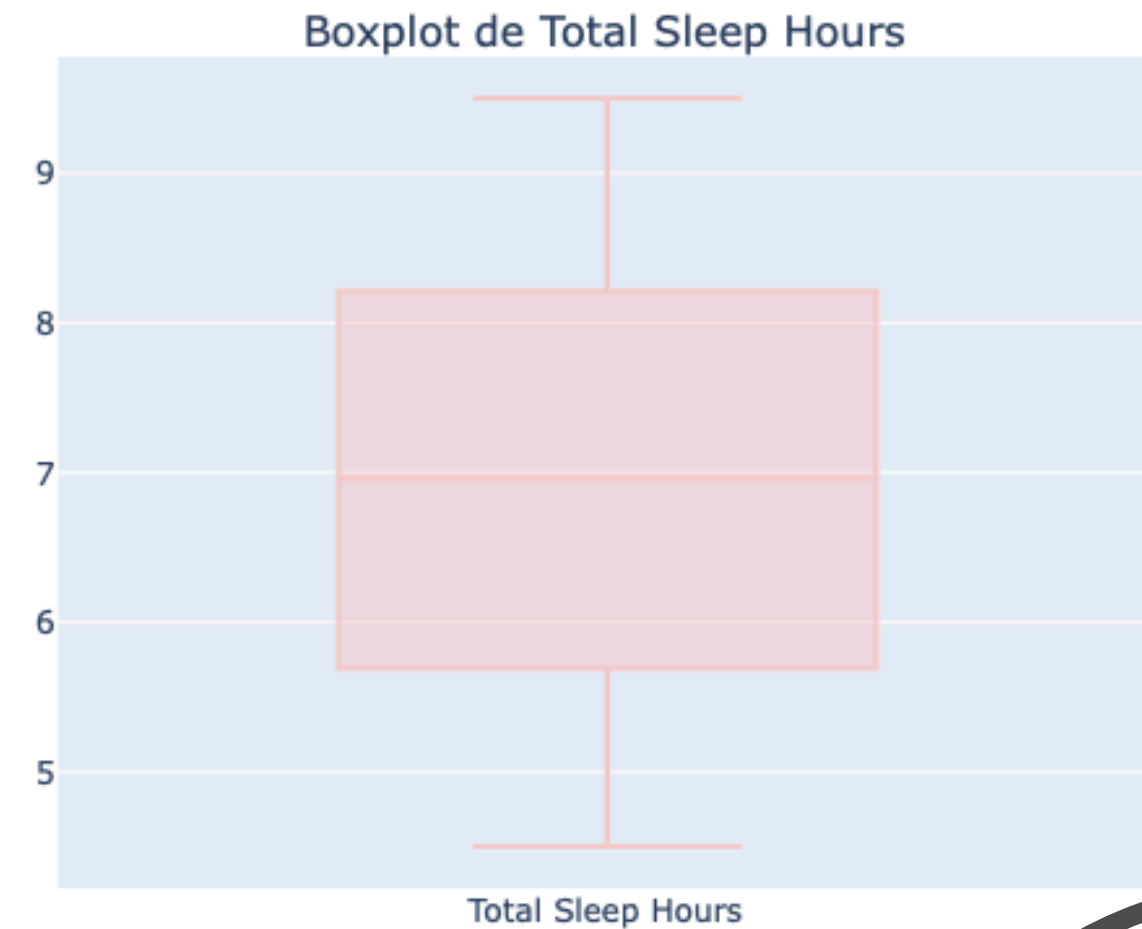
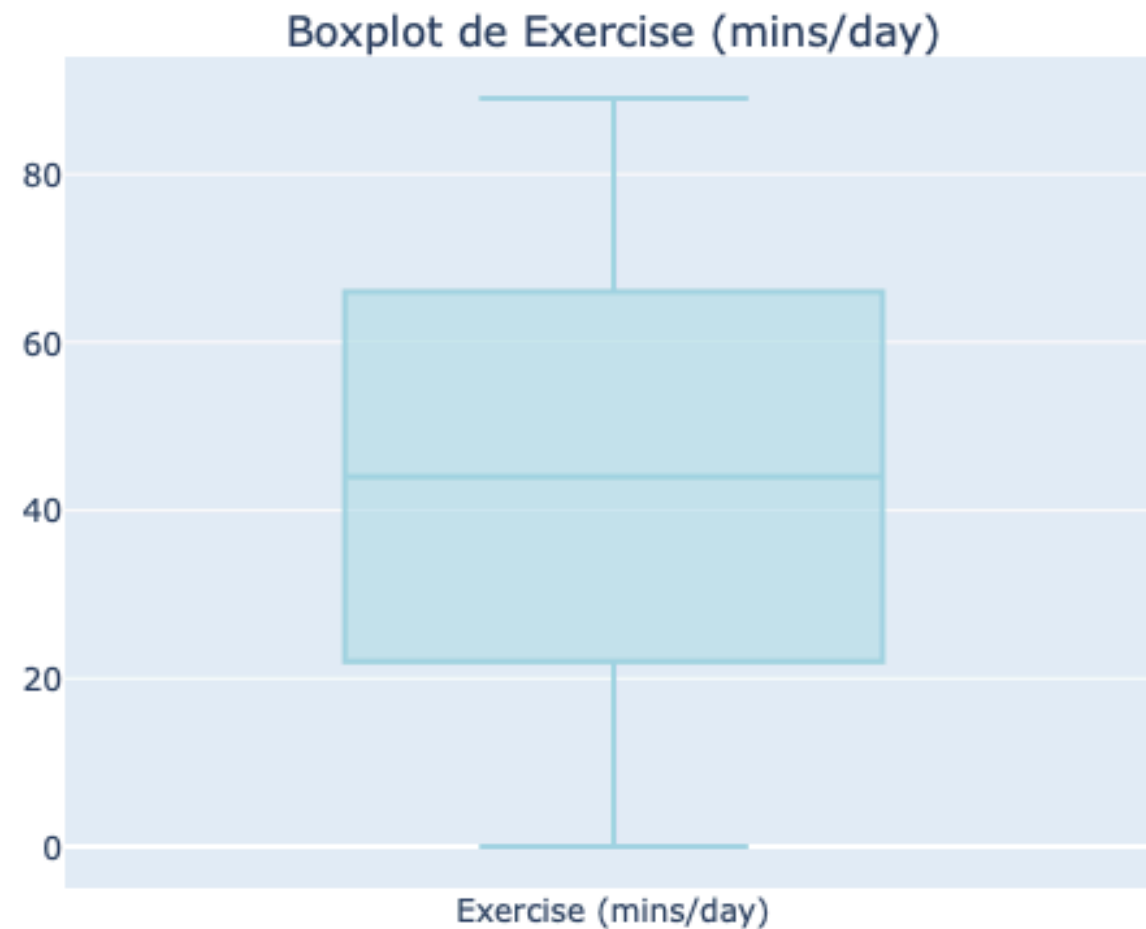


Análisis Estadística Descriptiva



Análisis Estadística Descriptiva

Boxplots de Variables Seleccionadas





Análisis Estadística

Descriptiva



	count	mean	std	min	25%	50%	75%
Exercise (mins/day)	5000.0	43.962600	25.798541	0.0	22.00	44.00	66.00
Total Sleep Hours	5000.0	6.974902	1.454033	4.5	5.69	6.96	8.21

	max	mode	variance	skewness	kurtosis
Exercise (mins/day)	89.0	86.00	665.564714	0.036178	-1.187155
Total Sleep Hours	9.5	8.02	2.114211	0.021970	-1.210223

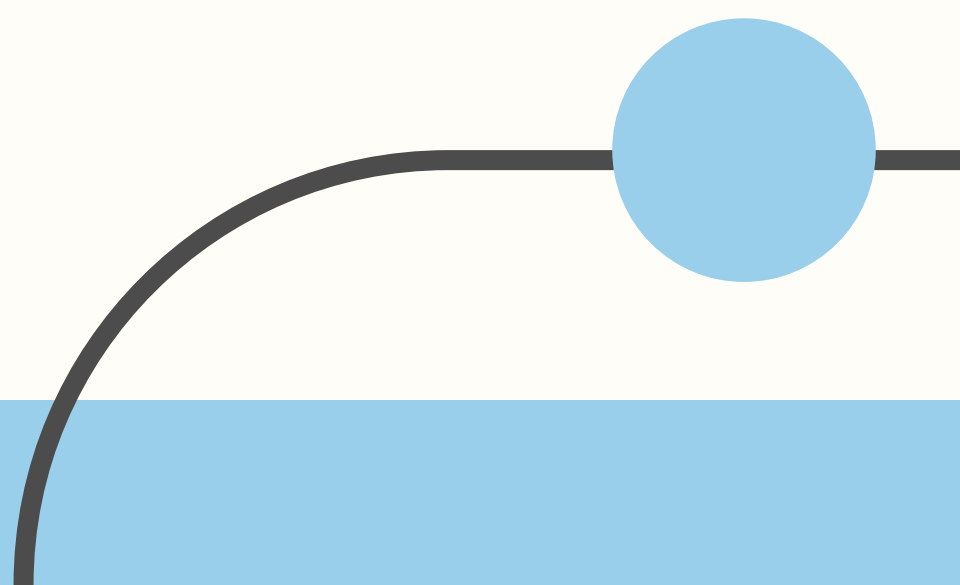


➔ Selección de Variables con ➔ Regresión Logística y ElasticNet

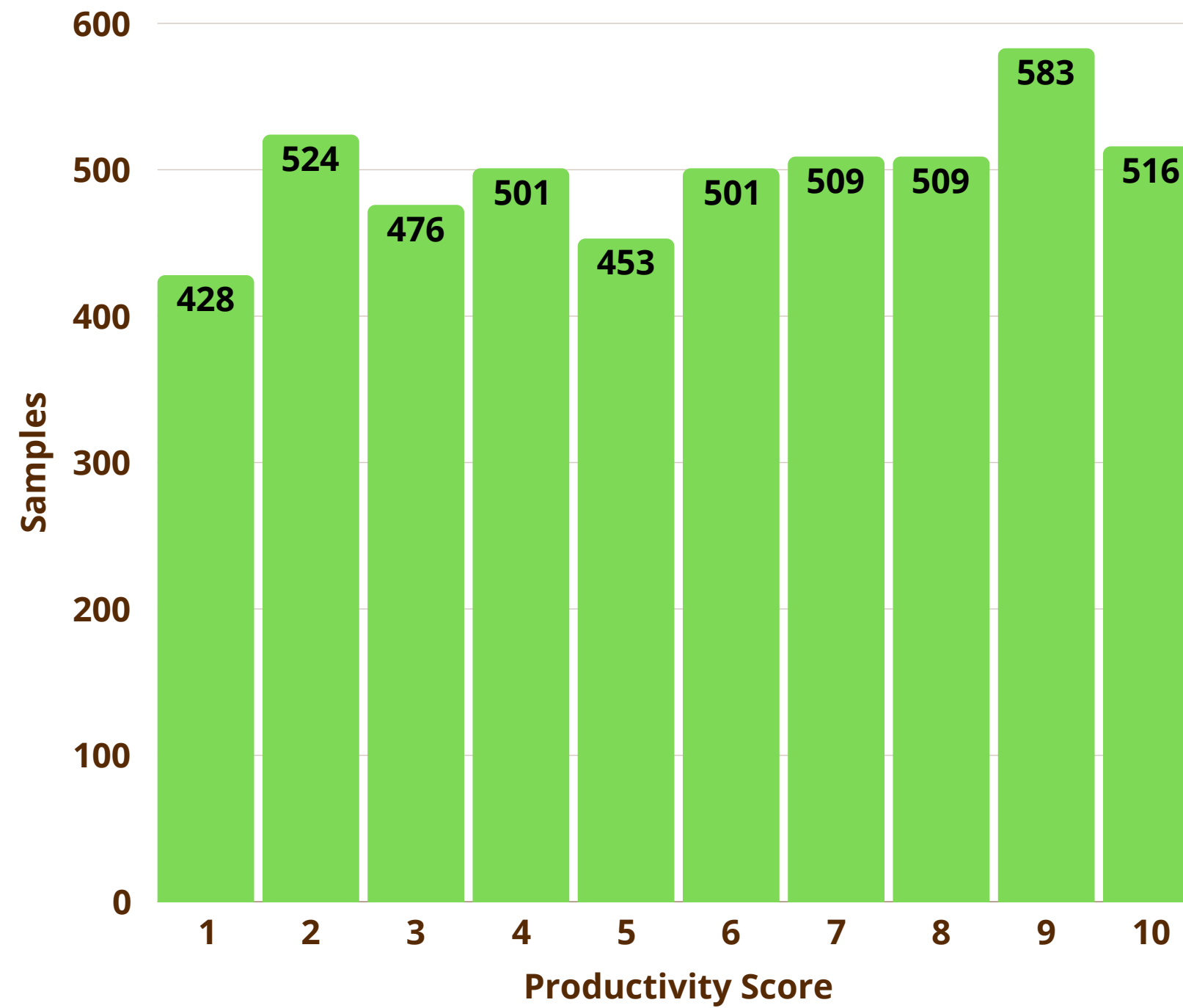
Planteamiento Inicial:

- Se seleccionaron dos variables con respecto a la correlación (Exercise, Total Sleep Hours)

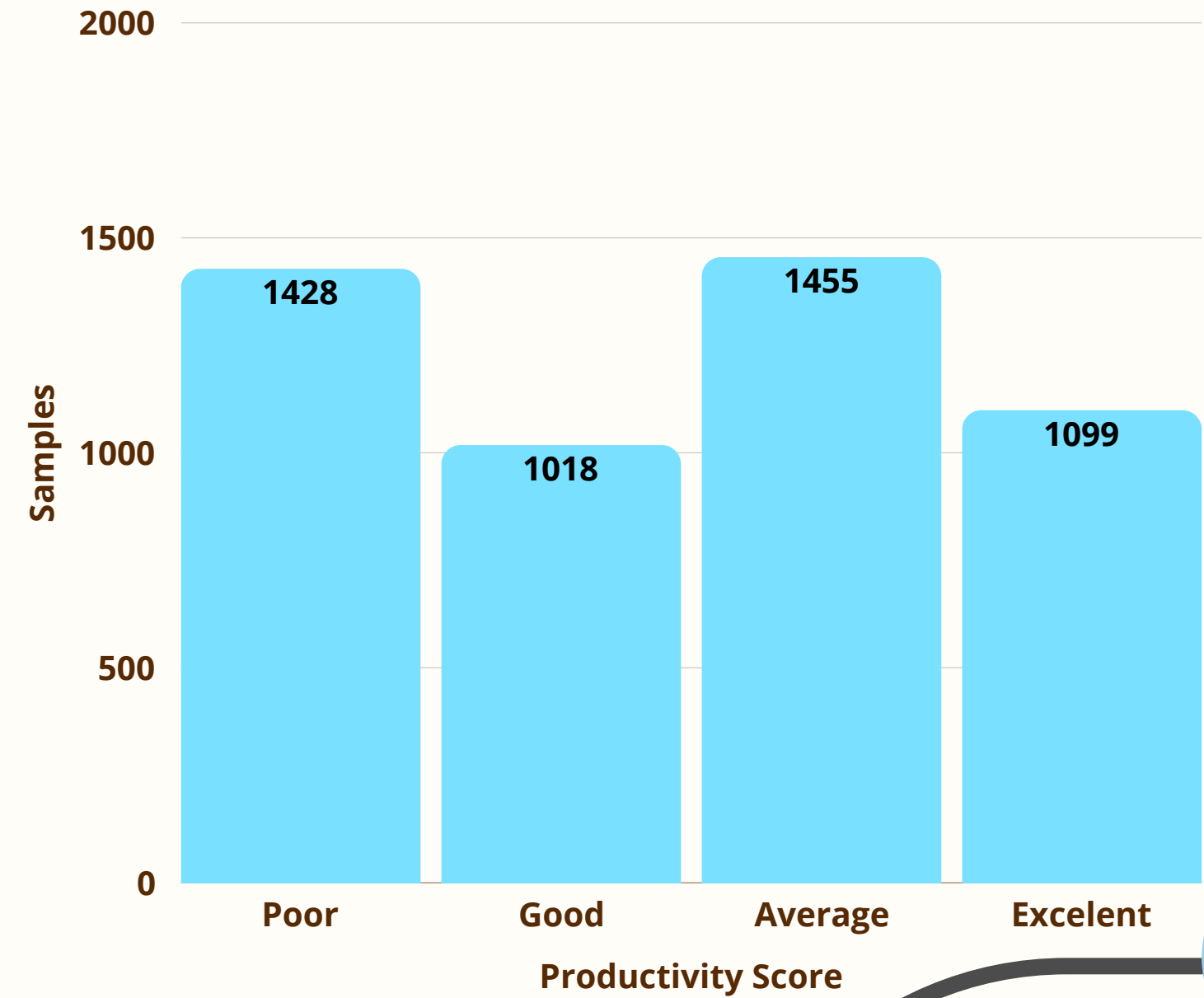
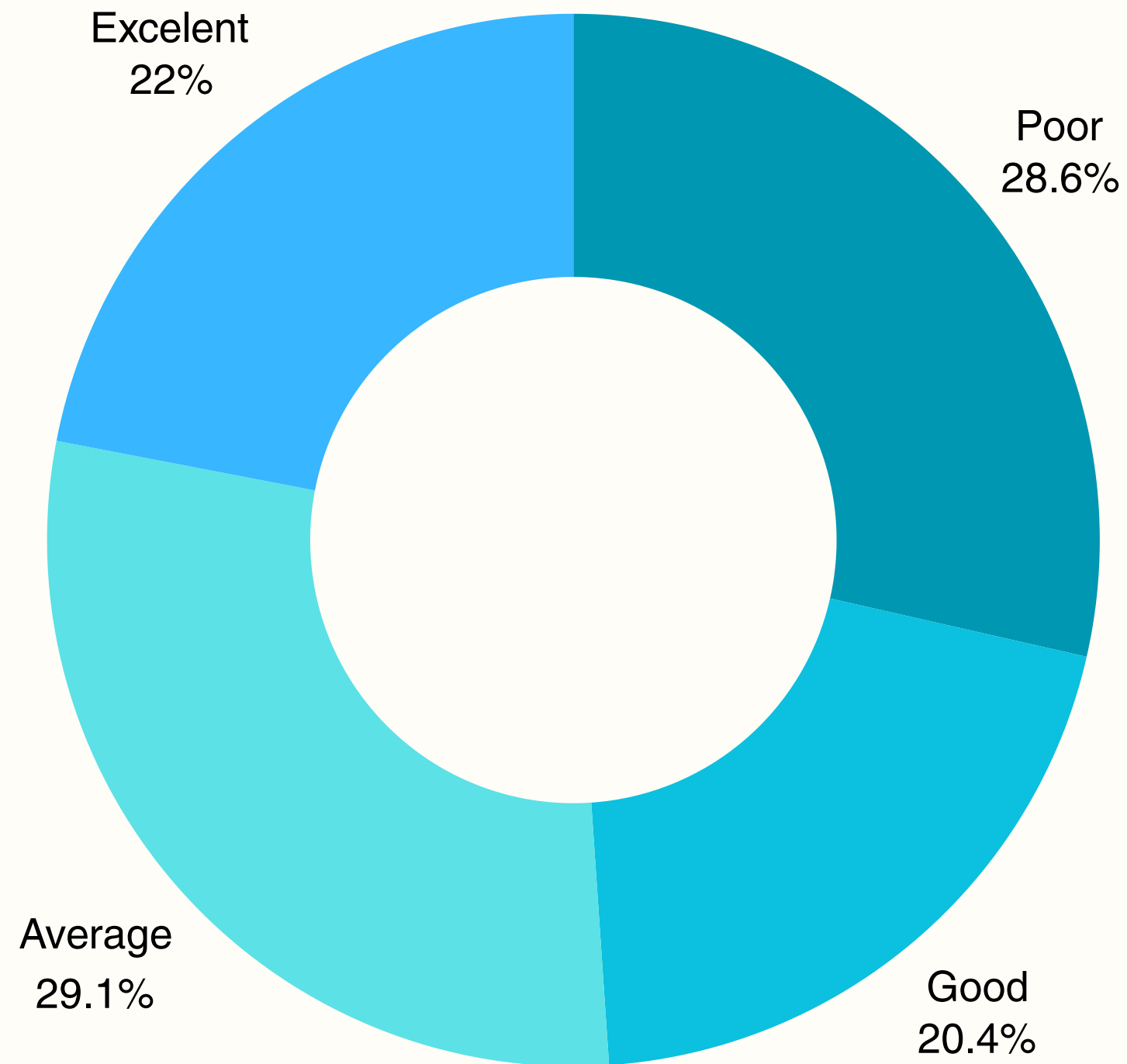
Resultados:

- El modelo nos entrega las variables más relevantes después del proceso de normalización con elasticNet ['Mood Score', 'Exercise (mins/day)', 'Total Sleep Hours', 'Screen Time Before Bed (mins)']
 - Reportando un **Accuracy** de alrededor del 32%
- 

Variable Objective Productivity score



Discretización Variable Objetivo



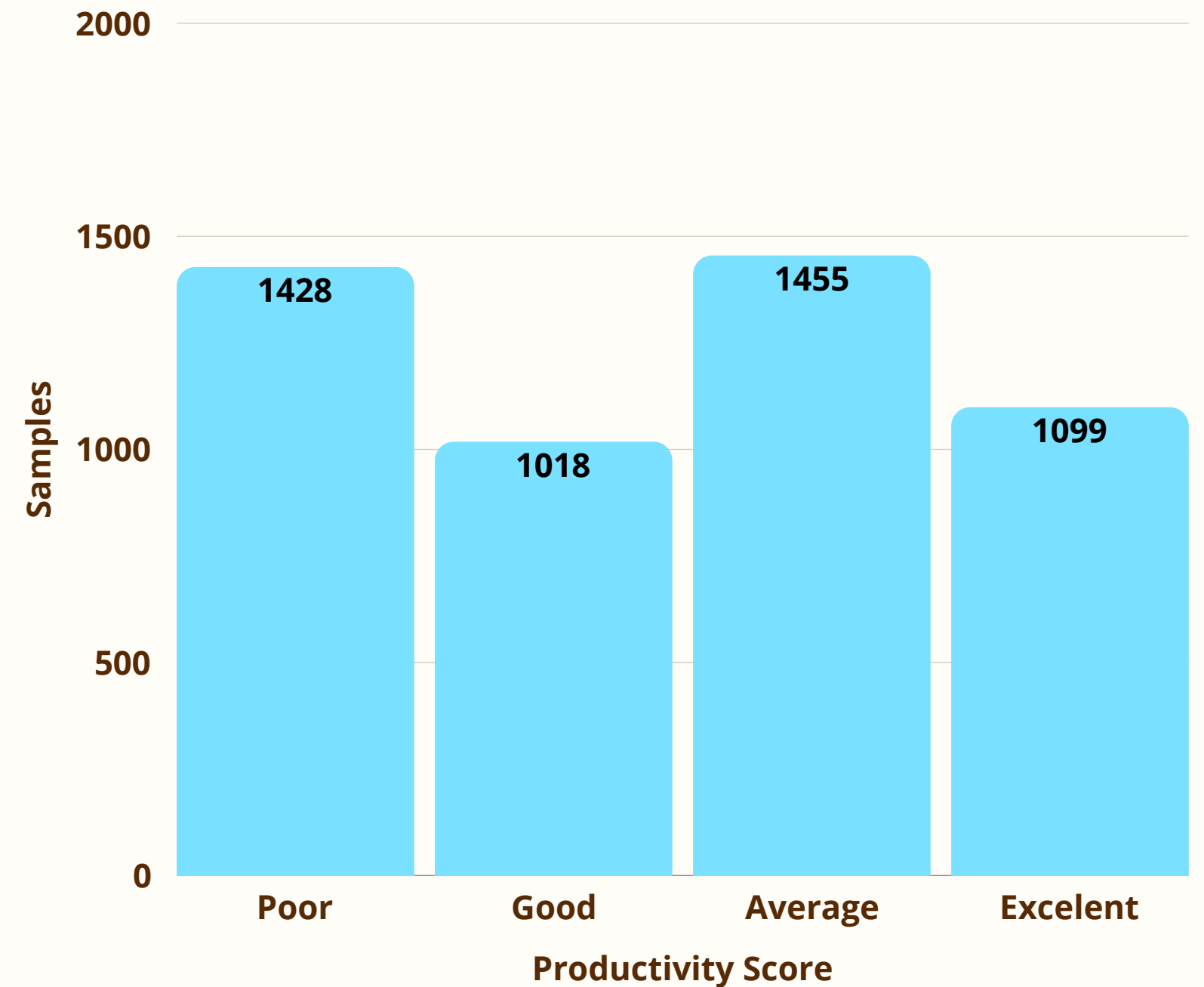
Resultados Modelo Inicial

Metodología:

- Implementación de Regresión Logística (ElasticNet)
- Definición de 4 categorías en la variable objetivo
- Normalización de los datos

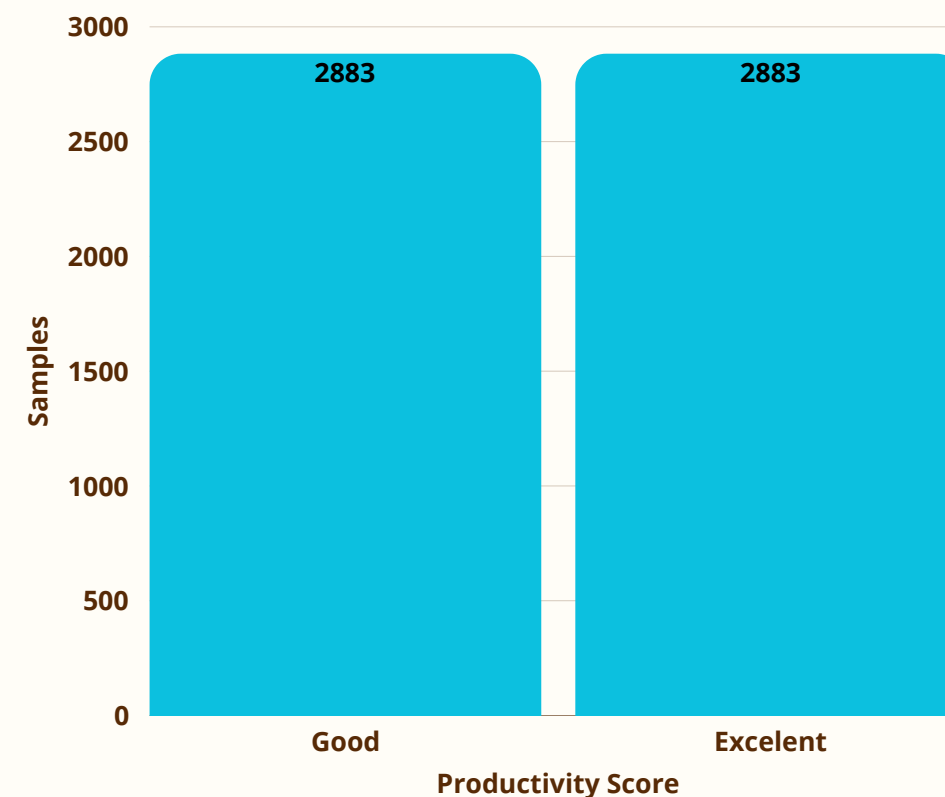
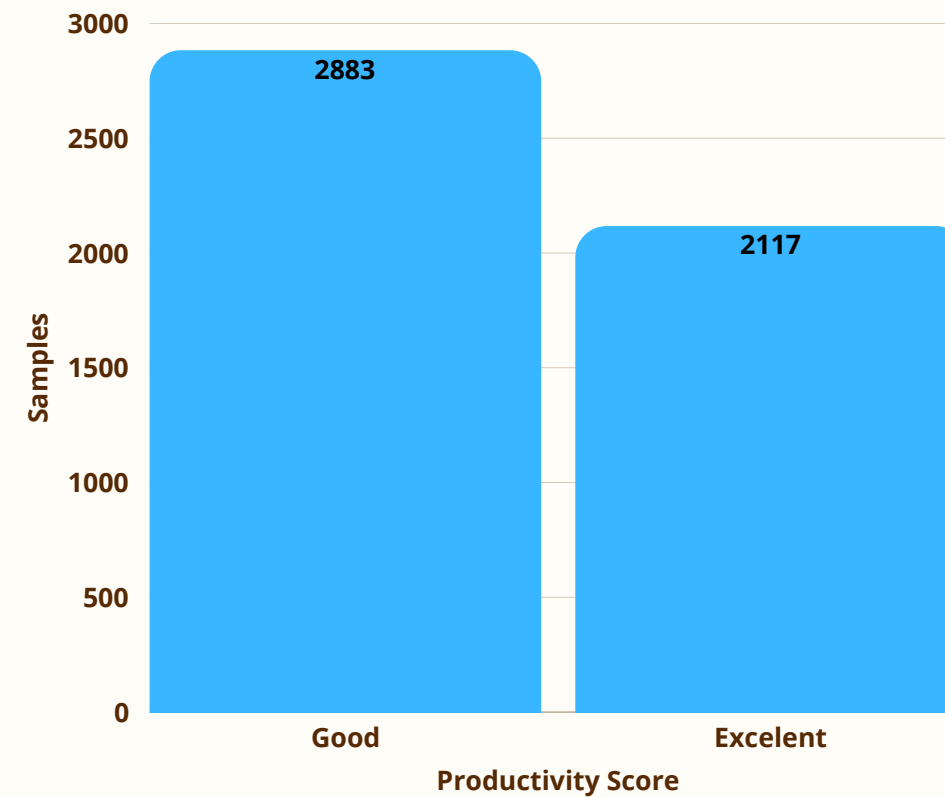
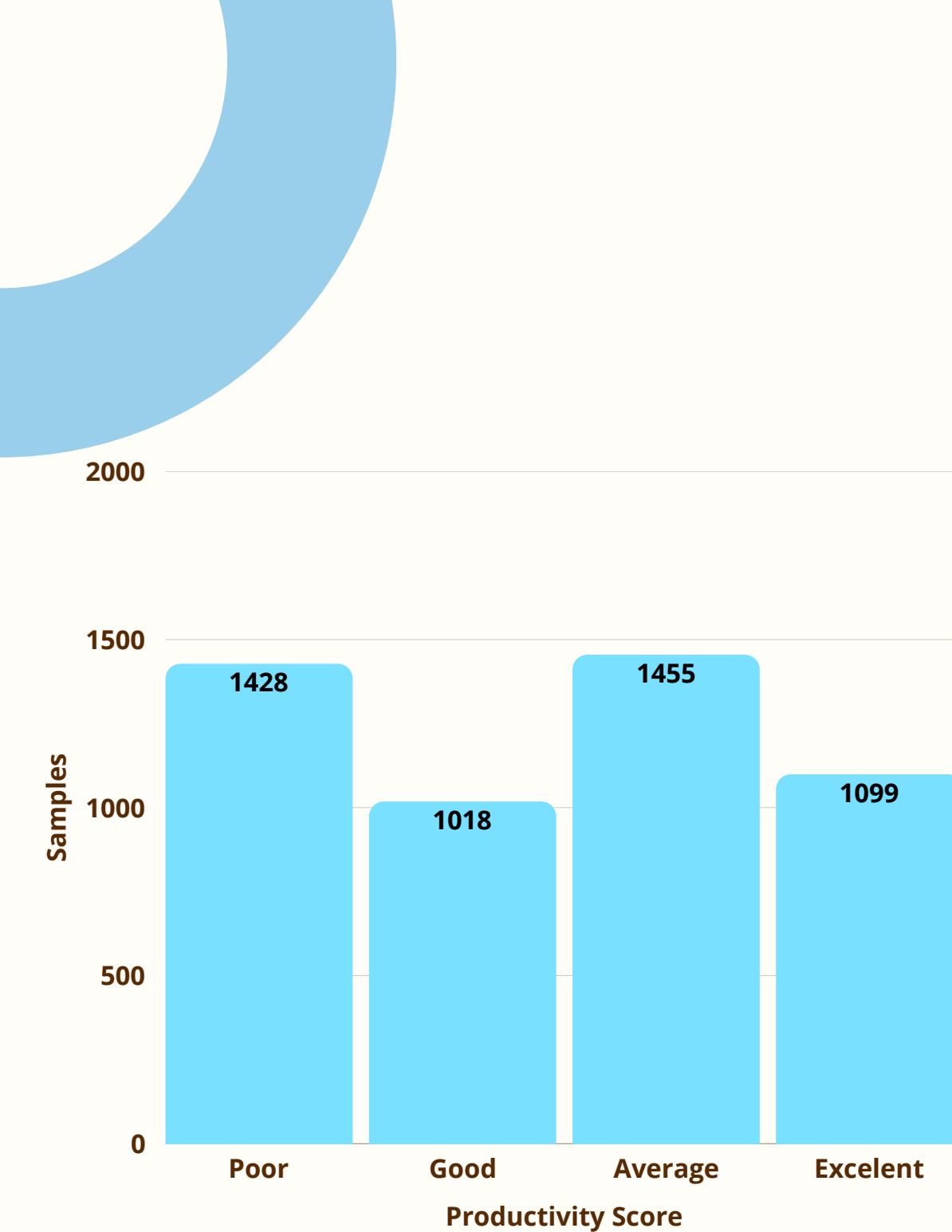
Resultados:

- Accuracy: 32%
- Variables más relevantes: ['Mood Score', 'Exercise (mins/day)', 'Total Sleep Hours', 'Screen Time Before Bed (mins)']



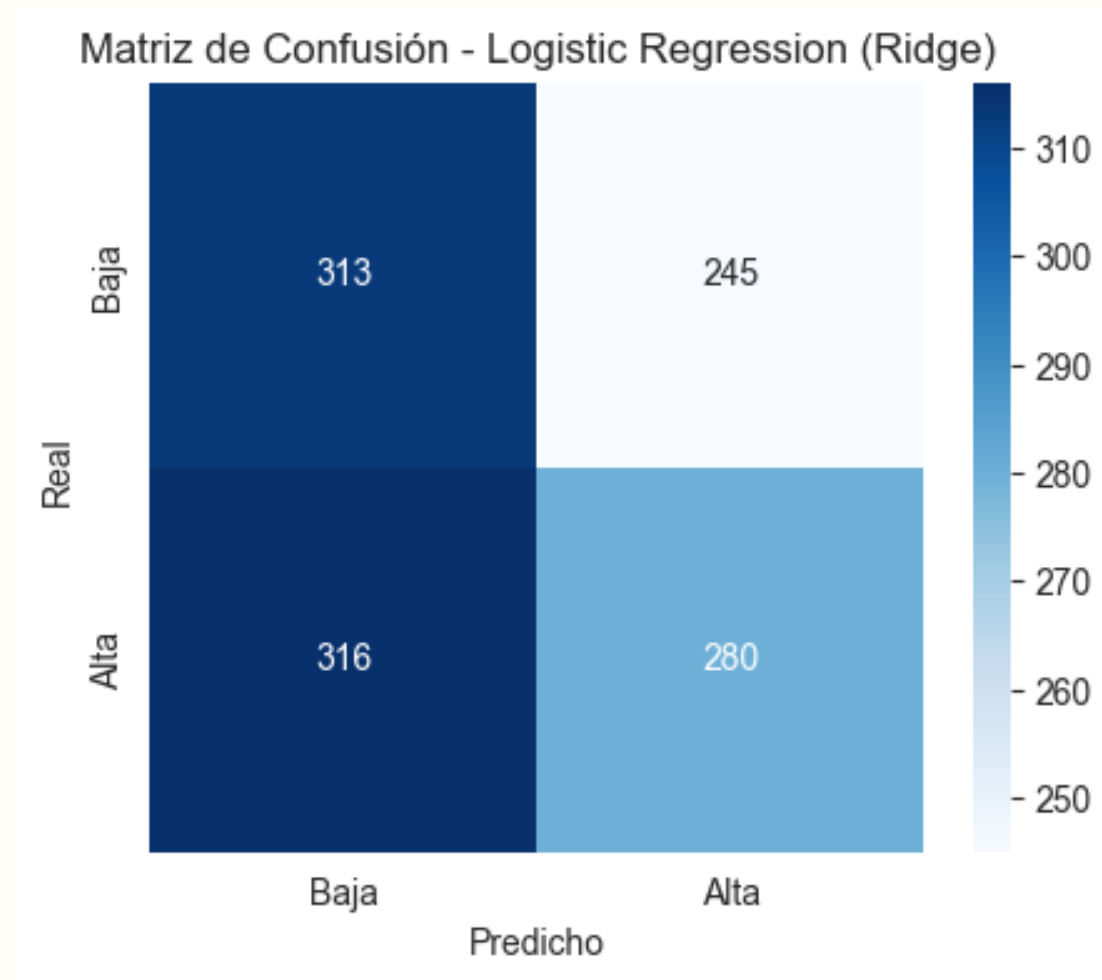
Featuring Engineering

- Validación de múltiples modelos (KNN, LR - Ridge, Random Forest)
- Balanceo de las clases con respecto a la variable objetivo (SMOTE)
- Agregamos transformaciones de características
- Usamos validación cruzada en el proceso de entrenamiento
- Reducimos la complejidad del modelo (cambiando las categorías)

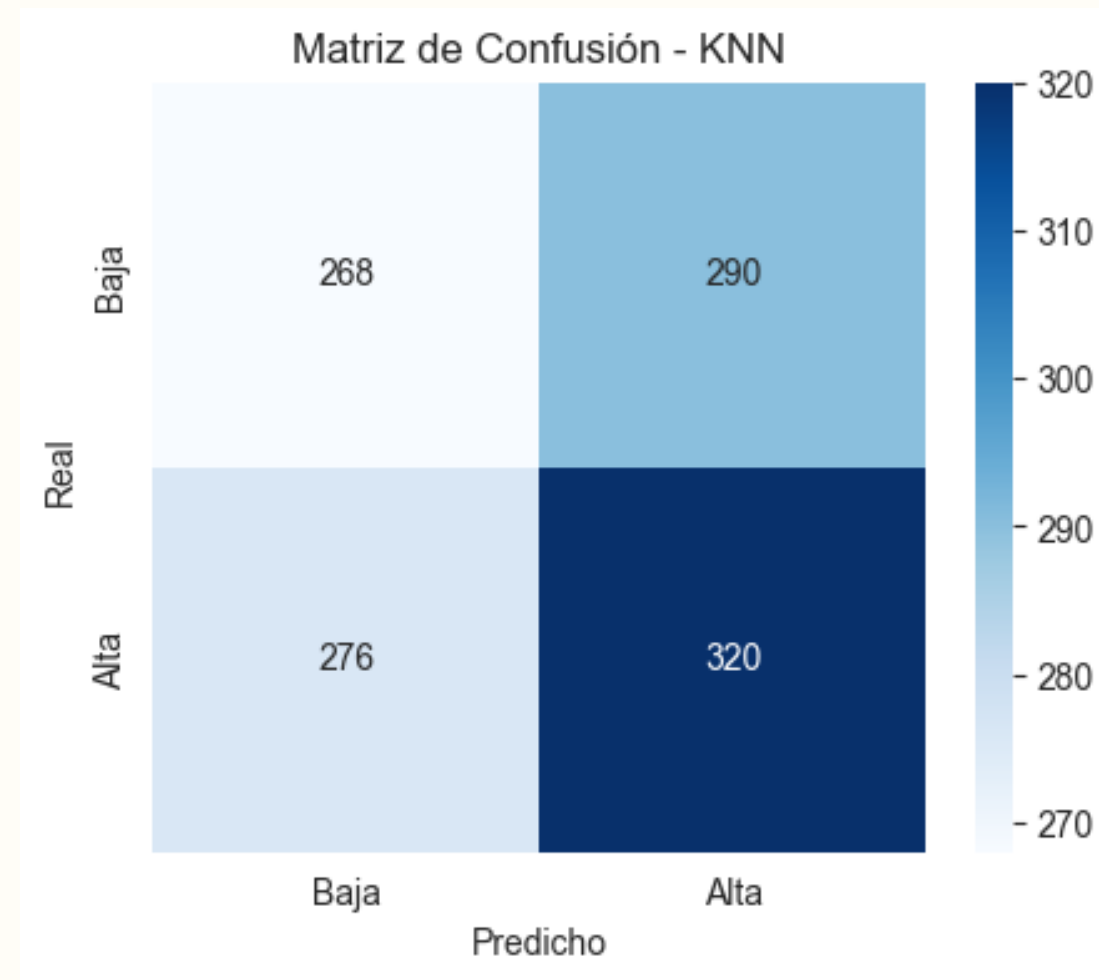




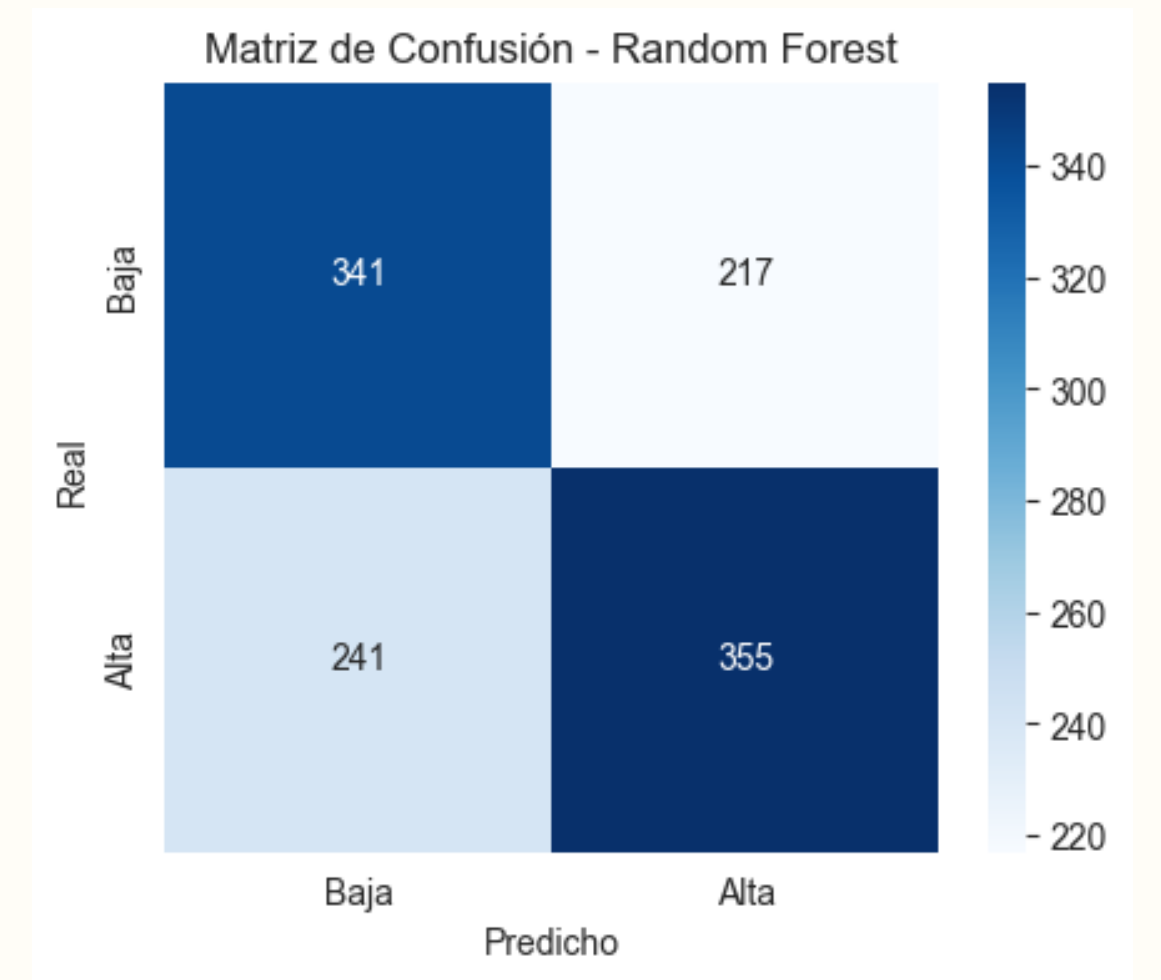
Featuring Engineering



Accuracy 51.4%



Accuracy 52%



Accuracy 60.3%

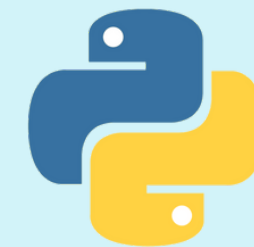


Pipeline



Notebook

Conversion



**Python
Functions**



Input



Function



Function



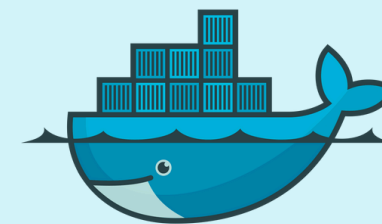
Function



Output



Apache
Airflow



docker



Conclusiones



- La **correlación no** siempre indica las mejores características.
 - La **normalización no** afecta el análisis realizado.
 - El uso de **técnicas** de **regularización** es clave para la **selección** de variables.
 - La **validación cruzada** ayuda a **evitar** el sobre-ajuste.
 - El pre-procesamiento influye en el desempeño del modelo.
 - El desempeño del modelo aumenta conforme balanceamos las clases pertenecientes a la variable objetivo
 - Disminuir la complejidad del modelo nos permitio tener un mejor desempeño
- 