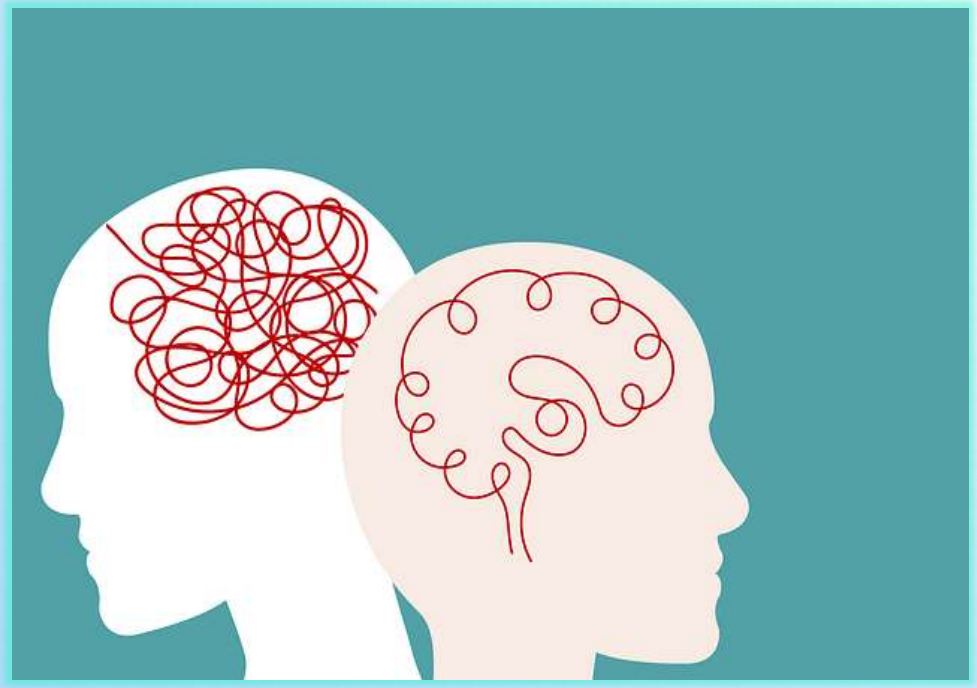
CODERHOUSE

# PROYECTO FINAL
# ANÁLISIS PREDICTIVO DE ANSIEDAD SOCIAL




- ***ESTUDIANTE:*** ALFREDO JASAUI CHERO
- ***CURSO:*** DATA SCIENCE I
- ***COMISIÓN:*** 67465
- ***PROFESOR:*** JUAN CRUZ ALRIC CORTABARRIA
- ***TUTOR:*** LUCIANO LISACHI

# CONTENIDOS

1. *Contexto y Objetivo*
2. *Importación de Librerías y Datos*
3. *Hipótesis*
4. *Análisis Exploratorio de Datos-EDA*
5. *Feature Engineering*
6. *Preprocesamiento y División de Datos*
7. *Construcción y Entrenamiento del Modelo*
8. *Validación del Modelo*
9. *Análisis de Importancia de Características*
10. *Conclusiones*

# CONTEXTO Y OBJETIVO

LA ANSIEDAD SOCIAL AFECTA A MILLONES DE PERSONAS EN EL MUNDO Y REPRESENTA UN DESAFÍO CRECIENTE EN MATERIA DE SALUD MENTAL. EN UN CONTEXTO DONDE LA ATENCIÓN TEMPRANA Y LA PREVENCIÓN SON CADA VEZ MÁS VALORADAS CONTAR CON HERRAMIENTAS PARA IDENTIFICAR FACTORES DE RIESGO Y PATRONES PUEDE APORTAR VALOR TANTO PARA INSTITUCIONES DE SALUD COMO PARA INVESTIGADORES Y DESARROLLADORES DE TECNOLOGÍA APLICADA.

LA DISPONIBILIDAD DE DATA SOBRE ANSIEDAD SOCIAL INCLUSO ABRE OPORTUNIDADES PARA GENERAR MODELOS DESCRIPTIVOS Y PREDICTIVOS GENERANDO BENEFICIOS TANTO SOCIALES DESDE EL LADO DE LOS PACIENTES; COMO COMERCIALES DESDE EL LADO DE LOS PROFESIONALES ACADÉMICOS Y FUNCIONARIOS PÚBLICOS DEDICADOS A LA SALUD MENTAL.

***OBJETIVO***

CONSTRUIR UN MODELO DE MACHINE LEARNING CAPAZ DE PREDECIR EL NIVEL DE ANSIEDAD DE UNA PERSONA EN BASE A UN CONJUNTO DE ATRIBUTOS. ESTA ES UNA TAREA DE REGRESIÓN YA QUE LA VARIABLE OBJETIVO **(NIVEL DE ANSIEDAD)** ES NUMÉRICA DISCRETA Y ORDENADA. EL MODELO A CONSTRUIR ESTÁ BASADO EN RANDOM FOREST.

# IMPORTACIÓN DE LIBRERÍAS Y DATOS

### *CARGA DE LIBRERÍAS*

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


### *CARGA DE DATOS*

DATASET  DESDE LA URL DE ORIGEN

**VISTA PRELIMINAR DE LOS PRIMEROS 5 REGISTROS*

| | 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 | 5.0 |
| 1 | 46 | Other | 6.2 | 5.7 | 200 | 8 | Yes | Yes | 1 | 62 | 23 | 2 | Yes | No | 2 | No | 8 | 3.0 |
| 2 | 64 | Male | 5.0 | 3.7 | 117 | 4 | No | Yes | 1 | 91 | 28 | 3 | No | No | 1 | Yes | 1 | 1.0 |
| 3 | 20 | Female | 5.8 | 2.8 | 360 | 6 | Yes | No | 4 | 86 | 17 | 3 | No | No | 0 | No | 1 | 2.0 |
| 4 | 49 | Female | 8.2 | 2.3 | 247 | 4 | Yes | No | 1 | 98 | 19 | 4 | Yes | Yes | 1 | No | 3 | 1.0 |

# HIPÓTESIS

## HIPÓTESIS NULA (H0)

LAS CARACTERÍSTICAS DISPONIBLES NO SON SUFICIENTES PARA PREDECIR EL NIVEL DE ANSIEDAD DE UNA PERSONA TENIENDO UN RENDIMIENTO DE MODELO SIGNIFICATIVAMENTE MEJOR QUE EL DE UN MODELO BASE. SE ESPERARÍA UN $R^2$ CERCANO A 0.

## HIPÓTESIS ALTERNATIVA (H1)

ES POSIBLE CONSTRUIR UN MODELO DE REGRESIÓN BASADO EN RANDOM FOREST UTILIZANDO LAS CARACTERÍSTICAS DISPONIBLES PARA PREDECIR EL NIVEL DE ANSIEDAD DE UNA PERSONA ALCANZANDO UN COEFICIENTE DE DETERMINACIÓN ($R^2$) SIGNIFICATIVAMENTE MAYOR A 0 IDEALMENTE MAYOR A 0.5 Y UN ERROR ABSOLUTO MEDIO (MAE) RAZONABLEMENTE BAJO LO QUE INDICARÍA UNA CAPACIDAD PREDICTIVA ÚTIL DEL MODELO.


# ANÁLISIS EXPLORATORIO DE DATOS-EDA

***DESCRIPCIÓN DE LAS 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)

# ANÁLISIS EXPLORATORIO DE DATOS-EDA (2)

```
RangeIndex: 999 entries, 0 to 998
Data columns (total 18 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Age                             999 non-null    int64
 1   Gender                          999 non-null    object
 2   Sleep Hours                     999 non-null    float64
 3   Physical Activity (hrs/week)    999 non-null    float64
 4   Caffeine Intake (mg/day)        999 non-null    int64
 5   Alcohol Consumption (drinks/week)  999 non-null  int64
 6   Smoking                         999 non-null    object
 7   Family History of Anxiety       999 non-null    object
 8   Stress Level (1-10)             999 non-null    int64
 9   Heart Rate (bpm)                999 non-null    int64
 10  Breathing Rate (breaths/min)    999 non-null    int64
 11  Sweating Level (1-5)            999 non-null    int64
 12  Dizziness                       999 non-null    object
 13  Medication                      999 non-null    object
 14  Therapy Sessions (per month)    999 non-null    int64
 15  Recent Major Life Event         999 non-null    object
 16  Diet Quality (1-10)             999 non-null    int64
 17  Anxiety Level (1-10)            999 non-null    float64
dtypes: float64(3), int64(9), object(6)
```

***INFORMACIÓN GENERAL DEL DATASET***

- 999 filas
- 18 columnas
- Sin valores nulos (NULL)
- Tipos de datos (datatypes):
  - *Int64*: Numérico entero (9 col.)
  - *Float64*: Numérico decimal (3 col.)
  - *Object*: No numérico (6 col.)

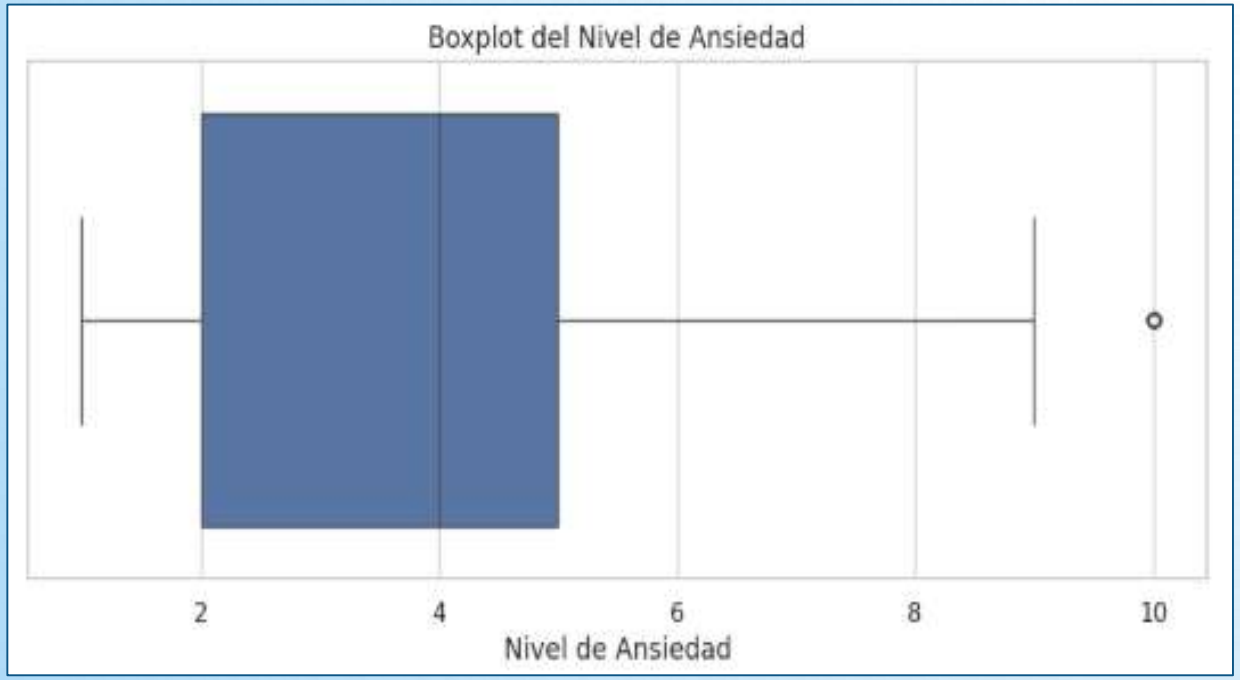
***ANÁLISIS DESCRIPTIVO POR VARIABLE***

- Conteos
- Medias
- Desviaciones estándar
- Valores mínimos
- Primeros cuartiles (25%)
- Medianas
- Terceros cuartiles (75%)
- 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 | 999.000000 | 999.000000 | 999.000000 | 999.000000 | 999.000000 | 999.000000 | 999.000000 | 999.000000 | 999.000000 | 999.000000 | 999.000000 | 999.000000 |
| mean | 39.842843 | 6.599800 | 2.881481 | 291.180180 | 10.060060 | 5.814815 | 89.814815 | 20.783784 | 3.127127 | 2.453453 | 5.197197 | 3.972973 |
| std | 12.883800 | 1.258112 | 1.811422 | 148.440509 | 5.752441 | 2.914744 | 17.565815 | 5.067806 | 1.389502 | 2.208924 | 2.876926 | 2.176636 |
| min | 18.000000 | 2.500000 | 0.000000 | 2.000000 | 0.000000 | 1.000000 | 60.000000 | 12.000000 | 1.000000 | 0.000000 | 1.000000 | 1.000000 |
| 25% | 29.000000 | 5.800000 | 1.400000 | 171.000000 | 6.000000 | 3.000000 | 75.000000 | 16.000000 | 2.000000 | 1.000000 | 3.000000 | 2.000000 |
| 50% | 39.000000 | 6.700000 | 2.700000 | 276.000000 | 10.000000 | 6.000000 | 90.000000 | 21.000000 | 3.000000 | 2.000000 | 5.000000 | 4.000000 |
| 75% | 50.500000 | 7.500000 | 4.100000 | 391.000000 | 15.000000 | 8.000000 | 105.500000 | 25.000000 | 4.000000 | 4.000000 | 8.000000 | 5.000000 |
| max | 64.000000 | 9.900000 | 9.900000 | 599.000000 | 19.000000 | 10.000000 | 119.000000 | 29.000000 | 5.000000 | 9.000000 | 10.000000 | 10.000000 |

# ANÁLISIS EXPLORATORIO DE DATOS-EDA (3)

## *DISTRIBUCIÓN DE LA VARIABLE OBJETIVO*




**La mayoría de personas presentan niveles de ansiedad entre 2 y 5. De la misma forma, la distribución está ligeramente sesgada hacia niveles elevados.**

## *CORRELACIÓN DE VARIABLES*

```
Correlaciones con la variable 'Anxiety Level (1-10)':
Anxiety Level (1-10)                1.000000
Stress Level (1-10)                 0.652753
Therapy Sessions (per month)        0.548083
Caffeine Intake (mg/day)            0.373950
Heart Rate (bpm)                    0.246266
Breathing Rate (breaths/min)        0.225927
Sweating Level (1-5)                0.150223
Smoking                             0.111429
Dizziness                           0.093040
Recent Major Life Event             0.085317
Alcohol Consumption (drinks/week)   0.079516
Medication                          0.033016
Gender                              0.009689
Age                                -0.126030
Diet Quality (1-10)                -0.224446
Physical Activity (hrs/week)       -0.240881
Sleep Hours                        -0.517240
Name: Anxiety Level (1-10), dtype: float64
```

**El Nivel de Estrés tiene la correlación positiva más alta con el Nivel de Ansiedad, mientras que las Horas de Sueño tienen la correlación negativa más alta.**

# FEATURE ENGINEERING

**SE GENERARON DOS NUEVAS CARACTERÍSTICAS DE VALOR AL DATASET:**

**'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

# PREPROCESAMIENTO Y DIVISIÓN DE DATOS

## PREPROCESAMIENTO

IMPLICA PREPARAR LOS DATOS EN UN FORMATO MÁS ADECUADO Y ÚTIL PARA SU POSTERIOR PROCESAMIENTO DE MODELADO

```
# Definir las características (en el eje X) y la variable objetivo (en el eje Y)
X = df.drop('Anxiety Level (1-10)', axis=1)
y = df['Anxiety Level (1-10)']
categorical_features_for_encoding = X.select_dtypes(include=['object', 'category']).columns.tolist()
numerical_features_for_scaling = X.select_dtypes(include=['int64', 'float64']).columns.difference(categorical_features_for_encoding).tolist()
```

```
# Crear el preprocesador
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numerical_features_for_scaling),
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features_for_encoding)
    ],
    remainder='passthrough'  # O 'drop'
)
```

## DIVISIÓN

IMPLICA DIVIDIR LOS DATOS EN CONJUNTOS DE ENTRENAMIENTO Y PRUEBA, PREVIA APLICACIÓN DEL PREPROCESAMIENTO

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
print(f"Tamaño del conjunto de entrenamiento (X_train): {X_train.shape}")
print(f"Tamaño del conjunto de prueba (X_test): {X_test.shape}")
print(f"Tamaño del conjunto de entrenamiento (y_train): {y_train.shape}")
print(f"Tamaño del conjunto de prueba (y_test): {y_test.shape}")
```

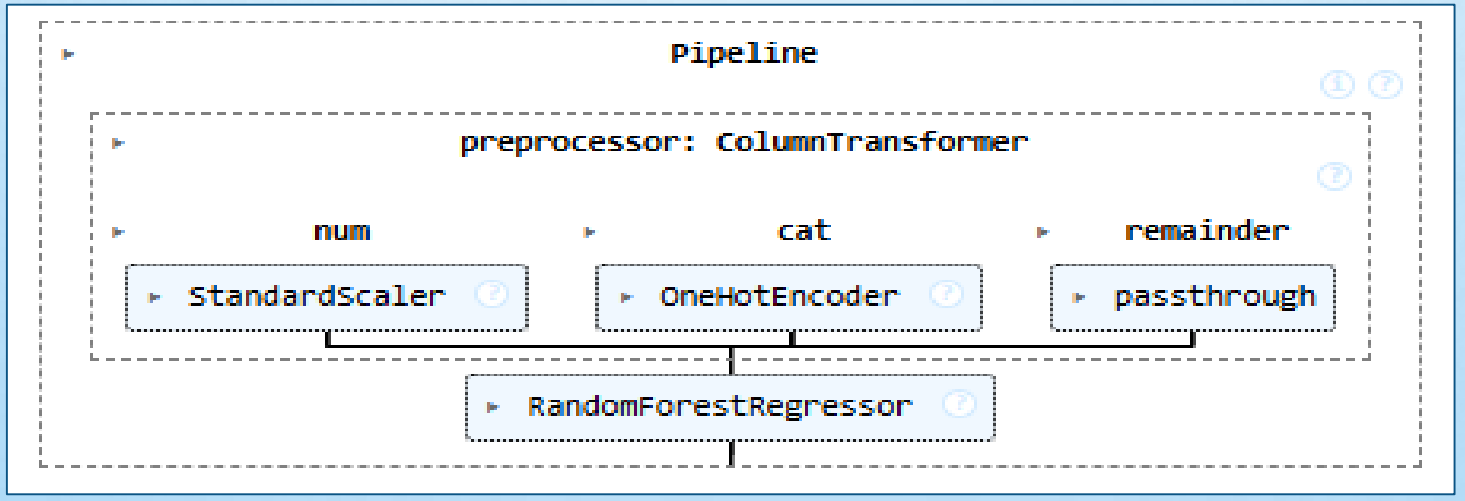
```
Tamaño del conjunto de entrenamiento (X_train): (799, 19)
Tamaño del conjunto de prueba (X_test): (200, 19)
Tamaño del conjunto de entrenamiento (y_train): (799,)
Tamaño del conjunto de prueba (y_test): (200,)
```

# CONSTRUCCIÓN Y ENTRENAMIENTO DEL MODELO

## CONSTRUCCIÓN

```
model_pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('regressor', RandomForestRegressor
     (n_estimators=100, # Refiere al número de árboles. 100 es un buen inicio ya que generalmente da buenos resultados sin requerir demasiado tiempo ni costo de cómputo
      random_state=42,  # Para asegurar reproducibilidad de los resultados
      n_jobs=-1,        # Usa todos los procesadores disponibles para entrenar más rápido
      max_depth=10,     # Profundidad máxima de los árboles. Ayuda a prevenir overfitting, al evitar que los árboles se ajusten demasiado a los datos de entrenamiento
      min_samples_split=5) #Número mínimo de muestras para dividir un nodo. Si es alto, los árboles serán más conservadores y menos complejos
     )
])
```

## ENTRENAMIENTO



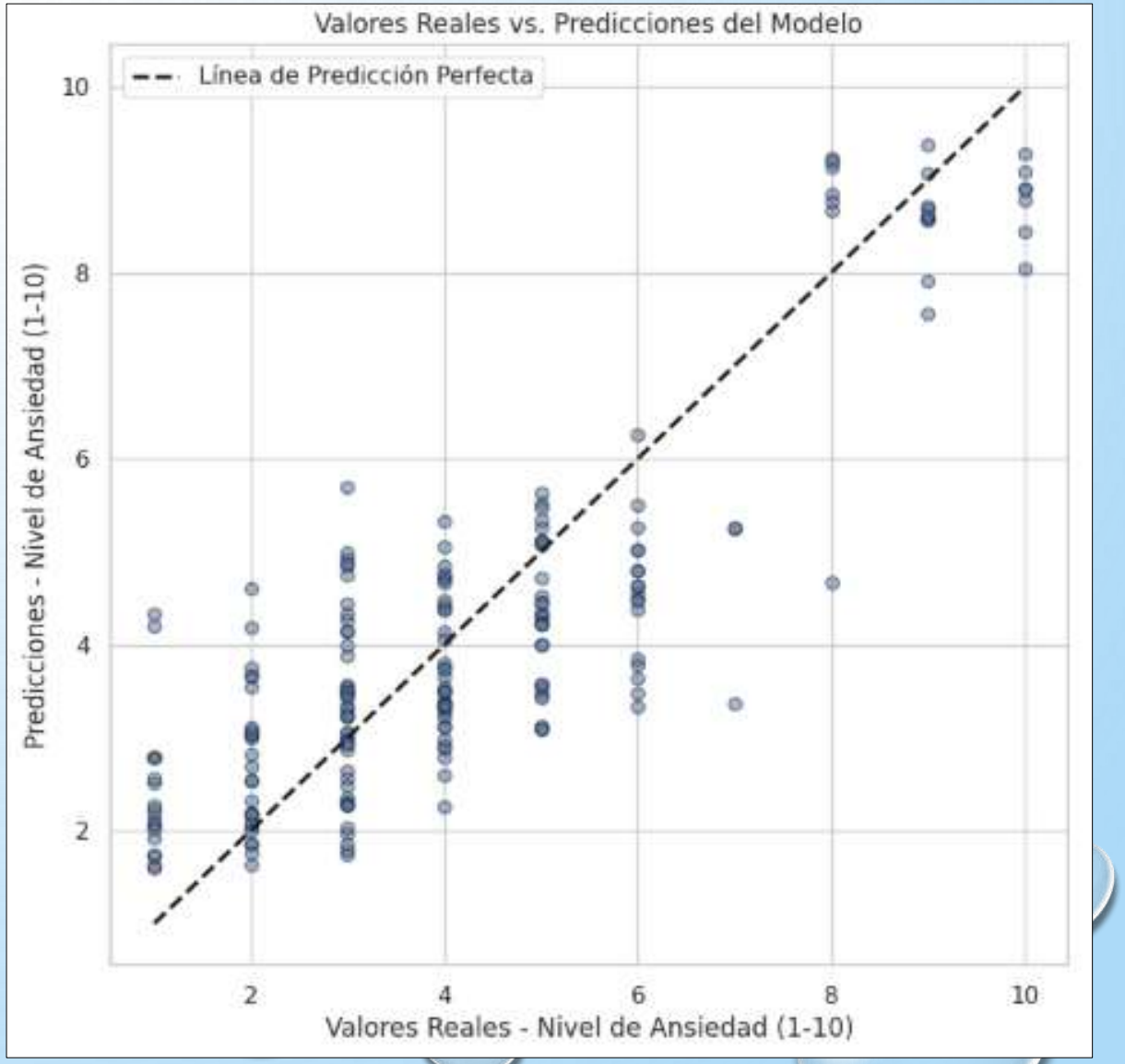
# VALIDACIÓN DEL MODELO

## MÉTRICAS

- ✓ Error Absoluto Medio (MAE): 0.9441 → Las predicciones del modelo se desvían en promedio, en 0.9441 puntos
- ✓ Error Cuadrático Medio (MSE): 1.3696
- ✓ Raíz del Error Cuadrático Medio (RMSE): 1.1703 → La desviación típica de los errores es de 1.1703
- ✓ Coeficiente de Determinación ($R^2$): 0.7376 → El 73.76% de la varianza en los niveles de ansiedad puede ser explicada por las características del modelo

## GRÁFICO DE VALORES VS. PREDICCIONES

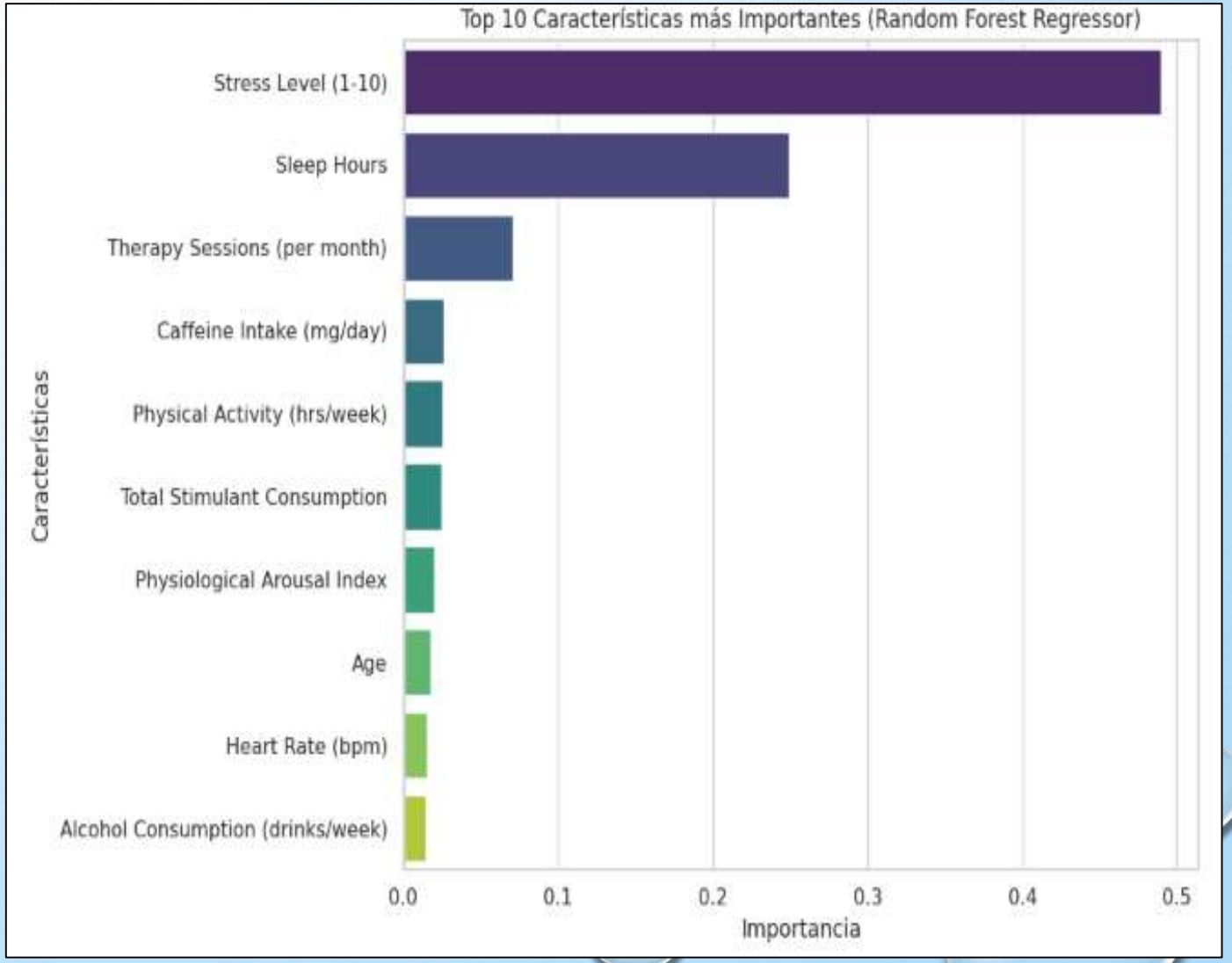
La mayoría de puntos están ubicados cerca de la línea de predicción perfecta, lo que sugiere que el modelo funciona razonablemente bien. Por otro lado, el modelo cuenta con mayor dispersión o sesgo a partir de niveles de ansiedad entre 5 y 7.



# ANÁLISIS DE IMPORTANCIA DE CARACTERÍSTICAS

Importancia de las características según Random Forest Regressor:

| | Feature | Importance |
|---|---|---|
| 14 | Stress Level (1-10) | 0.489277 |
| 12 | Sleep Hours | 0.249053 |
| 16 | Therapy Sessions (per month) | 0.070670 |
| 3 | Caffeine Intake (mg/day) | 0.026034 |
| 9 | Physical Activity (hrs/week) | 0.025203 |
| 17 | Total Stimulant Consumption | 0.024663 |
| 10 | Physiological Arousal Index | 0.020287 |
| 0 | Age | 0.017601 |
| 7 | Heart Rate (bpm) | 0.014887 |
| 1 | Alcohol Consumption (drinks/week) | 0.014701 |




- **Las características más relevantes al momento de predecir el Nivel de Ansiedad de una persona son el Nivel de Estrés, las Horas de Sueño y las Sesiones Mensuales de Terapia**

- **Las características menos relevantes para ello son la Edad, el Ritmo Cardíaco y el Consumo de Alcohol**

# CONCLUSIONES

## RESULTADOS DEL MODELO DE REGRESIÓN

- ***Error Absoluto Medio (MAE):*** 0.9441
- ***Raíz del Error Cuadrático Medio (RMSE):*** 1.1703
- ***Coeficiente de Determinación ($R^2$):*** 0.7376
- Las **características más importantes,** al momento de predecir el Nivel de Ansiedad de una persona, fueron:
  Stress Level (1-10), Sleep Hours, Therapy Sessions (per month)

## VALIDACIÓN DE LA HIPÓTESIS

- Se obtuvo un $R^2$ de 0.74 y un MAE de 0.94
- Por lo tanto, **la Hipótesis Alternativa (H1) es soportada por los resultados.** El modelo explica una porción considerable de la varianza y es significativamente mejor que un modelo base
- De la misma forma **se rechaza la Hipótesis Nula (H0)**
- **ENTONCES, SÍ es posible** construir un modelo de regresión basado en Random Forest utilizando las características disponibles para predecir el nivel de ansiedad de una persona. Se demuestra una capacidad predictiva útil del modelo.