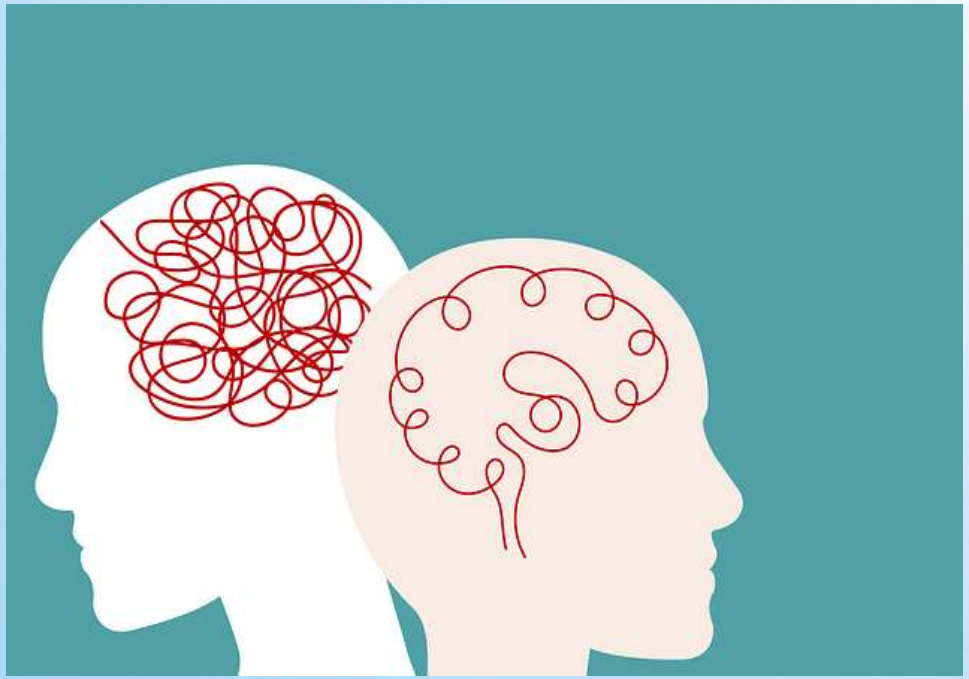
CODERHOUSE

PROYECTO FINAL
ANÁLISIS PREDICTIVO DE ANSIEDAD SOCIAL



- ***ESTUDIANTE:*** ALFREDO JASAUI CHERO
- ***CURSO:*** DATA SCIENCE II
- ***COMISIÓN:*** 67485
- ***PROFESOR:*** GUSTAVO BENITEZ
- ***TUTOR:*** GUILLERMO MALLO

CONTENIDOS

1. ***Presentación***
2. ***Lectura de Datos***
3. ***Data Wrangling*** – Limpieza & Transformación de Datos
4. ***Análisis Exploratorio de Datos*** - EDA
5. ***Feature Engineering*** – Ingeniería de Atributos
6. ***Feature Selection***
7. ***Modelado***
8. ***Optimización de Modelos***
9. ***Conclusiones Finales***

PRESENTACIÓN

MOTIVACIÓN

- ✓LA ANSIEDAD SOCIAL ES UN TRASTORNO COMÚN QUE LIMITA LA VIDA PERSONAL, ACADÉMICA Y PROFESIONAL
- ✓SU DETECCIÓN TEMPRANA ES CLAVE PARA PREVENIR COMPLICACIONES EMOCIONALES Y SOCIALES.
- ✓EL ANÁLISIS Y MODELADO DE DATOS PERMITE PREDECIR NIVELES DE ANSIEDAD CON MAYOR PRECISIÓN E IDENTIFICAR FACTORES [illegible] CULTOS.

AUDIENCIA

- ✓PROFESIONALES DE SALUD MENTAL
- ✓CLÍNICAS, HOSPITALES Y ASEGURADORAS
- ✓INVESTIGADORES Y ACADÉMICOS
- ✓ÁREAS DE RECURSOS HUMANOS
- ✓DESARROLLADORES DE APPS DE SALUD

PRESENTACIÓN

CONTEXTO ANALÍTICO & COMERCIAL

✓LA DISPONIBILIDAD DE DATA PERMITE ANÁLISIS DESCRIPTIVOS Y PREDICTIVOS

✓ENTRE ELLOS, MODELOS COMO MACHINE LEARNING QUE PERMITEN PREDECIR NIVELES DE ANSIEDAD SOCIAL, FACILITANDO LA TOMA DE DECISIONES CLÍNICAS INFORMADAS Y ANTICIPADAS

✓ESTO GENERA APOYO EN:

- ESTRATEGIAS DE INTERVENCIÓN Y TRATAMIENTOS PERSONALIZADOS
- PRIORIZACIÓN DE CASOS URGENTES
- INVESTIGACIÓN E INNOVACIÓN EN SALUD PÚBLICA Y PRIVADA
- POTENCIAL PARA APPS Y PLATAFORMAS DIGITALES EN SALUD, ASEGURADORAS Y EMPRESAS

PRESENTACIÓN

PREGUNTAS A RESOLVER

1. ¿CÓMO ES LA INFLUENCIA DEL NIVEL DE ESTRÉS SOBRE EL NIVEL DE ANSIEDAD REPORTADO?
2. ¿PUEDE LA CANTIDAD DE HORAS DE SUEÑO PREDECIR EL NIVEL DE ANSIEDAD DE UNA PERSONA?
3. ¿EXISTE RELACIÓN SIGNIFICATIVA ENTRE LA FRECUENCIA CARDIACA Y LOS NIVELES DE ANSIEDAD?
4. ¿QUÉ IMPACTO TIENEN LOS HÁBITOS DE CONSUMO DE CAFEÍNA Y ALCOHOL, ASÍ COMO LA ACTIVIDAD FÍSICA, EN LOS NIVELES DE ANSIEDAD?
5. ¿CUÁL ES EL EFECTO COMBINADO DEL NIVEL DE ESTRÉS Y DE LAS HORAS DE SUEÑO PARA CON LOS NIVELES DE ANSIEDAD?
6. ¿EN QUÉ MEDIDA EL CONSUMO DE ALCOHOL Y LA CALIDAD EN LA DIETA SE ASOCIAN CON LOS NIVELES DE ANSIEDAD?

OBJETIVO

CONSTRUIR UN MODELO DE MACHINE LEARNING CAPAZ DE PREDECIR EL NIVEL DE ANSIEDAD DE UNA PERSONA EN BASE A UN CONJUNTO DE ATRIBUTOS, A TRAVÉS DEL ENTRENAMIENTO DE UN APRENDIZAJE SUPERVISADO. ESTA ES UNA TAREA DE REGRESIÓN, YA QUE LA VARIABLE OBJETIVO 'ANXIETY LEVEL (1-10)' ES NUMÉRICA

LECTURA DE DATOS

CARGA DE LIBRERÍAS

HERRAMIENTAS NECESARIAS PARA LA MANIPULACIÓN ANÁLISIS Y VISUALIZACIÓN DE DATOS; INCLUYENDO FUNCIONES ESTADÍSTICAS MATEMÁTICA AVANZADA MODELADO Y APRENDIZAJE AUTOMÁTICO.

LECTURA DEL DATASET

DATASET DESDE LA URL DE ORIGEN

	Age	Gender	Sleep Hours	Physical Activity (hrs/week)	Caffeine Intake (mg/day)	Alcohol Consumption (drinks/week)	Smoking	Family History of Anxiety	Stress Level (1-10)	Heart Rate (bpm)	Breathing Rate (breaths/min)	Sweating Level (1-5)	Dizziness	Medication	Therapy Sessions (per month)	Recent Major Life Event	Diet Quality (1-10)	Anxiety Level (1-10)
0	29	Female	6.0	2.7	181	10	Yes	No	10	114	14	4	No	Yes	3	Yes	7	5
1	46	Other	6.2	5.7	200	8	Yes	Yes	1	62	23	2	Yes	No	2	No	8	3
2	64	Male	5.0	3.7	117	4	No	Yes	1	91	28	3	No	No	1	Yes	1	1
3	20	Female	5.8	2.8	360	6	Yes	No	4	86	17	3	No	No	0	No	1	2
4	49	Female	8.2	2.3	247	4	Yes	No	1	98	19	4	Yes	Yes	1	No	3	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10995	23	Female	6.1	3.1	566	9	Yes	No	8	91	28	1	Yes	Yes	1	No	3	6
10996	50	Other	6.6	3.6	64	17	Yes	No	7	95	17	3	No	No	2	No	7	3
10997	29	Male	6.7	6.9	159	14	No	No	8	72	16	1	Yes	Yes	2	Yes	7	4
10998	53	Other	5.7	2.7	248	8	No	No	4	112	28	3	Yes	Yes	1	Yes	2	4
10999	56	Other	6.1	1.1	205	11	No	No	1	66	13	3	No	No	2	Yes	8	2

11000 rows × 18 columns

LECTURA DE DATOS (2)

ANÁLISIS INICIAL DEL DATASET (1)

Dimensiones

```
Filas, columnas:  (11000, 18)
```

Nombres y tipos de datos de las variables, donde:

#int64: Numérico (Entero)
#float64: Numérico (Decimal)
#object: Categórico (Valor no numérico)

```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Age                             11000 non-null  int64
 1   Gender                          11000 non-null  object
 2   Sleep Hours                     11000 non-null  float64
 3   Physical Activity (hrs/week)    11000 non-null  float64
 4   Caffeine Intake (mg/day)        11000 non-null  int64
 5   Alcohol Consumption (drinks/week)  11000 non-null  int64
 6   Smoking                         11000 non-null  object
 7   Family History of Anxiety       11000 non-null  object
 8   Stress Level (1-10)             11000 non-null  int64
 9   Heart Rate (bpm)                11000 non-null  int64
 10  Breathing Rate (breaths/min)    11000 non-null  int64
 11  Sweating Level (1-5)            11000 non-null  int64
 12  Dizziness                       11000 non-null  object
 13  Medication                      11000 non-null  object
 14  Therapy Sessions (per month)    11000 non-null  int64
 15  Recent Major Life Event         11000 non-null  object
 16  Diet Quality (1-10)             11000 non-null  int64
 17  Anxiety Level (1-10)            11000 non-null  int64
dtypes: float64(2), int64(10), object(6)
```

Descripción de Variables

- **Age:** Edad
- **Gender:** Género (masculino, femenino u otro)
- **Sleep hours:** Horas de sueño diarias
- **Physical activity (hrs/week):** Actividad física (en horas por semana)
- **Caffeine intake (mg/day):** Ingesta de cafeína (en miligramos diarios)
- **Alcohol consumption (drinks/week):** Consumo de alcohol (en bebidas por semana)
- **Smoking:** Fumador (sí o no)
- **Diet quality (1-10):** Calidad de la dieta (del 1 al 10)
- **Stress level (1-10):** Nivel de estrés (del 1 al 10)
- **Heart rate (bpm):** Frecuencia cardíaca (en latidos por minuto)
- **Breathing rate (breaths/min):** Frecuencia respiratoria (en respiraciones por minuto)
- **Sweating level (1-5):** Nivel de sudoración (del 1 al 5)
- **Dizziness:** Sufre de mareos (sí o no)
- **Family history of anxiety:** Historial de ansiedad en la familia (sí o no)
- **Medication:** Uso de medicación (sí o no)
- **Therapy sessions (per month):** Número de sesiones de terapia (al mes)
- **Recent major life event:** Evento personal importante reciente (sí o no)
- **Anxiety level (1-10):** Nivel de ansiedad (del 1 al 10)

Frecuencia de valores únicos en las Variables Categóricas

```
Frecuencia de valores únicos de Gender
Gender
Female    3730
Male      3657
Other     3613
Name: count, dtype: int64

Frecuencia de valores únicos de Smoking
Smoking
Yes    5779
No     5221
Name: count, dtype: int64

Frecuencia de valores únicos de Family History of Anxiety
Family History of Anxiety
Yes    5847
No     5153
Name: count, dtype: int64

Frecuencia de valores únicos de Dizziness
Dizziness
Yes    5672
No     5328
Name: count, dtype: int64

Frecuencia de valores únicos de Medication
Medication
Yes    5666
No     5334
Name: count, dtype: int64

Frecuencia de valores únicos de Recent Major Life Event
Recent Major Life Event
Yes    5623
No     5377
Name: count, dtype: int64
```

LECTURA DE DATOS (2)

ANÁLISIS INICIAL DEL DATASET (2)

Estadísticas descriptivas de las Variables Numéricas

- **count=**Conteos
- **mean=**Medias
- **std=**Desviaciones estándar
- **min=**Valores mínimos
- **25%=** Primeros cuartiles
- **50%=** Medianas
- **75%=** Terceros cuartiles
- **max=**Valores máximos

	Age	Sleep Hours	Physical Activity (hrs/week)	Caffeine Intake (mg/day)	Alcohol Consumption (drinks/week)	Stress Level (1-10)	Heart Rate (bpm)	Breathing Rate (breaths/min)	Sweating Level (1-5)	Therapy Sessions (per month)	Diet Quality (1-10)	Anxiety Level (1-10)
count	11000.000000	11000.000000	11000.000000	11000.000000	11000.000000	11000.000000	11000.000000	11000.000000	11000.000000	11000.000000	11000.000000	11000.000000
mean	40.241727	6.650691	2.942136	286.090000	9.701636	5.856364	90.916000	20.957545	3.080636	2.427818	5.181818	3.929364
std	13.236140	1.227509	1.827825	144.813157	5.689713	2.927202	17.325721	5.160107	1.398877	2.183106	2.895243	2.122533
min	18.000000	2.300000	0.000000	0.000000	0.000000	1.000000	60.000000	12.000000	1.000000	0.000000	1.000000	1.000000
25%	29.000000	5.900000	1.500000	172.000000	5.000000	3.000000	76.000000	17.000000	2.000000	1.000000	3.000000	2.000000
50%	40.000000	6.700000	2.800000	273.000000	10.000000	6.000000	92.000000	21.000000	3.000000	2.000000	5.000000	4.000000
75%	51.000000	7.500000	4.200000	382.000000	15.000000	8.000000	106.000000	25.000000	4.000000	4.000000	8.000000	5.000000
max	64.000000	11.300000	10.100000	599.000000	19.000000	10.000000	119.000000	29.000000	5.000000	12.000000	10.000000	10.000000

DATA WRANGLING

VALORES DUPLICADOS

```
Filas duplicadas encontradas: 0

No existen filas duplicadas en el dataset.
```

VALORES

	0
Age	0
Gender	0
Sleep Hours	0
Physical Activity (hrs/week)	0
Caffeine Intake (mg/day)	0
Alcohol Consumption (drinks/week)	0
Smoking	0
Family History of Anxiety	0
Stress Level (1-10)	0
Heart Rate (bpm)	0
Breathing Rate (breaths/min)	0
Sweating Level (1-5)	0
Dizziness	0
Medication	0
Therapy Sessions (per month)	0
Recent Major Life Event	0
Diet Quality (1-10)	0
Anxiety Level (1-10)	0

dtype: int64

Resultado: Se observa que no existen valores nulos en el dataset

¿PODRÍA SER CONSIDERADO EL 0 COMO VALOR NULO EN ALGUNA(S) COLUMNA(S)?

	0
Therapy Sessions (per month)	2134
Alcohol Consumption (drinks/week)	506
Physical Activity (hrs/week)	87
Caffeine Intake (mg/day)	4
Gender	0
Age	0
Smoking	0
Family History of Anxiety	0
Stress Level (1-10)	0
Sleep Hours	0
Heart Rate (bpm)	0
Breathing Rate (breaths/min)	0
Dizziness	0
Sweating Level (1-5)	0
Medication	0
Recent Major Life Event	0
Diet Quality (1-10)	0
Anxiety Level (1-10)	0

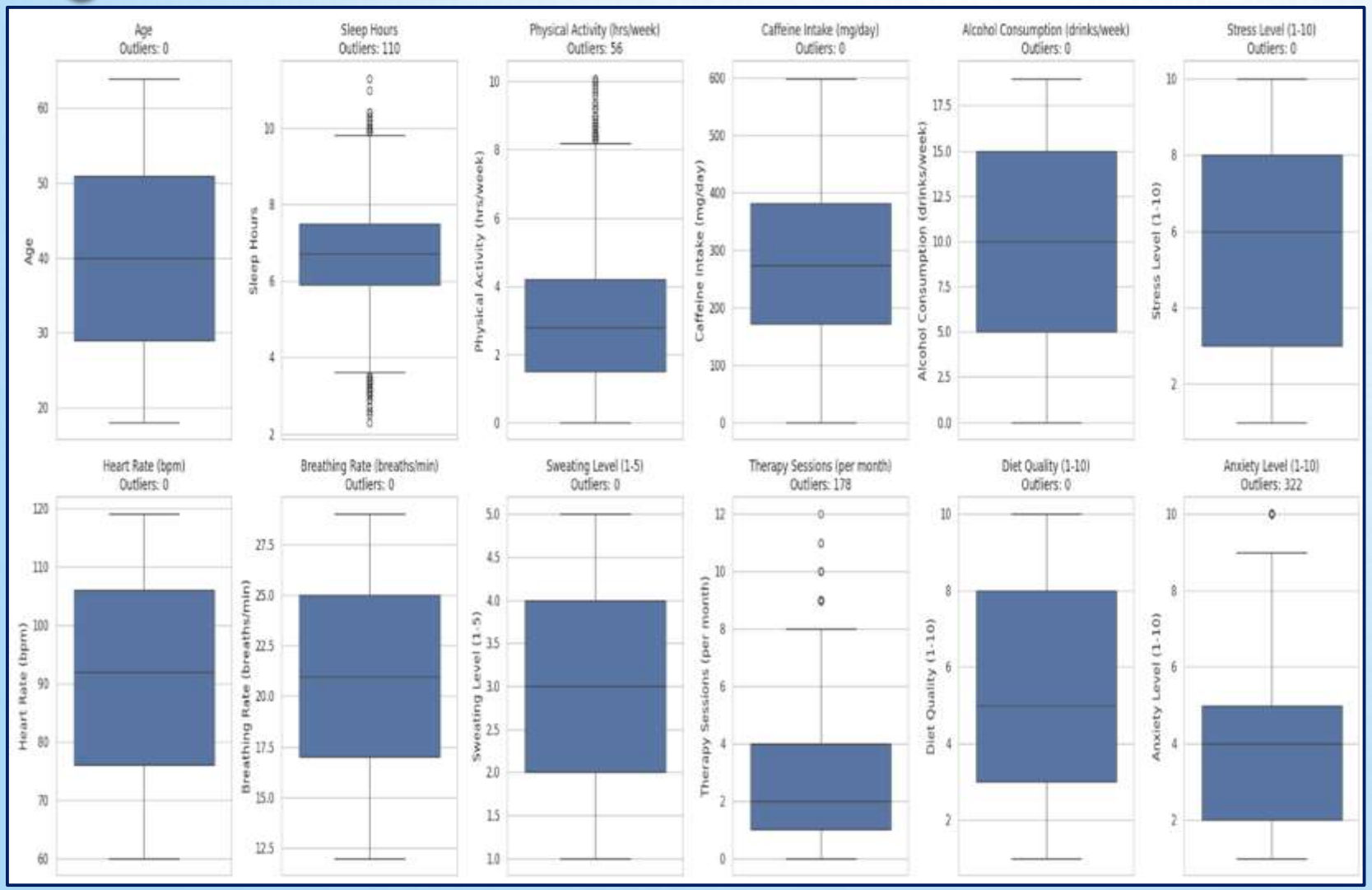
Variables que tienen valores cero (0):

- **Terapia:** Personas con 0 sesiones al mes
- **Alcohol:** Personas que no consumen (0 tragos/día)
- **Actividad física:** Personas sedentarias (0 horas/semana)
- **Cafeína:** Personas sin consumo (0 mg/día)

📌 **Por lo tanto :** Los valores en cero (0) son coherentes y deben conservarse, ya que reflejan realidades relevantes para el análisis

DATA WRANGLING (2)

TRATAMIENTO DE VALORES OUTLIERS



Variables que cuentan con Outliers:

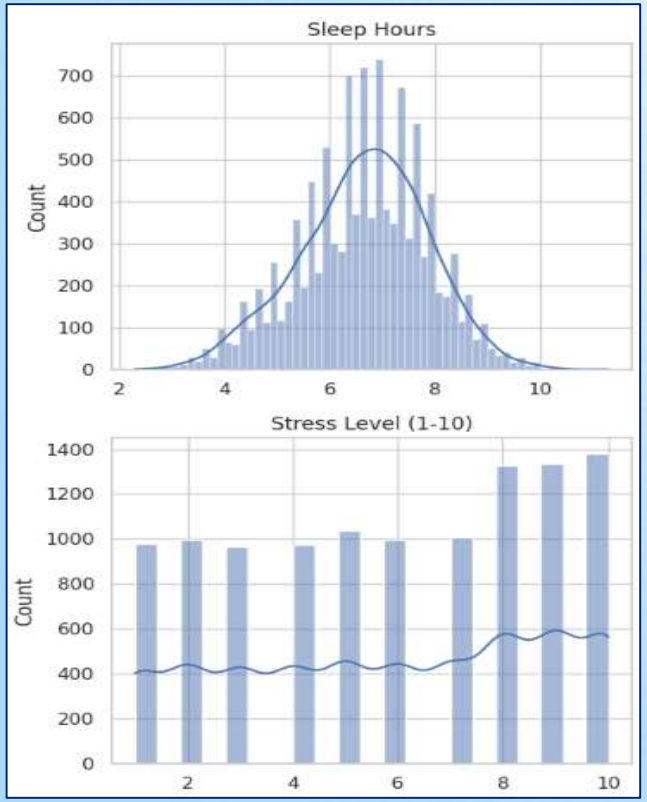
- **Sueño:** Personas con muy pocas o muchas horas de descanso
- **Actividad física:** Algunos superan las 8–9 horas semanales
- **Terapia:** Casos con más de 10 sesiones al mes
- **Ansiedad:** Niveles extremos, cercanos a 10

📌 **Por lo tanto:** Estos outliers no deben eliminarse, ya que aportan información valiosa

ANÁLISIS EXPLORATORIO DE DATOS

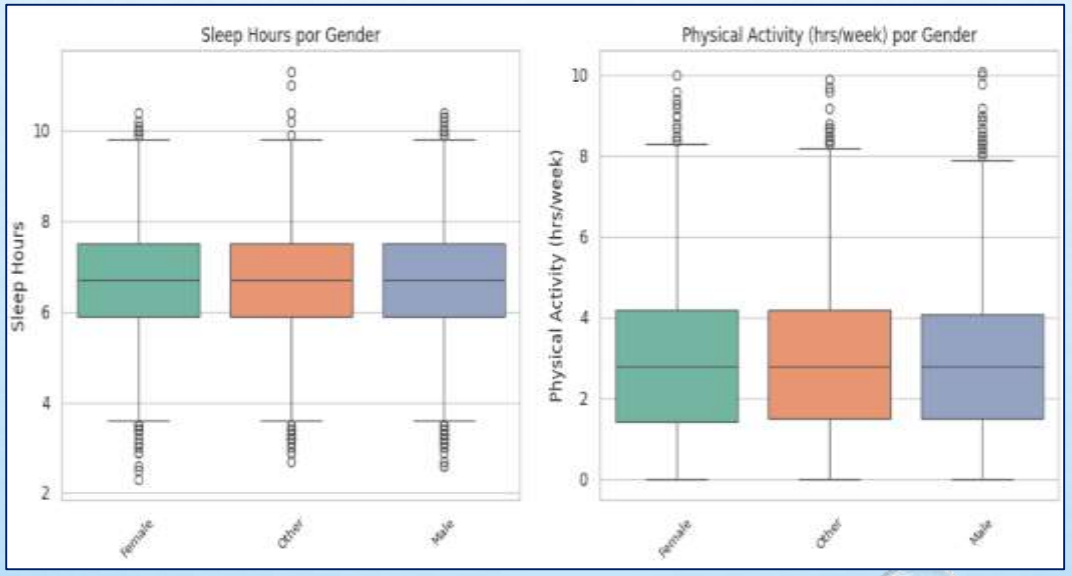
ANÁLISIS UNIVARIADO

Analiza una sola variable individualmente



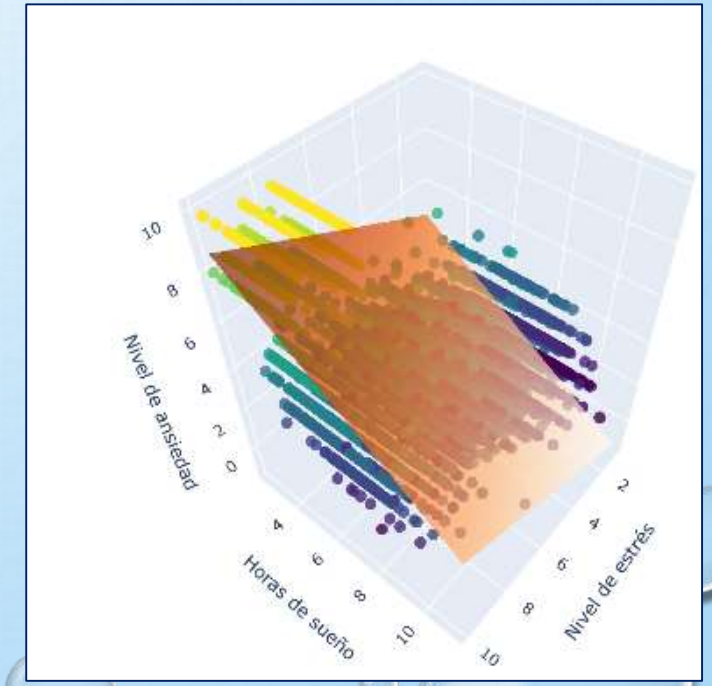
ANÁLISIS BIVARIADO

Analiza la relación entre dos variables



ANÁLISIS MULTIVARIADO

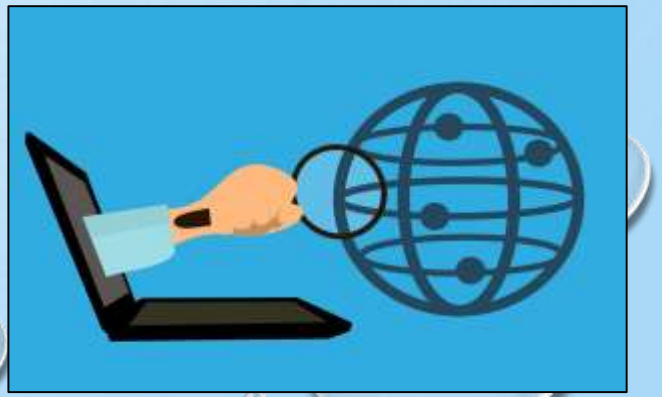
Analiza la relación entre más de dos variables simultáneamente



ANÁLISIS EXPLORATORIO DE DATOS (2)

CONCLUSIONES

- ✓ La mayoría de variables muestra distribuciones razonables y balanceadas en la data, sin outliers extremadamente atípicos o anómalos.
- ✓ El estrés y las horas de sueño son las variables más relacionadas con la ansiedad. A mayor **estrés**, mayor **ansiedad**. A menos **horas de sueño**, mayor **ansiedad** también. Gestionar el estrés y asegurar un sueño adecuado resultan clave para manejar la ansiedad.
- ✓ Factores como la dieta, la actividad física, el consumo de alcohol y de cafeína muestran una correlación muy débil con la ansiedad. Aunque la actividad física y una mejor dieta tienden a asociarse con menor ansiedad, el impacto no es tan significativo como el del estrés o el sueño.
- ✓ La frecuencia de sesiones de terapia está moderadamente relacionada con la ansiedad, probablemente debido a que las personas con más ansiedad sean quienes asistan a terapia con más frecuencia.
- ✓ Existe relación conjunta entre el consumo de alcohol, la calidad en la dieta y el nivel de ansiedad, donde tanto el aumento del consumo de alcohol como una menor calidad en la dieta contribuyen a un mayor nivel de ansiedad, tanto en conjunto como de forma independiente.



FEATURE ENGINEERING

SE OPTÓ POR GENERAR DOS VARIABLES NUEVAS CON EL FIN DE ENRIQUECER EL DATASET Y DARLE MAYOR VALOR

'TOTAL STIMULANT CONSUMPTION' - *CONSUMO TOTAL DE ESTIMULANTES*

FÓRMULA: 'Caffeine Intake (mg/day)' + 'Alcohol Consumption (drinks/week)' * alcohol_factor

***DONDE*:** alcohol_factor = 14

*El fin es escalar el alcohol para que tenga magnitud comparable a la cafeína. Un vaso de bebida alcohólica puede contener hasta 14 mg de alcohol según la OMS y el NIAAA

'PHYSIOLOGICAL AROUSAL INDEX' - *ÍNDICE DE ESTIMULACIÓN FISIOLÓGICA*

FÓRMULA: 'Heart Rate (bpm)' + 'Breathing Rate (breaths/min)' + 'Sweating Level (1-5)'

***DONDE*:** Previo a la suma , se estandarizan los valores de las columnas/variables para que tengan media 0 y desviación estándar 1, permitiendo compararlas en una misma escala, ya que tienen unidades distintas

FEATURE SELECTION

Se aplicó un ranking de importancia de las variables dependientes en la predicción de la variable objetivo, con el cual se decidió excluir aquellas con un nivel de importancia menor al 1%, pues la diferencia entre su peso y el de las variables principales (las que encabezan el ranking) justifica su pérdida de relevancia para el modelo.

Ranking de Importancia de Variables:

	Variable	Importancia (Peso)	Importancia (%)
5	Stress Level (1-10)	0.467628	46.76
1	Sleep Hours	0.217846	21.78
9	Therapy Sessions (per month)	0.092547	9.25
3	Caffeine Intake (mg/day)	0.033972	3.40
11	Physiological Arousal Index	0.024977	2.50
12	Total Stimulant Consumption	0.023228	2.32
2	Physical Activity (hrs/week)	0.023160	2.32
6	Heart Rate (bpm)	0.019969	2.00
10	Diet Quality (1-10)	0.019483	1.95
0	Age	0.019354	1.94
7	Breathing Rate (breaths/min)	0.013957	1.40
4	Alcohol Consumption (drinks/week)	0.013458	1.35
8	Sweating Level (1-5)	0.007568	0.76
13	Gender_Female	0.002275	0.23
14	Gender_Male	0.002240	0.22
15	Gender_Other	0.002156	0.22
18	Family History of Anxiety_No	0.001924	0.19
19	Family History of Anxiety_Yes	0.001684	0.17
21	Dizziness_Yes	0.001605	0.16
22	Medication_No	0.001603	0.16
20	Dizziness_No	0.001602	0.16
16	Smoking_No	0.001596	0.16
23	Medication_Yes	0.001586	0.16
24	Recent Major Life Event_No	0.001535	0.15
25	Recent Major Life Event_Yes	0.001524	0.15
17	Smoking_Yes	0.001524	0.15

Es así como el dataset termina quedando de la siguiente manera (de 18 a 13 variables).

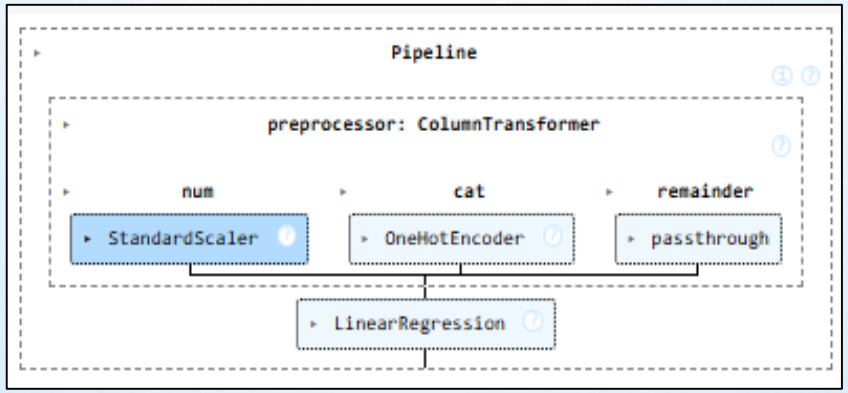
	Stress Level (1-10)	Sleep Hours	Therapy Sessions (per month)	Caffeine Intake (mg/day)	Physiological Arousal Index	Total Stimulant Consumption	Physical Activity (hrs/week)	Heart Rate (bpm)	Diet Quality (1-10)	Age	Breathing Rate (breaths/min)	Alcohol Consumption (drinks/week)	Anxiety Level (1-10)
0	10	6.0	3	181	0.641265	321	2.7	114	7	29	14	10	5
1	1	6.2	2	200	-2.045743	312	5.7	62	8	46	23	8	3
2	1	5.0	1	117	1.312053	173	3.7	91	1	64	28	4	1
3	4	5.8	0	360	-1.108384	444	2.8	86	1	20	17	6	2
4	1	8.2	1	247	0.686757	303	2.3	98	3	49	19	4	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10995	8	6.1	1	566	-0.117730	692	3.1	91	3	23	28	9	6
10996	7	6.6	2	64	-0.588902	302	3.6	95	7	50	17	17	3
10997	8	6.7	2	159	-3.540054	355	6.9	72	7	29	16	14	4
10998	4	5.7	1	248	2.524179	360	2.7	112	2	53	28	8	4
10999	1	6.1	2	205	-3.038003	359	1.1	66	8	56	13	11	2

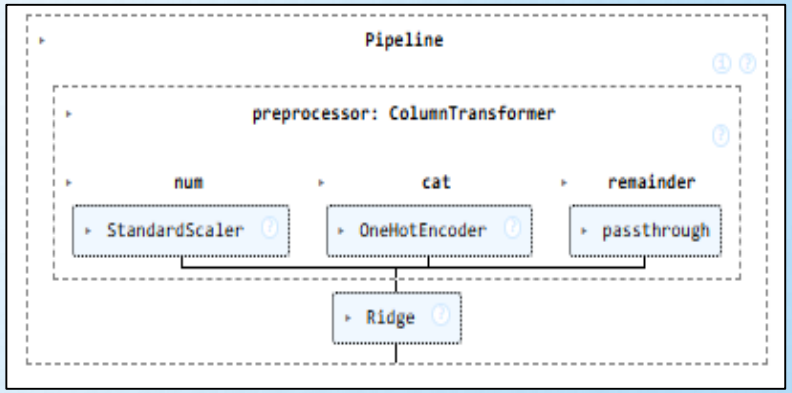
11000 rows × 13 columns

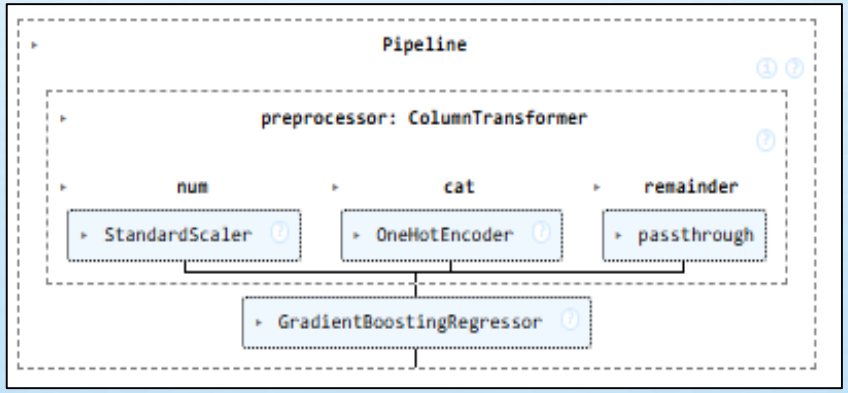
MODELADO

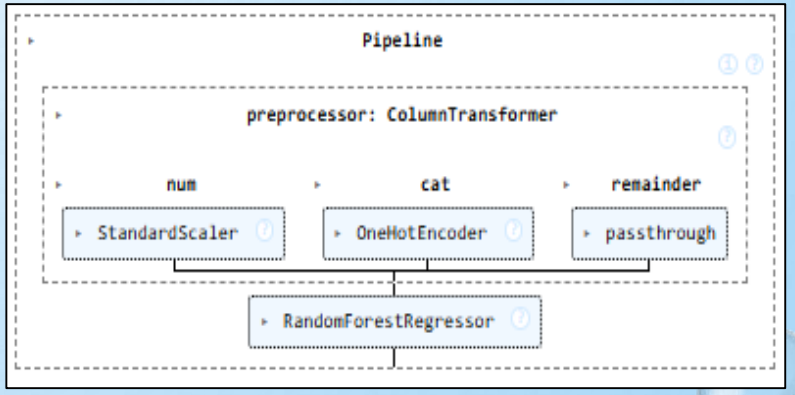
MODELOS DE REGRESIÓN EVALUADOS

- ✓ **Regresión Lineal**: Encuentra la recta que mejor explique la relación entre las variable
- ✓ **Regresión Ridge**: Variante de la regresión lineal que evita sobreajustes penalizando a los coeficientes
- ✓ **Random Forest (RF)**: Combina múltiples árboles de decisión sobre muestras aleatorias para lograr predicciones más precisas y estables
- ✓ **Gradient Boosting**: Construye árboles de forma secuencial, corrigiendo los errores del RF









Para desarrollar y evaluar cada modelo, se optó por el uso de **Pipelines** debido al flujo de trabajo secuencial y automatizado que éstos garantizan durante todo el proceso.

MODELADO (2)

CONCLUSIONES

El modelo **más conveniente** para predecir es el **Random Forest** ✓:

- ✓ Tiene el menor error absoluto medio (MAE): Predice los valores reales con mayor cercanía
- ✓ Tiene el menor error cuadrático medio (MSE): Minimiza el impacto de los errores grandes
- ✓ Tiene la menor raíz del error cuadrático medio (RMSE): Reduce errores significativos, ofreciendo predicciones consistentes y confiables
- ✓ Tiene el mayor coeficiente de determinación (R^2): Explica mejor la variabilidad de los datos, con mayor fidelidad el comportamiento real

	Regresión Lineal	Random Forest	Regresión Ridge	Gradient Boosting
MAE	0.894744	0.814277	0.894743	0.821425
MSE	1.272558	1.018432	1.272552	1.040597
RMSE	1.128077	1.009174	1.128075	1.020096
R^2	0.725368	0.780211	0.725369	0.775428

OPTIMIZACIÓN DE MODELOS

Métrica	Modelo Original	Validación Cruzada	Margen CV
MAE	0.814277	0.820668	0.017326
MSE	1.018432	1.042942	0.045075
RMSE	1.009174	1.021005	0.022150
R^2	0.780211	0.768194	0.012820

Métrica	Modelo Original	Ajuste de Hiperparámetros
MAE	0.814277	0.825230
MSE	1.018432	1.041998
RMSE	1.009174	1.020629
R^2	0.780211	0.768540

- La **Validación Cruzada** muestra ligeros incrementos en métricas de error (MAE, MSE y RMSE), mientras que el coeficiente de determinación (R^2) presenta disminución mínima. Esto transmite que el modelo mantiene su capacidad predictiva, evidenciando estabilidad y generalización adecuadas.
- El **Ajuste de Hiperparámetros** muestra resultados similares en métricas, indicando en este caso que el modelo no experimentó mejoras significativas en su desempeño, indicando que la configuración original del modelo ya era cercana al óptimo.
- **Ambos casos** evidencian que el modelo Random Forest es robusto, estable y generalizable, con bajo riesgo de overfitting, manteniendo su capacidad predictiva sin variaciones relevantes.

CONCLUSIONES FINALES

✓ **Valores cero:** Coherentes con la realidad, no requieren limpieza.
✓ **Outliers:** Casos de interés, mantenerlos permitió capturar escenarios reales y extremos de ansiedad.
✓ **Variables clave:** Ansiedad aumenta con estrés alto y sueño insuficiente. Estas dos últimas son las variables más importantes para predecir la ansiedad en una persona.
✓ **Mejor modelo:** De los 4, Random Forest es el mejor para predecir con precisión y explicar mejor la variabilidad de los datos.
✓ **Confiabilidad del modelo:** Los métodos de optimización confirman su estabilidad, generabilidad y bajo riesgo de overfitting. La Validación Cruzada muestra estabilidad, mientras que el Ajuste de Hiperparámetros al no mejorar resultados, indica que la configuración del modelo original ya era óptima.