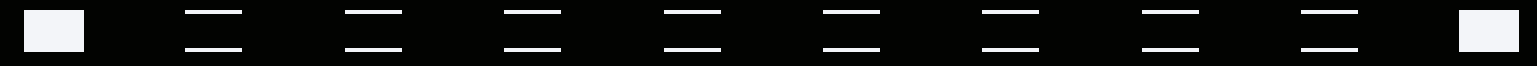


DULCES SUEÑOS PREDICTIVOS

DETECTANDO TRASTORNOS DEL
SUEÑO CON CIENCIA DE DATOS

Naiara Saratxaga





Se estima que alrededor del **27,3 % de la población mundial** padece al menos un trastorno del sueño, lo que equivale a aproximadamente **2.100 millones de personas** (basado en una población global de ~7.700 millones).

Algunos trastornos concretos destacan por su magnitud:

Apnea obstructiva del sueño (OSA): afecta cerca de 1.000 millones de adultos de entre 30 y 69 años en todo el mundo.

Insomnio: entre el 10 % y el 30 % de los adultos tienen síntomas de insomnio en un momento dado, lo que supondría entre 770 millones y 2.300 millones de personas.

En conjunto, estos datos reflejan que los trastornos del sueño constituyen un **problema de salud pública de primer orden**, con miles de millones de afectados en todas las edades y regiones.

OBJETIVOS

Este proyecto tiene como objetivo desarrollar un sistema que permita detectar trastornos del sueño a partir de datos obtenidos en DataSets de Kaggle.

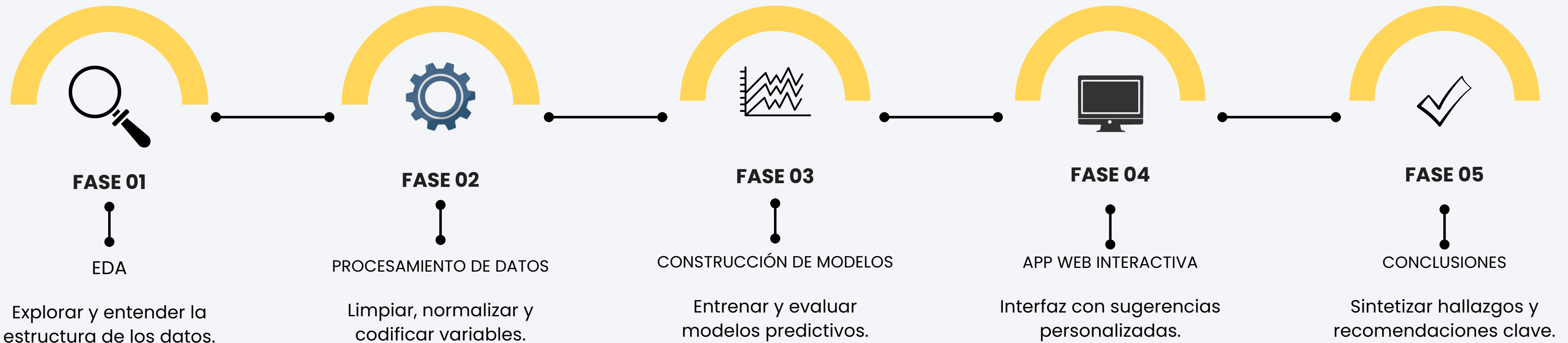
El enfoque se basa en un proceso de 3 etapas:

- 1. Detección binaria:** determinar si una persona presenta o no un trastorno del sueño.
- 2. Clasificación del tipo de trastorno:** Identificar de qué trastorno se trata (por ejemplo, insomnio, apnea, narcolepsia, etc.).
- 3. Segmentación de pacientes (Clustering):** Identificar perfiles de riesgo o patrones de comportamiento.

Finalmente se ha realizado **una aplicación web** interactiva, que, según el tipo de trastorno diagnosticado por el modelo, **ofrece sugerencias prácticas** y adaptadas para mejorar la calidad del descanso.



PARTES DEL PROYECTO

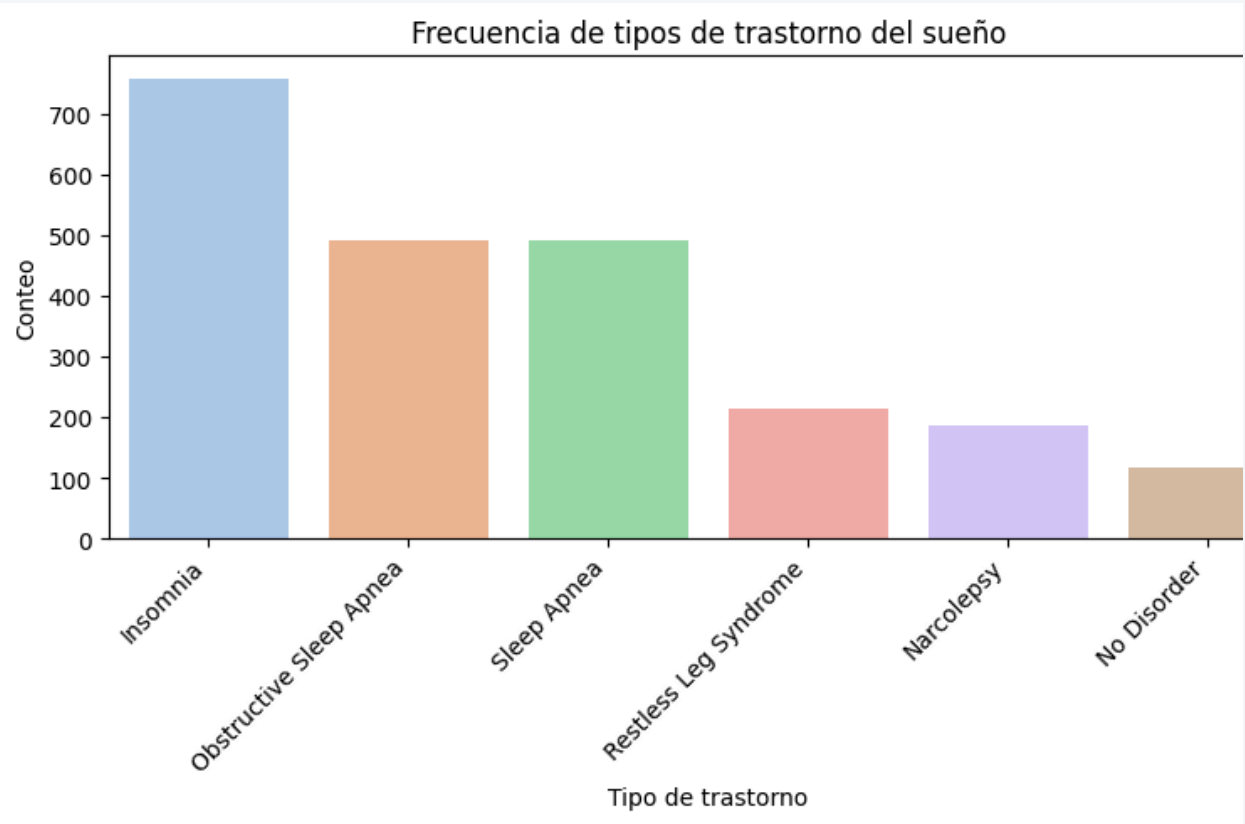


EDA

01

Análisis Univariante:

- Distribución de la variable objetivo (tipo de trastorno)



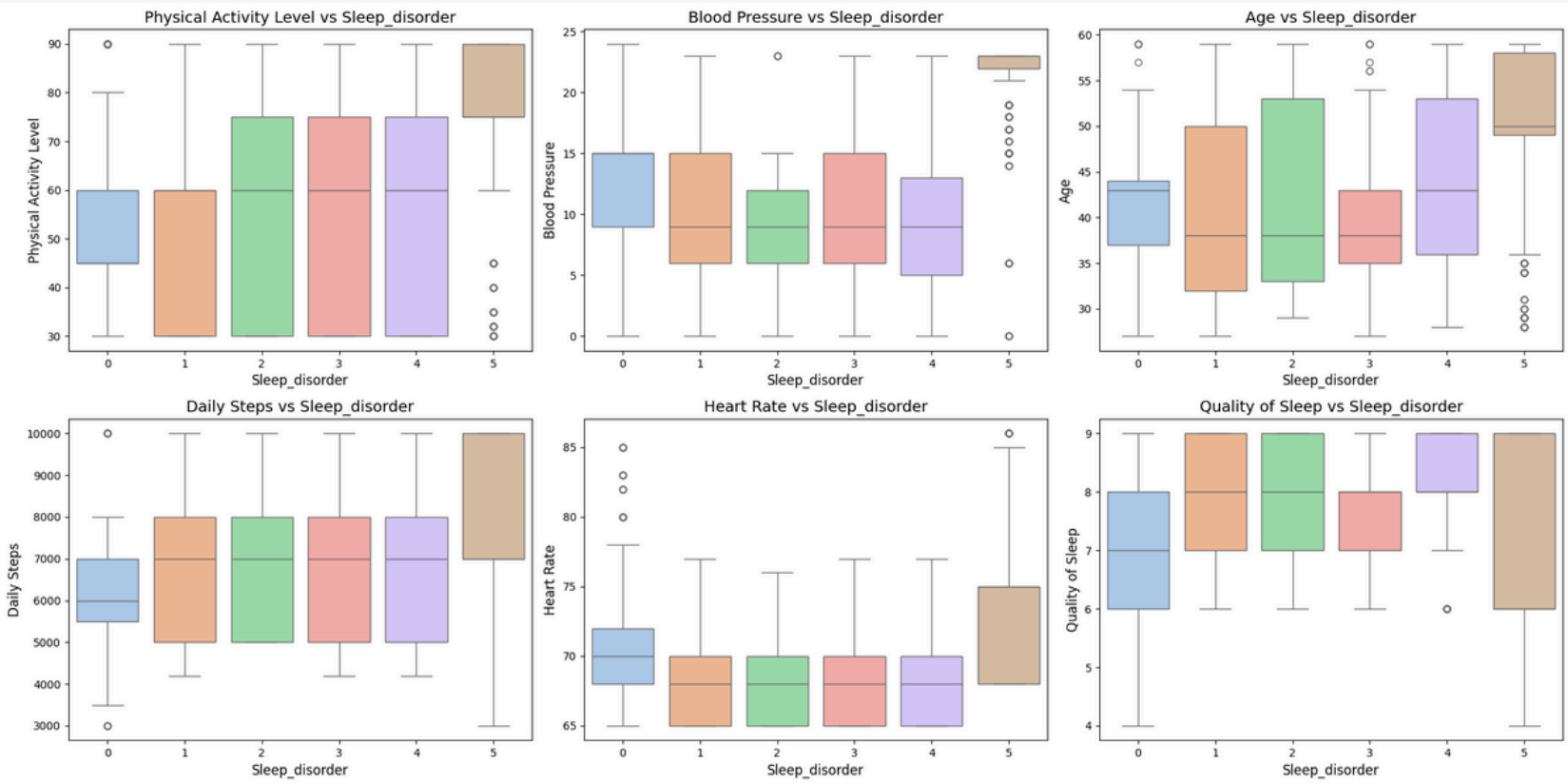
Conteo de cada categoría: Trastornos del sueño

- Insomnio: 757
- Apnea obstructiva del sueño: 492
- Apnea del sueño: 490
- Síndrome de piernas inquietas: 213
- Narcolepsia: 187
- Sin trastorno: 117

02

Análisis Bivariante:

- Relacionar dos variables para ver patrones, en este caso la variable objetivo con distintos features



Clases de trastornos

0 Insomnio • 1 Narcolepsia • 2 Sin trastorno • 3 Apnea obstructiva • 4 Piernas inquietas • 5 Apnea del sueño

Puntos clave de los boxplots

- Actividad física & pasos diarios: Apnea del sueño (5) muestra la mediana y dispersión más altas.
- Edad: Grupo 5 es el más mayor (mediana \approx 50 años), 0–4 rondan 33–38 años.
- Calidad del sueño: Peor en Insomnio (0, mediana \approx 7); el resto \approx 8–9.

Los pacientes con Apnea del sueño son mayores, con mayor presión arterial, frecuencia cardíaca y actividad física muy variable; los de Insomnio destacan por la peor calidad de sueño.

CONSTRUCCIÓN DE MODELOS






Diagnóstico Confirmado (Clasificación Binaria)

Objetivo: Predecir si un paciente tiene un diagnóstico confirmado (Diagnosis_Confirmed).

- Modelos entrenados:
 - Random Forest (base · balanced · SMOTE)
 - XGBoost (ajuste scale_pos_weight)
 - LightGBM (is_unbalanced=True)
 - Red Neuronal (SMOTE + class_weight)
- Métricas clave: Precisión · Recall · F1 · Exactitud · ROC-AUC
- Selección final:
 - Red Neuronal con umbral 0.35** → F1-macro 0.72 · AUC 0.7852 · Recall 73% en clase negativa

Modelo	Recall clase 0	F1 macro	ROC AUC	Accuracy
Random Forest (base)	0.24	0.65	—	0.89
Random Forest + balanced	0.75	0.52	—	0.60
Random Forest + SMOTE	0.69	0.54	—	0.63
XGBoost	0.78	0.52	—	0.59
LightGBM	0.24	0.65	—	0.89
 Red Neuronal (Keras)	0.73 (umbral 0.35)	0.72	0.7852	0.72



Tipo de Trastorno (Clasificación Multiclase)

Objetivo: Identificar el tipo de trastorno (Sleep_disorder).

- Modelos entrenados: RF (balanced) · XGBoost · LightGBM · Stacking · Red Neuronal
- Métrica principal: F1-ponderado · Exactitud
- Resultado:
 - Random Forest mejor rendimiento → F1 weighted 0.563
 - Red Neuronal se quedó cerca, pero no superó al ensemble

Modelo	Accuracy	F1-score (w)
RandomForest	0.531	0.563
XGBoost	0.588	0.560
LightGBM	0.531	0.563
Stacking	0.540	0.495
Red Neuronal	0.546	0.560



Segmentación de Pacientes (Clustering No Supervisado)

- Métodos probados:
 - K-Means (k óptimo = 4)
 - DBSCAN (demasiado ruido)
 - Clustering jerárquico (Ward)
- Preprocesamiento: StandardScaler + PCA(2D)
- Elección final:
 - K-Means (k=4) → perfiles diferenciados, validado con PCA

<u>Cluster</u>	Edad	Sueño (Duración/Calidad)	Actividad	Estrés	Salud Física
0	~41	Largo y de calidad	Moderada	Bajo	IMC y presión bajos
1	~49	Corto y de baja calidad	Muy alta	Alto	IMC y presión altos
2	~37	Moderado y de baja calidad	Baja	Medio	Presión media
3	~53	Largo y de muy buena calidad	Alta	Muy bajo	IMC alto

PREDICCIÓN Y RESULTADOS FINALES

01

Diagnóstico Binario

Red Neuronal optimizada
(umbral **0.35**)

- F1 = **0.73**
- Alta sensibilidad en la clase minoritaria (SMOTE + ajuste de umbral)

02

Clasificación Multiclase

Random Forest balanceado

- F1-ponderado \approx **0.56**
- Excelente para **“Sin trastorno” y “Apnea”**
- Dificultad en clases minoritarias (p. ej. Narcolepsia)

03

Segmentación de Pacientes

K-Means (k=4)

- Perfiles diferenciados según edad, calidad de sueño, actividad y salud
- Validado con PCA (separación clara de clusters)

APLICACIÓN INTERACTIVA EN STREAMLIT



Menú

Inicio

Analizar

Recomendador de Sueño Personalizado

 Introduce tus datos y descubramos juntos qué tipo de trastorno de sueño podrías estar padeciendo.

¿Cómo te llamas?

Género

Femenino

Edad

30

18 80

Duración del sueño (h)

6.00

0.00 12.00

Calidad del sueño (1-10)

5

1 10

Actividad física (0-100)

50

0 100

Nivel de estrés (0-10)

5

0 10

Presión arterial

120

CONCLUSIONES FINALES

Conclusiones Principales

- Es viable detectar y clasificar trastornos del sueño con datos clínicos y de estilo de vida, logrando buen rendimiento tras balancear clases (SMOTE, class_weight).
- La aplicación en Streamlit facilita la interpretación y despliegue de los resultados para el usuario final.

Áreas de Mejora

- Datos sintéticos y origen limitado: falta reflejar la variabilidad clínica real.

Futuros Pasos

1. Enriquecer el dataset con fuentes reales (hospitales, clínicas).
2. Incorporar nuevas variables (historial médico, hábitos nocturnos...).
3. Probar arquitecturas avanzadas (redes recurrentes, transformers) para captar dependencias temporales.
4. Mejorar la app añadiendo generación automática de informes y conexión con dispositivos de monitoreo.



MUCHAS GRACIAS!!